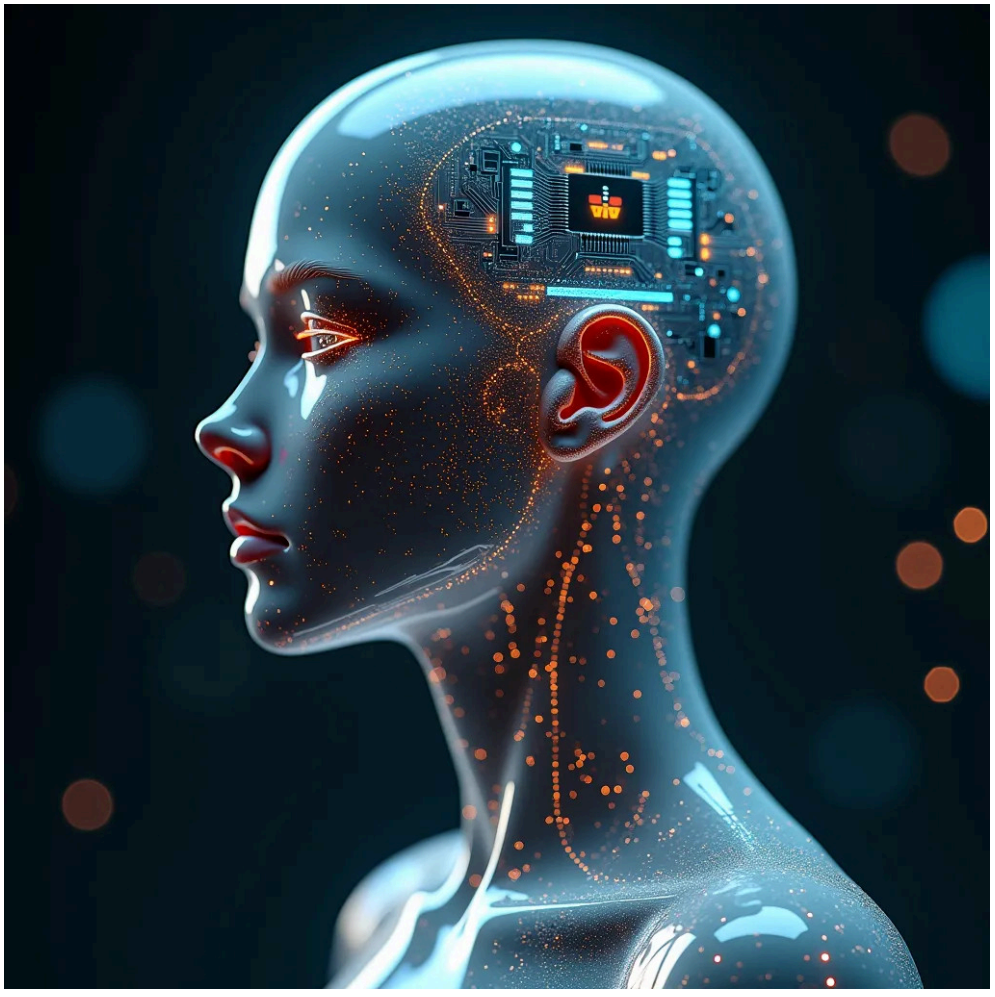
UD06: Principios legales y éticos de la Inteligencia Artificial



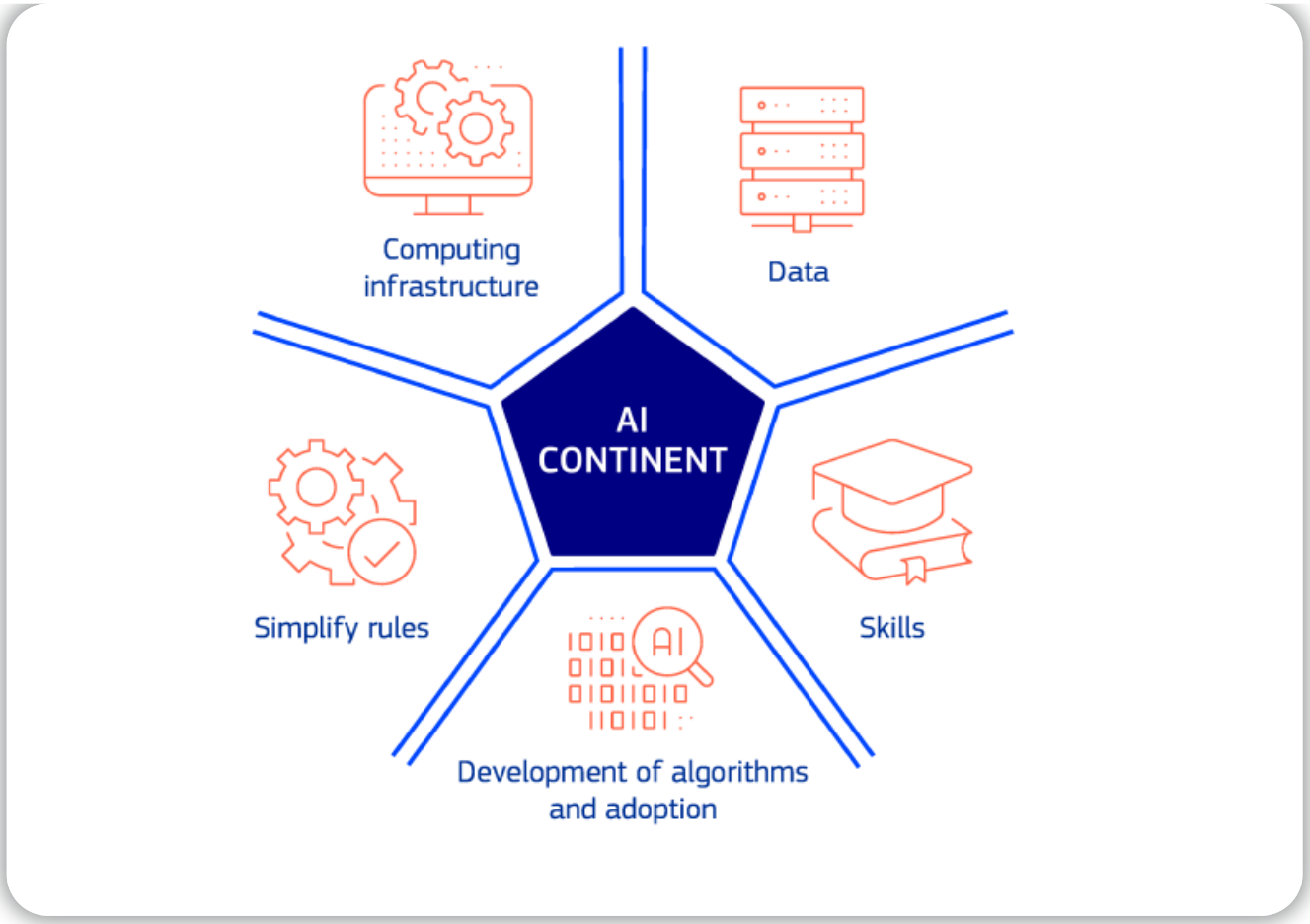
1. Introducción

Al igual que en muchas otras disciplinas científicas y técnicas, a la hora de llevar acabo desarrollos en el campo de la inteligencia artificial, todo profesional debe ser capaz de diferenciar entre lo que sería posible desarrollar en este campo desde el punto de vista técnico y lo que es ético y legal.

El Libro Blanco sobre la inteligencia artificial, realizado bajo la dirección de la Comisión Europea, pone de manifiesto cómo el desarrollo de la inteligencia artificial ha cambiado nuestra vida en los últimos años. Gracias a la rápida evolución de estas tecnologías, se están consiguiendo efectos muy beneficiosos para la sociedad como, por ejemplo, se ha mejorado la atención sanitaria haciendo más precisos los diagnósticos, se han producido mejoras en la agricultura, en los sistemas de producción, en la seguridad, etc.

Es necesario que desde Europa se pueda plantear una respuesta a los retos de la inteligencia artificial, por ello la Comisión Europea ha implementado una estrategia conjunta. Dicha estrategia propone un enfoque para el desarrollo de la inteligencia artificial en Europa, haciendo también énfasis en la regulación de su desarrollo y uso. El enfoque de la UE con respecto a la inteligencia artificial se centra en la excelencia y la confianza, con el objetivo de impulsar la investigación y la capacidad industrial, garantizando al mismo tiempo la seguridad y los derechos fundamentales. El Reglamento (UE) 2024/1689, conocido como **AI Act**, se publicó en el DOUE el 12 de julio de 2024 y, conforme a su art. 113, **entró en vigor el 1 de agosto de 2024**. Sin embargo, la gran mayoría de sus obligaciones solo serán exigibles una vez transcurridos veinticuatro meses, de modo que **su aplicación plena llegará el 2 de agosto de 2026**. El propio reglamento fija fases intermedias: las prohibiciones absolutas y las obligaciones de alfabetización en IA se activan a los seis meses (2 febrero 2025), mientras que las normas para los modelos de IA de propósito general y el nuevo régimen sancionador lo harán al año (2 agosto 2025).

Así, muy recientemente, en abril del 2025 se ha presentado el Plan Continental de Acción para la IA. Este plan establece 5 áreas estratégicas.



1. Infraestructura de computación

- **Fábricas de IA**
  - **Objetivo**: entrenar y afinar modelos de IA
- **Gigafábricas de IA**
  - **Objetivo**: entrenar y desarrollar modelos complejos de IA
    Desarrollo de la Nube e IA
  - **Objetivo**: impulsar la investigación en infraestructura altamente sostenible

2. Datos

- **Estrategia de Unificación de Datos**
  - Mejorar el acceso a datos para empresas y administraciones y simplificar las reglas de datos
  - Los Laboratorios de Datos en las Fábricas de IA recopilarán y curarán datos de alta calidad de diferentes fuentes

3. Habilidades

- **Talento en IA**
  - Formar asociaciones para reclutar internacionalmente

- Ofrecer becas de IA a estudiantes destacados, investigadores y profesionales de IA
- Impulsar habilidades y adopción de IA a través de la Academia de Habilidades de IA
- Programa piloto de un título enfocado en IA generativa
- Apoyar la recualificación a través de Centros Europeos de Innovación Digital

4. Desarrollo de algoritmos y adopción

- **Estrategia de Aplicación de IA**
  - Acelerar la adopción de IA en sectores estratégicos, como salud, automoción y fabricación avanzada
  - Apoyar a empresas y administraciones públicas en el desarrollo y despliegue de soluciones prometedoras de IA

5. Simplificación de reglas

- **Facilitación de la Implementación de la Ley de IA**
  - Aumentar la confianza de los ciudadanos y proporcionar certeza jurídica a través de la Ley de IA
  - Lanzar el Centro de Servicio de la Ley de IA en verano de 2025
  - Proporcionar herramientas y consejos gratuitos y personalizados a las empresas

Al mismo tiempo, debe tenerse en cuenta que **la inteligencia artificial conlleva una serie de riesgos potenciales**. Entre los posibles riesgos de la inteligencia artificial, son citados por el Libro Blanco, a modo de ejemplo cabe señalar, la opacidad en la toma de todo tipo de decisiones, la posible discriminación de género, el uso de los modelos de inteligencia artificial y de los conjuntos de datos para fines delictivos, así como la pérdida de parte de nuestra privacidad.

1. **Opacidad en la toma de decisiones**
   La dificultad para entender cómo y por qué un sistema de IA toma determinadas decisiones, lo que puede dificultar la supervisión y el control humano sobre los resultados.
2. **Discriminación y sesgos**
   Riesgos de discriminación de género, racial u otro tipo, derivados de sesgos tanto en los datos como en los algoritmos utilizados por los sistemas de IA.
   Esto puede traducirse en decisiones injustas o desiguales, especialmente en ámbitos sensibles como la contratación laboral, la educación o el acceso a servicios.
3. **Intromisión en la privacidad y protección de datos**
   Amenazas a la privacidad de las personas y a la protección de los datos personales, debido a la capacidad de la IA para procesar grandes volúmenes de información sensible.

4. **Seguridad y responsabilidad**
   Riesgos para la seguridad física y digital, incluyendo la posibilidad de daños materiales o inmateriales importantes, lesiones o incluso muerte, especialmente en aplicaciones críticas como vehículos autónomos o sistemas médicos.
   Dificultades para atribuir la responsabilidad civil en caso de fallos o daños causados por sistemas de IA, debido a la autonomía y complejidad de estos sistemas.

5. **Uso con fines delictivos**
   Potencial uso de la IA para actividades ilícitas, como fraudes, ciberataques, manipulación de la información o vigilancia masiva.

6. **Impacto sobre los derechos fundamentales**
   Riesgos para los derechos fundamentales de los ciudadanos, incluyendo la libertad, la dignidad, la igualdad y la no discriminación.

.

2. Deontología profesional en inteligencia artificial

Se entiende por deontología la parte de la ética que trata de los deberes, especialmente de los que rigen una actividad profesional. Así, la deontología es el conjunto de deberes relacionados con el ejercicio de una determinada profesión.

Todo profesional dentro de su ámbito de trabajo se enfrenta a dilemas morales y éticos. Existe una serie de problemas éticos que se han presentado en el campo de la inteligencia artificial desde el comienzo de su desarrollo como disciplina científica. Algunos de los más comunes son los que se relacionan a continuación:

- Los procesos de automatización inteligente pueden causar la pérdida de puestos de trabajo, sobre todo los del personal con menor cualificación. De una u otra forma, este problema ético ha estado presente al menos desde la creación de la máquina de vapor en el siglo xviii y el comienzo de la revolución industrial. En la actualidad, todas las industrias son fundamentalmente dependientes del uso de ordenadores y, en algunos casos, también del empleo de algoritmos de inteligencia artificial. Como ejemplos de esta dependencia se pueden citar los programas de credit scoring, que son quienes de forma automática deciden a qué personas se les puede conceder un crédito bancario. Este trabajo, hasta la creación de estos sistemas, era realizado por humanos. Otro ejemplo podría ser cómo, en la actualidad, el uso de la robótica permite que ciertas operaciones rutinarias que se realizan en los almacenes sean llevadas a cabo por robots. Pero, si bien el uso de aplicaciones de inteligencia artificial puede conducir a la destrucción de algunos empleos, no es menos cierto que sirve para la generación de otros trabajos, en general más cualificados y, por tanto, con mayor remuneración.
- El que la inteligencia artificial reemplace a los humanos en muchas de sus tareas podría hacer que el ser humano quedase relegado, perdiendo
  toda utilidad para el trabajo. Aunque esta idea se puso de manifiesto hace ya muchas décadas por parte de escritores de ciencia-ficción como Alvin Toffler o Arthur C. Clarke, parece complicado que se llegue a convertir en una realidad.
- Los sistemas de inteligencia artificial se podrían emplear para usos ilegales o perjudiciales. Efectivamente, toda tecnología puede ser usada para fines ilegales, inmorales o destructivos y evitar que esto ocurra resulta prácticamente imposible.
- La utilización de sistemas de inteligencia artificial puede conducir a una pérdida de la responsabilidad individual. Un ejemplo de ello podría ser un médico que tome la decisión de aplicar un tratamiento a un paciente a partir de un diagnóstico equivocado proporcionado por un sistema de inteligencia artificial. En este caso, se produce un dilema relativo a si el médico, profesional formado, es el culpable por haber aceptado ese diagnóstico como bueno o si bien es el programador del sistema el único responsable de ese error. Además, si se dispone de sistemas expertos que son capaces de diagnosticar mejor que un médico, este podría estar siendo negligente en caso de no usar este tipo de sistemas para llevar a cabo sus diagnósticos.

- El éxito de la inteligencia artificial podría suponer el fin de la raza humana. En el caso de la inteligencia artificial, se trata de una tecnología que, por su propia definición, posiblemente llegase a tomar conciencia de sí misma y llevar a cabo sus propias decisiones. Así por ejemplo, en muchos libros de ciencia ficción se habla acerca de robots asesinos que tratan de acabar con la raza humana. Sin llegar a este extremo, por ejemplo, un error de estimación de un coche autónomo podría causar un accidente de graves consecuencias para los ocupantes de un vehículo. Pero, a favor del vehículo autónomo, se puede decir que ese tipo de errores los cometen los humanos con mucha mayor frecuencia.

En la actualidad, Europa produce más de un cuarto de todos los robots de servicios industriales y profesionales que se fabrican en el mundo y desempeña un papel importante en el desarrollo y uso de las aplicaciones informáticas para empresas y organizaciones, así como de las aplicaciones para el fomento de la administración digital y las aplicaciones de «empresas inteligentes».

Dentro del ecosistema de excelencia en inteligencia artificial que se pretende desarrollar en Europa, resulta necesario disponer de profesionales cualificados en este campo. Esto supone el desarrollo de las habilidades necesarias para trabajar en este campo y mejorar las cualificaciones profesionales de los trabajadores para adaptarlas a la transformación que implica esta tecnología. Dentro del currículo de los profesionales de este campo, el Libro Blanco sugiere la inclusión de competencias éticas.

Con el fin de lograr este propósito, en el año 2019 la Unión Europea publicó el documento titulado Directrices éticas para una IA fiable. El objetivo de estas directrices es promover una inteligencia artificial fiable. Según dicho documento, la fiabilidad de la inteligencia artificial se apoya en tres componentes que han de satisfacerse a lo largo de todo el ciclo de vida del sistema. En primer lugar, la inteligencia artificial debe ser lícita, es decir, cumplir todas las leyes y reglamentos aplicables; además, ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos y, finalmente, tiene que ser robusta, tanto desde el punto de vista técnico como social, puesto que los sistemas de inteligencia artificial, incluso si las intenciones son buenas, pueden provocar daños a terceros.

3. La privacidad de los datos

Si bien la mayor parte del desarrollo legislativo y la regulación de la protección de los datos de carácter personal, tanto en España como en Europa, se produjo a partir de 2010, ya la Constitución Española, promulgada en 1978, recoge en su artículo 18.4 que «La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos».

Es bien sabido que Internet ofrece grandes oportunidades a la ciudadanía, pero también supone un riesgo considerable en todo lo referente a su privacidad. Resulta por tanto necesaria la existencia de un marco normativo que haga efectivos los derechos de la ciudadanía en Internet, promoviendo la igualdad.

En la actualidad, en el Reino de España, la concreción y desarrollo del derecho fundamental de protección de las personas físicas en relación con el tratamiento de los datos personales; se encuentra recogido en la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales. Además, dentro del ámbito de la Unión Europea, en el año 2016 se adoptó el Reglamento (UE) del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de esos datos.

La ya mencionada Ley Orgánica 3/2018 recoge aspectos de gran interés e importancia para la ciudadanía. Se resumen a continuación algunos de los considerados como más relevantes:

- **Derecho de acceso:** todo ciudadano tiene derecho a solicitar acceso a los datos que de él dispone cualquier persona física o jurídica. Esa solicitud se canalizará a través del correspondiente responsable de los datos. Cuando el responsable disponga de una gran cantidad de datos relativos al afectado y este ejercite su derecho de acceso, deberá especificar si el acceso se refiere a la totalidad o a parte de esa información.
- **Derecho de acceso a los datos de personas fallecidas:** los familiares de una persona fallecida pueden dirigirse al responsable o encargado del tratamiento de los datos de una compañía que posea información relativa a su familiar, al objeto de solicitar el acceso a los datos del fallecido y permitir su rectificación o supresión, salvo que el fallecido hubiese designado expresamente un responsable de sus datos o bien hubiera prohibido expresamente el acceso a los mismos tras su fallecimiento.
- **Derecho de rectificación, supresión y limitación del tratamiento:** toda persona física tiene derecho a la rectificación, supresión o limitación del tratamiento que se haga de sus datos personales por parte de cualquier persona física o jurídica.

- **Sistemas de información crediticia:** en el caso de los sistemas de información crediticia, salvo prueba en contra, se presumirá lícito el tratamiento de datos relativo al incumplimiento de obligaciones dinerarias, financieras o de crédito por sistemas comunes de información crediticia, cuando los datos hayan sido facilitados por el acreedor o su representante, los datos se refieran a deudas vencidas que no hayan sido objeto de reclamación administrativa o judicial por parte del deudor y siempre que el acreedor haya informado al afectado en el contrato o en el momento de requerir el pago acerca de la posibilidad de inclusión en dichos sistemas. Finalmente, cabe destacar que los datos se mantendrán en el sistema únicamente mientras persista el incumplimiento y con un límite máximo de 5 años desde la fecha de vencimiento de la obligación dineraria, financiera o de crédito.
- **Tratamiento de la información con fines de videovigilancia:** en lo relativo a la videovigilancia, solamente se permite la captación de imágenes en la vía pública en la medida en que esta resulte imprescindible para garantizar la seguridad de las personas y de los bienes, teniendo los datos que ser suprimidos en el plazo máximo de 1 mes desde su captación, salvo que exista causa que justifique la necesidad de conservarlos para acreditar la comisión de actos que atenten contra la integridad de bienes o personas, en cuyo caso se tendrán que poner a disposición judicial en menos de 72 horas.
- **Sistemas de exclusión publicitaria:** cuando un ciudadano se pone en contacto con una compañía para pedir su exclusión publicitaria, esta le habrá de informar de los sistemas disponibles para dicha exclusión. Se puede restringir la recepción de publicidad no deseada mediante la inscripción gratuita y voluntaria en un fichero de exclusión publicitaria. Actualmente solo existe el fichero denominado Lista Robinson que está gestionado por la Asociación Española de Economía Digital. Esta lista debe ser consultada por quienes vayan a realizar una campaña publicitaria para excluir de la misma a las personas inscritas. Aquel ciudadano que esté inscrito en un fichero de exclusión publicitaria, únicamente recibirá publicidad de las compañías de las que sea cliente o de aquellas a las que haya autorizado expresamente a enviarle publicidad.
- **Derecho a la neutralidad de Internet y de acceso universal:** este derecho significa que los proveedores de servicios de Internet proporcionarán una oferta transparente de servicios sin discriminación por motivos técnicos o económicos. Además, se garantiza el acceso a Internet, independientemente de la condición personal, social, económica o geográfica de cada individuo.
- **Protección de los menores en Internet:** dados los posibles peligros a los que se exponen los menores en sus comunicaciones a través de Internet, la Ley Orgánica 3/2018 especifica que deben ser los tutores de los menores los que procuren que estos hagan un uso responsable de la red.

- **Derecho a la desconexión digital e intimidad frente al uso de dispositivos de videovigilancia, grabación y geolocalización:** la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales proporciona un marco regulatorio para los derechos de los trabajadores, alguno de ellos tan recientes como el derecho a la desconexión digital. El derecho a la desconexión digital consiste en que tanto los trabajadores como los empleados públicos tienen derecho a la desconexión de sus dispositivos electrónicos de trabajo a fin de garantizar, fuera de su jornada laboral, el respeto de su descanso e intimidad personal y familiar. Dentro del derecho a la desconexión digital se indica expresamente que este derecho se preservará también en el supuesto de realización total o parcial del trabajo desde el domicilio del empleado. También se regula en qué lugar y bajo qué condiciones resulta posible la instalación de dispositivos para la grabación de imágenes y sonido. Así, según se indica en la Ley, en ningún caso se admitirá la instalación de sistemas de grabación de sonidos ni de videovigilancia en lugares destinados al descanso o esparcimiento de los trabajadores tales como vestuarios, aseos, comedores, etc. En lo referente a la instalación de dispositivos geolocalizadores, los empleadores habrán de informar con carácter previo de la existencia de dichos dispositivos a sus empleados aunque podrán emplear la información obtenida de los mismos para el control de las labores del empleado aunque siempre dentro del marco legal.
- **El derecho al olvido:** consiste en que toda persona tiene derecho a que los motores de búsqueda en Internet eliminen de las listas de resultados que se obtuvieran los enlaces relativos a una persona cuando estos fuesen inadecuados, inexistentes, no pertinentes, etc. De igual manera también existe el derecho al olvido en los servicios de las redes sociales y equivalentes, teniendo el usuario derecho a que sean suprimidos de toda red sus datos personales cuando así lo solicite.

4. Protección frente a errores

Los sistemas de inteligencia artificial cometen errores. Muchos de estos fallos son fácilmente detectables durante el desarrollo de los sistemas, pero otros pasan inadvertidos y se manifiestan durante el uso del mismo. Este tipo de errores pueden ser de muy diversa índole.

Así, por ejemplo, hoy en día existen sistemas de reconocimiento de imágenes muy sofisticados que son, capaces de etiquetar objetos. Gracias a la existencia de estas tecnologías se han conseguido desarrollos como el coche autónomo, pero dichos sistemas no están exentos de fallos. Seguidamente se exponen algunos errores relacionados con aplicaciones de la inteligencia artificial en distintos campos.

Ejemplo En 2015, el ingeniero informático Jacky Alciné se percató de que el algoritmo de reconocimiento de imágenes de Google Photos había etiquetado a algunos de sus amigos de raza negra como gorilas. Se trataba de un error producido por el algoritmo de inteligencia artificial de dicha aplicación e inmediatamente, Google pidió disculpas por ello prometiendo resolver el problema. Sin embargo, 3 años después, un reportaje publicado en la revista Wired demostró que este problema no estaba todavía resuelto. Realmente, la solución por la que había optado Google con el fin de evitar problemas con sus usuarios fue eliminar la etiqueta gorila así como la de algunos otros primates como chimpancé o mono. Es decir, las funcionalidades de la aplicación se habían limitado dado que no habían sido capaces de corregir el fallo. Este hecho pone de manifiesto la existencia de ciertas dificultades técnicas en el tratamiento de imágenes que todavía no han sido superadas.

Ejemplo El 18 de marzo de 2018 tuvo lugar el primer accidente con víctimas mortales causado por un coche autónomo. Este accidente se produjo cuando un coche autónomo de Uber arrolló en Tempe (Arizona, Estados Unidos) a una mujer que atravesaba una carretera de cuatro carriles empujando una bicicleta. El vehículo de Uber que atropelló a esta mujer operaba en modo autónomo pero asistido por un conductor quien, en caso de que existieran posibilidades de sufrir un accidente, debería de asumir los mandos del vehículo. Tras la investigación del suceso, se encontraron distintas causas que podrían haber influido en el mismo, algunas de las cuales se relacionan a continuación:

- En el momento del accidente, el coche se desplazaba a una velocidad de 69 km/h.
- El sistema podría haber visto al peatón 6 segundos antes de la colisión, pero durante 4,7 segundos no hizo nada por evitarla.
- El conductor que iba en el vehículo para garantizar la seguridad del mismo, dado que se trataba de un vehículo en pruebas, estaba distraído en el momento en el que produjo el accidente.
- La forma en la que el peatón cruzaba la carretera, empujando una bicicleta y en una hora de baja visibilidad, no era segura.
- La zona de la carretera por donde el peatón estaba cruzando tenía una señal que indicaba que estaba prohibido el paso a los peatones.
- Un peritaje realizado por expertos mostró que, en un primer momento, el software del coche autónomo falló en la detección del peatón. En este contexto, el que el peatón estuviese empujando una bicicleta (objeto metálico) pudo haber influido en la deficiente detección que el sistema hizo del mismo.

Tras el accidente, Uber suspendió las pruebas de su coche autónomo. Transcurridos unos meses, Uber emitió un informe en el que se afirmaba que con el grado de desarrollo que en esos momentos tenían sus coches autónomos, estos eran más seguros que los conducidos por humanos. El 20 de diciembre de 2018 Uber reanudó sus pruebas.

Ejemplo De entre de todos los posibles fallos que hoy en día presentan los asistentes de voz como Siri, Alexa o Cortana, se expone a continuación el que se produjo en una presentación que realizó el consejero delegado de Microsoft Satya Nadella en 2015. Se trataba de una conferencia de la empresa Salesforce.com, compañía dedicada al desarrollo de aplicaciones de software para la relación con el cliente.

En dicha conferencia Nadella trataba de mostrar las capacidades de análisis y entendimiento del habla de Cortana, asistente virtual de Microsoft, para lo que le preguntó al sistema de viva voz en inglés «Show me my most at-risk opportunities» lo que se podría traducir como «muéstrame mis oportunidades de mayor riesgo», lo que Cortana interpretó como «Show me to buy milk at this opportunity» que quiere decir algo así como «enséñame a comprar leche en esta oportunidad».

Aunque Nadella se lo repitió varias veces, Cortana no fue capaz de entenderlo y hubo de reformular la frase. La causa raíz de este fallo es que el sistema de reconocimiento de voz de Cortana, entrenado para entender a una persona hablando en inglés con acento norteamericano, no fue capaz de entender con la misma precisión lo que decía Nadella, nacido en la India aunque residente en Estados Unidos desde finales de la década de 1980.

En ingeniería de software, cuando se desarrolla un programa que no hace uso de inteligencia artificial, es el programador quien escribe el código fuente que interacciona con el usuario para llevar a cabo cierta tarea. En este contexto, las pruebas de software ayudan a garantizar que el programa funcione tal y como se espera. Sin embargo, en los sistemas de aprendizaje automático, la programación consiste en proporcionar al sistema ejemplos de cómo debe comportarse y hacer que este aprenda. Por tanto, es necesario llevar a cabo un minucioso proceso de prueba que garantice que con la información proporcionada el sistema ha sido capaz de aprender correctamente. Aunque este problema no es exclusivo de la IA, ya que algunos algoritmos pueden no ser públicos, lo que en algunas situaciones puede presentar graves problemas como la opacidad de los algoritmos que nos juzgan

5. Principios éticos

Dado que la IA es una tecnología poderosa, tenemos la obligación moral de utilizarla bien, promover los aspectos positivos y evitar o mitigar los negativos.

Los aspectos positivos son muchos. Por ejemplo, la IA puede salvar vidas mediante mejores diagnósticos médicos, nuevos descubrimientos médicos, una mejor predicción de fenómenos meteorológicos extremos y una conducción más segura con asistencia al conductor y (eventualmente) tecnologías de conducción autónoma. También hay muchas oportunidades para mejorar vidas. El programa AI for Humanitarian Action de Microsoft aplica la IA para recuperarse de desastres naturales, abordar las necesidades de los niños, proteger a los refugiados y promover los derechos humanos. El programa AI for Social Good de Google apoya el trabajo sobre la protección de la selva tropical, la jurisprudencia de derechos humanos, el monitoreo de la contaminación, la medición de las emisiones de combustibles fósiles, el asesoramiento en crisis, la verificación de noticias, la prevención del suicidio, el reciclaje y otros temas. El Centro de Ciencia de Datos para el Bien Social de la Universidad de Chicago aplica el aprendizaje automático a problemas de justicia penal, desarrollo económico, educación, salud pública, energía y medio ambiente.

Las aplicaciones de IA en el manejo de cultivos y la producción de alimentos ayudan a alimentar al mundo. La optimización de los procesos empresariales mediante el aprendizaje automático hará que las empresas sean más productivas, aumentará la riqueza y generará más empleo. La automatización puede reemplazar las tareas tediosas y peligrosas que enfrentan muchos trabajadores y liberarlos para concentrarse en aspectos más interesantes. Las personas con discapacidad se beneficiarán de la asistencia basada en inteligencia artificial para ver, oír y moverse. La traducción automática ya permite comunicarse entre personas de diferentes culturas. Las soluciones de IA basadas en software tienen un costo marginal de producción casi nulo y, por lo tanto, tienen el potencial de democratizar el acceso a la tecnología avanzada (incluso cuando otros aspectos del software tienen el potencial de centralizar el poder).

A pesar de estos muchos aspectos positivos, no debemos ignorar los negativos. Muchas tecnologías nuevas han tenido efectos secundarios negativos no deseados: la fisión nuclear provocó Chernobyl y la amenaza de destrucción global; el motor de combustión interna trajo contaminación del aire, calentamiento global y la pavimentación del paraíso. Otras tecnologías pueden tener efectos negativos incluso cuando se utilizan según lo previsto, como el gas sarín, los rifles AR-15 y las solicitudes telefónicas. La automatización creará riqueza, pero en las condiciones económicas actuales gran parte de esa riqueza fluirá hacia los propietarios de los sistemas automatizados, lo que conducirá a una mayor desigualdad de ingresos. Esto puede ser perjudicial para una sociedad que funcione bien. En los países en desarrollo, el camino tradicional hacia el crecimiento a través de la fabricación de bajo costo para la exportación puede verse cortado, a medida que los países ricos adopten instalaciones de fabricación en el país totalmente automatizadas. Nuestras decisiones éticas y de gobernanza dictarán el nivel de desigualdad que generará la IA.

Todos los científicos e ingenieros enfrentan consideraciones éticas sobre qué proyectos deben o no emprender y cómo pueden asegurarse de que la ejecución del proyecto sea segura y beneficiosa. En 2010, el Consejo de Investigación en Ingeniería y Ciencias Físicas del Reino Unido celebró una reunión para desarrollar un conjunto de Principios de la Robótica. En los años siguientes, otras agencias gubernamentales,
organizaciones sin fines de lucro y empresas crearon conjuntos de principios similares. La esencia es que cada organización que crea tecnología de IA, y todos los miembros de la organización, tienen la responsabilidad de asegurarse de que la tecnología contribuya al bien y no al daño. Los principios más comúnmente citados son:

- Garantizar la seguridad
- Establecer responsabilidad
- Garantizar la equidad
- Defender los derechos y valores humanos.
- Respetar la privacidad
- Reflejar diversidad/inclusión
- Promover la colaboración
- Evitar la concentración de poder.
- Proporcionar transparencia.
- Reconocer las implicaciones legales/políticas.
- Limitar los usos nocivos de la IA.
- Contemplar las implicaciones para el empleo.

Tenga en cuenta que muchos de los principios, como "garantizar la seguridad", son aplicables a todos los sistemas de software o hardware, no solo a los sistemas de inteligencia artificial. Varios principios están redactados de manera vaga, lo que dificulta su medición o aplicación. Esto se debe en parte a que la IA es un gran campo con muchos subcampos, cada uno de los cuales tiene un conjunto diferente de normas históricas y diferentes relaciones entre los desarrolladores de IA y las partes interesadas. Mittelstadt (2019) sugiere que cada uno de los subcampos debería desarrollar pautas procesables y precedentes de casos más específicos.

5.1. Armas autónomas letales

La ONU define un arma letal autónoma como aquella que localiza, selecciona y ataca (es decir, mata) objetivos humanos sin supervisión humana. Varias armas cumplen algunos de estos criterios. Por ejemplo, las minas terrestres se utilizan desde el siglo XVII: pueden seleccionar y atacar objetivos de forma limitada según el grado de presión ejercida o la cantidad de metal presente, pero no pueden salir y localizar objetivos por sí mismas. (Las minas terrestres están prohibidas en virtud del Tratado de Ottawa). Los misiles guiados, utilizados desde la década de 1940, pueden perseguir objetivos, pero un ser humano debe apuntarlos en la dirección general correcta. Los cañones de disparo automático controlados

por radar se han utilizado para defender buques de guerra desde la década de 1970; Su objetivo principal es destruir los misiles entrantes, pero también podrían atacar aviones tripulados. Aunque la palabra "autónoma" se utiliza a menudo para describir vehículos aéreos no tripulados o drones, la mayoría de estas armas son pilotadas de forma remota y requieren la actuación humana de la carga letal.

En el momento de escribir este artículo, varios sistemas de armas parecen haber cruzado la línea hacia la plena autonomía. Por ejemplo, el misil Harop de Israel es una "munición merodeadora" con una envergadura de tres metros y una ojiva de cincuenta libras. Busca durante hasta seis horas en una región geográfica determinada cualquier objetivo que cumpla un criterio determinado y luego lo destruye. El criterio podría ser "emite una señal de radar que se asemeja a un radar antiaéreo" o "parece un tanque". El fabricante turco STM anuncia su cuadricóptero Kargu, que transporta hasta 1,5 kg de explosivos, como capaz de "impactar de forma autónoma... objetivos seleccionados en imágenes... rastrear objetivos en movimiento... antipersonal... reconocimiento facial".

Las armas autónomas han sido llamadas la "tercera revolución en la guerra" después de la pólvora y las armas nucleares. Su potencial militar es obvio. Por ejemplo, pocos expertos dudan de que los aviones de combate autónomos derrotarían a cualquier piloto humano. Los aviones, tanques y submarinos autónomos pueden ser más baratos, más rápidos, más maniobrables y tener mayor alcance que sus homólogos tripulados. La Guerra de Ucrania es un claro ejemplo del desarrollo de este tipo de armamento autónomo.

Desde 2014, las Naciones Unidas en Ginebra han llevado a cabo debates periódicos bajo los auspicios de la Convención sobre Ciertas Armas Convencionales (CAC) sobre la cuestión de si se deben prohibir las armas letales autónomas. En el momento de redactar este informe, 30 naciones, cuyo tamaño varía desde China hasta la Santa Sede, han declarado su apoyo a un tratado internacional, mientras que otros países clave, incluidos Israel, Rusia, Corea del Sur y Estados Unidos, se oponen a un tratado. prohibición.

El debate sobre las armas autónomas incluye aspectos legales, éticos y prácticos. Las cuestiones jurídicas se rigen principalmente por la CCW, que exige la posibilidad de discriminar entre combatientes y no combatientes, el juicio sobre la necesidad militar de un ataque y la evaluación de la proporcionalidad entre el valor militar de un objetivo y la posibilidad de daños colaterales.

La viabilidad de cumplir estos criterios es una cuestión de ingeniería, cuya respuesta sin duda cambiará con el tiempo. En la actualidad, la discriminación parece factible en algunas circunstancias y sin duda mejorará rápidamente, pero la necesidad y la proporcionalidad no son factibles actualmente: requieren que las máquinas hagan juicios subjetivos y situacionales que son considerablemente más difíciles que las tareas relativamente simples de buscar y atacar objetivos potenciales. . Por estas razones, sería legal usar armas autónomas solo en circunstancias en las que un operador humano pueda predecir razonablemente que la ejecución de la misión no resultará en que los civiles sean atacados o que las armas realicen ataques innecesarios o desproporcionados. Esto significa que, por el momento, sólo se pueden realizar misiones muy restringidas con armas autónomas.

Desde el punto de vista ético, algunos consideran simplemente moralmente inaceptable delegar la decisión de matar humanos a una máquina. Por ejemplo, el embajador de Alemania en Ginebra ha declarado que "no aceptará que la decisión sobre la vida o la muerte sea tomada únicamente por un sistema autónomo", mientras que Japón "no tiene ningún plan para desarrollar robots con humanos fuera del circuito, que tal vez sean capaces de cometer asesinato" . El general Paul Selva, en ese momento el segundo oficial militar de Estados Unidos, dijo en 2017: "No creo que sea razonable que pongamos a los robots a cargo de si matamos o no una vida humana". Finalmente, António Guterres, jefe de las Naciones Unidas, afirmó en 2019 que "las máquinas con el poder y la discreción de quitar vidas sin participación humana son políticamente inaceptables, moralmente repugnantes y deberían estar prohibidas por el derecho internacional" .

Más de 140 ONG en más de 60 países forman parte de la Campaña para detener a los robots asesinos, y una carta abierta organizada en 2015 por el Future of Life Institute fue firmada por más de 4.000 investigadores de IA2 y 22.000 más.

En contra de esto, se puede argumentar que a medida que la tecnología mejore, debería ser posible desarrollar armas que tengan menos probabilidades que los soldados o pilotos humanos de causar víctimas civiles. (También existe el importante beneficio de que las armas autónomas reducen la necesidad de soldados humanos y pilotos corren el riesgo de morir.) Los sistemas autónomos no sucumbirán a la fatiga, la frustración, la histeria, el miedo, la ira o la venganza, y no necesitan "disparar primero, hacer preguntas después" (Arkin, 2015). Así como las municiones guiadas han reducido el daño colateral En comparación con las bombas no guiadas, se puede esperar que las armas inteligentes mejoren aún más la precisión de los ataques. (En contra de esto, ver Benjamin (2013) para un análisis de las víctimas de la guerra con aviones no tripulados). Esta, aparentemente, es la posición de Estados Unidos en la última ronda de negociaciones en Ginebra.

Quizás contrariamente a lo intuitivo, Estados Unidos es también una de las pocas naciones cuyas propias políticas actualmente excluyen el uso de armas autónomas. La hoja de ruta del Departamento de Defensa de Estados Unidos (Directiva 3000.09 de 2012, actualizada 2017) dice: "En el futuro previsible, las decisiones sobre el uso de la fuerza [por sistemas autónomos] y la elección de qué objetivos individuales atacar con fuerza letal se mantendrán bajo control humano" . La razón principal de esta política es práctica: los sistemas autónomos no son lo suficientemente confiables como para confiarles decisiones militares.

La cuestión de la confiabilidad pasó a primer plano el 26 de septiembre de 1983, cuando la pantalla de la computadora del oficial de misiles soviético Stanislav Petrov mostró una alerta de un ataque con misiles inminente. Según el protocolo, Petrov debería haber iniciado un contraataque nuclear, pero sospechaba que la alerta era un error y lo traté como tal. Tenía razón y la Tercera Guerra Mundial se evitó (por poco). No sabemos qué habría pasado si no hubiera habido ningún ser humano en el circuito.

La confiabilidad es una preocupación muy seria para los comandantes militares, quienes conocen bien la complejidad de las situaciones en el campo de batalla. Los sistemas de aprendizaje automático que funcionan perfectamente durante la capacitación pueden tener un rendimiento deficiente cuando se implementan. Los ciberataques contra armas autónomas podrían provocar bajas por fuego amigo; desconectar el arma de toda comunicación puede

evitarlo (suponiendo que aún no haya sido comprometida), pero entonces el arma no puede recuperarse si no funciona correctamente.

La cuestión práctica primordial con las armas autónomas es que son armas de destrucción masiva escalables, en el sentido de que la escala de un ataque que se puede lanzar es proporcional a la cantidad de hardware que uno puede permitirse desplegar. Un cuadricóptero de cinco centímetros de diámetro puede transportar una carga explosiva letal, y un millón de ellos caben en un contenedor de transporte normal. Precisamente porque son autónomas, estas armas no necesitarían un millón de supervisores humanos para hacer su trabajo.

Como armas de destrucción masiva, las armas autónomas escalables tienen ventajas para el atacante en comparación con las armas nucleares y los bombardeos masivos: dejan la propiedad intacta y pueden usarse selectivamente para eliminar sólo a aquellos que puedan amenazar a una fuerza ocupante. Sin duda, podrían utilizarse para eliminar a todo un grupo étnico o a todos los seguidores de una religión determinada. En muchas situaciones, también serían imposibles de rastrear. Estas características los hacen particularmente atractivos para los actores no estatales.

Estas consideraciones –particularmente aquellas características que benefician al atacante– sugieren que las armas autónomas reducirán la seguridad global y nacional de todas las partes. La respuesta racional de los gobiernos parece ser involucrarse en discusiones sobre control de armas en lugar de una carrera armamentista.

Sin embargo, el proceso de diseño de un tratado no está exento de dificultades. La IA es una tecnología de doble uso: las tecnologías de IA que tienen aplicaciones pacíficas, como control de vuelo, seguimiento visual, cartografía, navegación y planificación multiagente, pueden aplicarse fácilmente con fines militares. Es fácil convertir un cuadricóptero autónomo en un arma simplemente colocando un explosivo y ordenándole que busque un objetivo. Para abordar esto será necesario implementar cuidadosamente regímenes de cumplimiento con la cooperación de la industria, como ya ha demostrado con cierto éxito la Convención sobre Armas Químicas.

5.2. Vigilancia, seguridad y privacidad

En 1976, Joseph Weizenbaum advirtió que la tecnología de reconocimiento automatizado de voz podría dar lugar a escuchas telefónicas generalizadas y, por tanto, a una pérdida de libertades civiles. Hoy en día, esa amenaza se ha hecho realidad: la mayoría de las comunicaciones electrónicas pasan por servidores centrales que pueden ser monitoreados y las ciudades están repletas de micrófonos y cámaras que pueden identificar y rastrear a las personas basándose en su voz, rostro y modo de andar. La vigilancia que solía requerir recursos humanos costosos y escasos ahora puede realizarse a gran escala mediante máquinas.

En 2025, las cámaras de vigilancia en China superan los 700 millones, frente a los 350 millones que había en 2018. En 2024 se estima que había un valor superior a los 85 millones en Estados Unidos. En la Unión Europea se estima un valor entre los 10 y 15 millones de cámaras (≈ 1 por cada 30-45 habitantes). China y otros países han comenzado a exportar tecnología de vigilancia a países de baja tecnología, algunos con reputación de maltratar a sus ciudadanos y apuntar desproporcionadamente a comunidades marginadas. Los ingenieros de IA deben tener claro qué usos de la vigilancia son compatibles con los derechos humanos y negarse a trabajar en aplicaciones que sean incompatibles.

A medida que más instituciones nuestras operan en línea, nos volvemos más vulnerables al cibercrimen (phishing, fraude con tarjetas de crédito, redes de bots, ransomware) y al ciberterrorismo (incluidos ataques potencialmente mortales como el cierre de hospitales y plantas de energía o el control de vehículos autónomos). El aprendizaje automático puede ser una herramienta poderosa para ambas partes en la batalla de la ciberseguridad. Los atacantes pueden utilizar la automatización para detectar inseguridades y pueden aplicar aprendizaje reforzado para intentos de phishing y chantaje automatizado. Los defensores pueden utilizar el aprendizaje no supervisado para detectar patrones de tráfico entrante anómalos (Chandola et al., 2009; Malhotra et al., 2015) y diversas técnicas de aprendizaje automático para detectar fraude (Fawcett y Provost, 1997; Bolton y Hand, 2002). A medida que los ataques se vuelven más sofisticados, todos los ingenieros, no solo los expertos en seguridad, tienen una mayor responsabilidad de diseñar sistemas seguros desde el principio. Un pronóstico (Kanal, 2017) situaban el mercado del aprendizaje automático en ciberseguridad en aproximadamente 100 mil millones de dólares para 2021. La realidad ha sido sensiblemente más modesta: el estudio de referencia de **Markets & Markets** cifra el mercado en **22,4 M US$ en 2023** y pronostica que, pese a crecer a un sólido 21,9 % anual compuesto, solo alcanzará **60,6 M US$ en 2028**. Esta brecha entre la expectativa inicial y los datos efectivos refleja hasta qué punto la adopción de soluciones de IA defensiva depende de factores como la madurez tecnológica, la disponibilidad de talento especializado y la progresiva entrada en vigor de marcos normativos específicos.

A medida que interactuamos con las computadoras durante una cantidad cada vez mayor de nuestra vida diaria, los gobiernos y las corporaciones recopilan más datos sobre nosotros. Los recolectores de datos tienen la responsabilidad moral y legal de ser buenos administradores de los datos que poseen. En los Estados Unidos, la Ley de Responsabilidad y Portabilidad del Seguro Médico (HIPAA) y la Ley de Privacidad y Derechos Educativos de la Familia (FERPA) protegen la privacidad de los registros médicos y de los estudiantes. El Reglamento General de Protección de Datos (GDPR) de la Unión Europea exige que las empresas diseñen sus sistemas teniendo en cuenta la protección de datos y exige que obtengan el consentimiento del usuario para cualquier recopilación o procesamiento de datos.

En contraposición al derecho del individuo a la privacidad está el valor que la sociedad obtiene al compartir datos. Queremos poder detener a los terroristas sin oprimir la disidencia pacífica y queremos curar enfermedades sin comprometer el derecho de ningún individuo a mantener en privado su historial médico. Una práctica clave es la desidentificación: eliminar la información de identificación personal (como el nombre y el número de seguro social) para que los investigadores médicos puedan utilizar los datos para promover el bien común. El problema es que los datos no identificados compartidos pueden estar sujetos a reidentificación. Por ejemplo, si los datos excluyen el nombre, el número de seguro social y la dirección postal, pero incluyen la fecha de nacimiento, el sexo y el código postal, entonces, como lo muestra Latanya Sweeney (2000), el 87% de la población estadounidense puede ser reidentificado de forma

única. Sweeney enfatizó este punto al volver a identificar el historial médico del gobernador de su estado cuando ingresó en el hospital. En la competencia del Premio Netflix, se publicaron registros anónimos de calificaciones de películas individuales y se pidió a los competidores que idearan un algoritmo de aprendizaje automático que pudiera predecir con precisión qué películas le gustarían a un individuo. Pero los investigadores pudieron volver a identificar a usuarios individuales haciendo coincidir la fecha de una clasificación en la base de datos de Netflix con la fecha de una clasificación similar en Internet Movie Database (IMDB), donde los usuarios a veces usan sus nombres reales (Narayanan y Shmatikov, 2006).).

Este riesgo se puede mitigar en cierta medida generalizando campos: por ejemplo, reemplazando la fecha de nacimiento exacta solo con el año de nacimiento, o un rango más amplio como "20-30 años". Eliminar un campo por completo puede verse como una forma de generalizar a "cualquiera". Pero la generalización por sí sola no garantiza que los registros estén a salvo de una reidentificación; Es posible que solo haya una persona en el código postal 94720 que tenga entre 90 y 100 años. Una propiedad útil es el k-anonimato: una base de datos se k-anonimiza si cada registro en la base de datos es indistinguible de al menos k-1 otros registros. Si hay registros que son más únicos que este, habría que generalizarlos aún más.

Una alternativa a compartir registros no identificados es mantener todos los registros privados, pero permitir consultas agregadas. Se proporciona una API para consultas a la base de datos y las consultas válidas reciben una respuesta que resume los datos con un recuento o promedio. Pero no se da ninguna respuesta si ello violaría ciertas garantías de privacidad. Por ejemplo, podríamos permitir que un epidemiólogo preguntara, para cada código postal, el porcentaje de personas con cáncer. Para códigos postales con al menos n personas, se daría un porcentaje (con una pequeña cantidad de ruido aleatorio), pero no se daría respuesta para códigos postales con menos de n personas.

Ejemplo Se debe tener cuidado para protegerse contra la desidentificación mediante consultas múltiples. Por ejemplo, si la consulta "salario promedio y número de empleados de la empresa XYZ de 30 a 40 años" da la respuesta [81.234,12 €] y la consulta "salario promedio y número de empleados de la empresa XYZ de 30 a 41 años" da la respuesta [81.199,13€], y si usamos LinkedIn para encontrar al hombre de 41 años en la compañía XYZ, entonces lo hemos identificado con éxito y podemos calcular su salario exacto, a pesar de que todas las respuestas involucraron a 12 o más personas. El sistema debe diseñarse cuidadosamente para protegerse contra esto, con una combinación de límites en las consultas que se pueden realizar (tal vez sólo se pueda consultar un conjunto predefinido de rangos de edad que no se superpongan) y la precisión de los resultados (tal vez ambas consultas den la respuesta "alrededor de 81.000 €").

Una garantía más sólida es la privacidad diferencial, que garantiza que un atacante no pueda utilizar consultas para volver a identificar a ningún individuo en la base de datos, incluso si el atacante puede realizar múltiples consultas y tiene acceso a bases de datos vinculadas separadas. La respuesta a la consulta emplea un algoritmo aleatorio que agrega una pequeña cantidad de ruido al resultado. En otras palabras, el hecho de que una persona decida participar o no en la base de datos no supone una diferencia apreciable en las respuestas que cualquiera pueda obtener y, por lo tanto, no existe ningún desincentivo de privacidad para participar. Muchas bases de datos están diseñadas para garantizar una privacidad diferencial.

Hasta ahora hemos considerado la cuestión de compartir datos no identificados de una base de datos central. Un enfoque llamado aprendizaje federado no tiene una base de datos central; en cambio, los usuarios mantienen sus propias bases de datos locales que mantienen la privacidad de sus datos. Sin embargo, pueden compartir parámetros de un modelo de aprendizaje automático que se mejora con sus datos, sin el riesgo de revelar ninguno de los datos privados. Imagine una aplicación de comprensión del habla que los usuarios puedan ejecutar localmente en su teléfono. La aplicación contiene una red neuronal básica, que luego se mejora mediante entrenamiento local sobre las palabras que se escuchan en el teléfono del usuario. Periódicamente, los propietarios de la aplicación encuestan a un subconjunto de usuarios y les preguntan los valores de los parámetros de su sistema local mejorado. red, pero no para ninguno de sus datos sin procesar. Los valores de los parámetros se combinan para formar un nuevo modelo mejorado que luego se pone a disposición de todos los usuarios, para que todos obtengan el beneficio de la capacitación realizada por otros usuarios.

Para que este esquema preserve la privacidad, debemos poder garantizar que los parámetros del modelo compartidos por cada usuario no puedan someterse a ingeniería inversa. Si enviamos los parámetros sin procesar, existe la posibilidad de que un adversario que los inspeccione pueda deducir si, por ejemplo, una determinada palabra había sido escuchada por el teléfono del usuario. Una forma de eliminar este riesgo es mediante la agregación segura (Bonawitz et al., 2017). La idea es que el servidor central no necesite conocer

el valor exacto del parámetro de cada usuario distribuido; sólo necesita saber el valor promedio de cada parámetro, entre todos los usuarios encuestados. De modo que cada usuario puede disfrazar los valores de sus parámetros agregando una máscara única a cada valor; Siempre que la suma de las máscaras sea cero, el servidor central podrá calcular el promedio correcto. Los detalles del protocolo garantizan que sea eficiente en términos de comunicación (menos de la mitad de los bits transmitidos corresponden a enmascaramiento), que sea robusto ante usuarios individuales que no responden y que sea seguro frente a usuarios adversarios, espías o incluso un servidor central adversario.

5.3. Equidad y parcialidad

El aprendizaje automático está aumentando y, a veces, reemplazando la toma de decisiones humana en situaciones importantes: qué préstamo se aprueba, en qué vecindarios se despliegan los agentes de policía, quién obtiene la libertad provisional o la libertad condicional. Pero los modelos de aprendizaje automático pueden perpetuar los prejuicios sociales. Consideremos el ejemplo de un algoritmo para predecir si es probable que los acusados reincidan y, por tanto, si deberían ser puestos en libertad antes del juicio. Bien podría ser que tal sistema recoja los prejuicios raciales o de género de los jueces humanos a partir de los ejemplos del conjunto de capacitación. Los diseñadores de sistemas de aprendizaje automático tienen la responsabilidad moral de garantizar que sus sistemas sean realmente justos. En ámbitos regulados como el crédito, la educación, el empleo y la vivienda, también tienen una responsabilidad legal. Pero ¿qué es la justicia? Hay múltiples criterios; Aquí hay seis de los conceptos más utilizados:

- **Justicia individual**: Requisito de que los individuos sean tratados de manera similar a otros individuos similares, independientemente de en qué clase se encuentren.
- **Equidad de grupo**: requisito de que dos clases sean tratadas de manera similar, según lo medido por alguna estadística resumida.
- **Equidad por desconocimiento**: si eliminamos los atributos de raza y género del conjunto de datos, entonces podría parecer que el sistema no puede discriminar esos atributos. Desafortunadamente, sabemos que los modelos de aprendizaje automático pueden predecir variables latentes (como la raza y el género). dadas otras variables correlacionadas (como el código postal y la ocupación). Además, eliminar esos atributos hace imposible verificar la igualdad de oportunidades o la igualdad de resultados. Aún así, algunos países (por ejemplo, Alemania) han elegido este enfoque para sus estadísticas demográficas (independientemente de si se utilizan modelos de aprendizaje automático o no).

- **Igual resultado**: La idea de que cada clase demográfica obtiene los mismos resultados; Tienen paridad demográfica. Por ejemplo, supongamos que tenemos que decidir si debemos aprobar las solicitudes de préstamo; el objetivo es aprobar a aquellos solicitantes que pagarán el préstamo y no a aquellos que no cumplirán con el mismo. La paridad demográfica dice que tanto hombres como mujeres deberían tener el mismo porcentaje de préstamos aprobados. Tenga en cuenta que este es un criterio de equidad grupal que no garantiza la equidad individual; un solicitante bien calificado podría ser rechazado y un solicitante mal calificado podría ser aprobado, siempre y cuando los porcentajes generales sean iguales. Además, este enfoque favorece la corrección de sesgos pasados por encima de la precisión de la predicción. Si un hombre y una mujer son iguales en todos los sentidos, excepto que la mujer recibe un salario más bajo por el mismo trabajo, ¿debería aprobarse porque sería igual si no fuera por sesgos históricos, o debería denegarse porque el salario más bajo no influye en ¿Esto la hace más propensa a incumplir?
- **Igualdad de oportunidades**: La idea de que las personas que realmente tienen la capacidad de pagar el préstamo deben tener las mismas posibilidades de ser correctamente clasificadas como tales, independientemente de su sexo. Este enfoque también se llama "equilibrio". Puede conducir a resultados desiguales e ignora el efecto del sesgo en los procesos sociales que produjeron los datos de capacitación.
- **Igual impacto**: Personas con similar probabilidad de pagar el préstamo deberían tener la misma utilidad esperada, independientemente de la clase a la que pertenezcan. Esto va más allá de la igualdad de oportunidades en el sentido de que considera tanto los beneficios de una predicción verdadera como los costos de una predicción falsa.

Examinemos cómo se desarrollan estas cuestiones en un contexto particular. COMPAS es un sistema comercial de puntuación de reincidencia (reincidencia). Asigna a un acusado en un caso penal una puntuación de riesgo, que luego el juez utiliza para ayudar a tomar decisiones: ¿Es seguro liberar al acusado antes del juicio o debería encarcelarlo? En caso de ser declarado culpable, ¿cuánto debería durar la sentencia? ¿Debería concederse la libertad condicional? Dada la importancia de estas decisiones, el sistema ha sido objeto de un intenso escrutinio (Dressel y Farid, 2018).

COMPAS está diseñado para estar bien calibrado: todos los individuos a los que el algoritmo les da la misma puntuación deben tener aproximadamente la misma probabilidad de reincidir, independientemente de su raza. Por ejemplo, entre todas las personas a las que el modelo asigna una puntuación de riesgo de 7 sobre 10, el 60% de los blancos y el 61% de los negros reinciden. Por tanto, los diseñadores afirman que cumple con el objetivo de equidad deseado.

Por otro lado, COMPAS no logra la igualdad de oportunidades: la proporción de aquellos que no reincidieron pero fueron calificados falsamente como de alto riesgo fue del 45% para los negros y del 23% para los blancos. En el caso Estado contra Loomis, donde un juez se basó en COMPAS para determinar la sentencia del acusado, Loomis argumentó que el secreto funcionamiento interno del algoritmo violaba sus derechos al debido proceso. Aunque la Corte Suprema de Wisconsin determinó que la sentencia dictada no sería diferente sin COMPAS en este caso, sí emitió advertencias sobre la precisión del algoritmo y los riesgos para los acusados minoritarios. Otros investigadores han cuestionado si es apropiado utilizar algoritmos en aplicaciones como la sentencia.

Podríamos esperar un algoritmo que esté bien calibrado y ofrezca igualdad de oportunidades, pero, como Kleinberg et al. (2016), eso es imposible. Si las clases base son diferentes, entonces cualquier algoritmo que esté bien calibrado no necesariamente brindará igualdad de oportunidades, y viceversa. ¿Cómo podemos sopesar los dos criterios? El mismo impacto es una posibilidad. En el caso de COMPAS, esto significa sopesar la utilidad negativa de que los acusados sean clasificados falsamente como de alto riesgo y pierdan su libertad, versus el costo para la sociedad de cometer un delito adicional, y encontrar el punto que optimice la compensación. Esto es complicado porque hay múltiples costos a considerar. Hay costos individuales: un acusado que es encarcelado injustamente sufre una pérdida, al igual que la víctima de un acusado que fue liberado injustamente y reincide. Pero más allá de eso, hay costos grupales: todos tienen cierto temor de ser encarcelados injustamente o de ser víctimas de un delito, y todos los contribuyentes contribuyen a los costos de las cárceles y los tribunales. Si valoramos esos temores y costos en proporción al tamaño de un grupo, entonces la utilidad para la mayoría puede llegar a expensas de una minoría.

Otro problema con la idea de la puntuación de reincidencia, independientemente del modelo utilizado, es que no disponemos de datos reales imparciales. Los datos no nos dicen quién ha cometido un delito; todo lo que sabemos es quién ha sido condenado por un delito. Si los agentes, el juez o el jurado que lo arrestan están sesgados, entonces los datos estarán sesgados. Si más agentes patrullan algunos lugares, los datos estarán sesgados en contra de las personas en esos lugares. Sólo los acusados que son liberados son candidatos a volver a ser condenados, por lo que si los jueces que toman las decisiones de liberación están sesgados, los datos pueden estar sesgados. Si se supone que detrás del conjunto de datos sesgados hay un conjunto de datos subyacente, desconocido e imparcial que ha sido corrompido por un agente con sesgos, entonces existen técnicas para recuperar una aproximación a los datos imparciales. Jiang y Nachum (2019) describen varios escenarios y las técnicas involucradas.

Un riesgo más es que el aprendizaje automático pueda utilizarse para justificar sesgos. Si las decisiones las toma un humano sesgado después de consultar con un sistema de aprendizaje automático, el humano puede decir "así es como mi interpretación del modelo respalda mi decisión, por lo que no deberías cuestionar mi decisión". Pero otras interpretaciones podrían conducir a una decisión contraria.

A veces, la justicia significa que debemos reconsiderar la función objetivo, no los datos o el algoritmo. Por ejemplo, al tomar decisiones de contratación laboral, si el objetivo es contratar candidatos con las mejores calificaciones, corremos el riesgo de recompensar injustamente a quienes han tenido oportunidades educativas ventajosas a lo largo de su vida, imponiendo así los límites de clase. Pero si el objetivo es contratar candidatos con la mejor capacidad para aprender en el trabajo, tenemos más posibilidades de trascender las fronteras de clase y elegir entre un grupo más amplio. Muchas empresas tienen programas diseñados para estos solicitantes y descubren que después de un año de capacitación, los empleados contratados de esta manera obtienen tan buenos resultados como los candidatos tradicionales. De manera similar, sólo el 18% de los graduados en ciencias de la computación en EE.UU. son mujeres, pero algunas escuelas, como la Universidad Harvey Mudd, han logrado una paridad del 50% con un enfoque que se centra en alentar y retener a quienes inician el programa de ciencias de la computación, especialmente aquellos que comienzan con menos experiencia en programación.

Una última complicación es decidir qué clases merecen protección. En Estados Unidos, la Ley de Vivienda Justa reconoció siete clases protegidas: raza, color, religión, origen nacional, sexo, discapacidad y situación familiar. Otras leyes locales, estatales y federales reconocen otras clases, incluida la orientación sexual y el embarazo, el estado civil y el estado de veterano. ¿Es justo que estas clases cuenten para algunas leyes y no para otras? El derecho internacional de los derechos humanos, que abarca un amplio conjunto de clases protegidas, es un marco potencial para armonizar las protecciones entre varios grupos.

Incluso en ausencia de sesgo social, la disparidad en el tamaño de la muestra puede generar resultados sesgados. En la mayoría de los conjuntos de datos habrá menos ejemplos de entrenamiento de individuos de clases minoritarias que de individuos de clases mayoritarias. Los algoritmos de aprendizaje automático brindan mayor precisión con más datos de entrenamiento, lo que significa que los miembros de clases minoritarias experimentarán una menor precisión. Por ejemplo, Buolamwini y Gebru (2018) examinaron un servicio de identificación de género por visión por computadora y descubrieron que tenía una precisión casi perfecta para los hombres de piel clara y una tasa de error del 33 % para las mujeres de piel oscura. Es posible que un modelo restringido no pueda ajustarse simultáneamente a la clase mayoritaria y minoritaria; un modelo de regresión lineal podría minimizar el error promedio ajustando solo la clase mayoritaria, y en un modelo SVM, todos los vectores de soporte podrían corresponder a miembros de la clase mayoritaria.

El sesgo también puede entrar en juego en el proceso de desarrollo de software (ya sea que el software implique aprendizaje automático o no). Es más probable que los ingenieros que depuran un sistema noten y solucionen los problemas que les son aplicables. Por ejemplo, es difícil darse cuenta de que el diseño de una interfaz de usuario no funcionará para personas daltónicas a menos que usted sea daltónico, o que una traducción al idioma urdu sea defectuosa si no habla urdu.

¿Cómo podemos defendernos de estos prejuicios? Primero, comprenda los límites de los datos que está utilizando. Se ha sugerido que los conjuntos de datos (Gebru et al., 2018; Hind et al., 2018) y los modelos (Mitchell et al., 2019) deberían venir con anotaciones: declaraciones de procedencia, seguridad, conformidad y aptitud para el uso. Esto es similar a las hojas de datos que acompañan a los componentes electrónicos como las resistencias; permiten a los diseñadores decidir qué componentes utilizar. Además de las hojas de datos, es importante capacitar a los ingenieros para que sean conscientes de las cuestiones de equidad y parcialidad, tanto en la escuela como en la capacitación en el trabajo. Tener una diversidad de ingenieros de diferentes orígenes les facilita notar problemas en los datos o modelos. Un estudio del AI Now Institute (West et al., 2019) encontró que solo el 18% de los autores de las principales conferencias sobre IA y el 20% de los profesores de IA son mujeres. Los trabajadores negros de IA representan menos del 4%. Las tarifas en los laboratorios de investigación de la industria son similares. La diversidad podría aumentar mediante programas en etapas más tempranas (en la universidad o la escuela secundaria) y mediante una mayor conciencia a nivel profesional. Joy Buolamwini fundó la Liga de Justicia Algorítmica para crear conciencia sobre este tema y desarrollar prácticas de rendición de cuentas.

Una segunda idea es eliminar el sesgo de los datos (Zemel et al., 2013). Podríamos sobremuestrear de clases minoritarias para defendernos de la disparidad en el tamaño de la muestra. Técnicas como SMOTE, la técnica de sobremuestreo minoritario sintético (Chawla et al., 2002) o A DASYN, el enfoque de muestreo sintético adaptativo para el aprendizaje desequilibrado (He et al., 2008), proporcionan formas de sobremuestreo basadas en principios. Podríamos examinar la procedencia de los datos y, por ejemplo, eliminar ejemplos de jueces que hayan mostrado parcialidad en sus casos judiciales anteriores. Algunos analistas se oponen a la idea de descartar datos y, en cambio, recomendarían construir un modelo jerárquico de los datos que incluya fuentes de sesgo, para que puedan modelarse y compensarse. Google y NeurIPS han intentado crear conciencia sobre este problema patrocinando el Concurso de Imágenes Inclusivas, en el que los competidores entrenan una red con un conjunto de datos de imágenes etiquetadas recopiladas en América del Norte y Europa, y luego la prueban con imágenes tomadas de todo el mundo. . El problema es que, dado este conjunto de datos, es fácil aplicar la etiqueta “novia” a una mujer con un vestido de novia occidental estándar, pero es más difícil reconocer la vestimenta matrimonial tradicional africana e india.

Una tercera idea es inventar nuevos modelos y algoritmos de aprendizaje automático que sean más resistentes al sesgo; y la idea final es dejar que un sistema haga recomendaciones iniciales que puedan estar sesgadas, pero luego entrenar a un segundo sistema para que elimine el sesgo de las recomendaciones del primero. Bellamy et al. (2018) introdujeron el sistema IBM AI FAIRNESS 360, que proporciona un marco para todas estas ideas. Esperamos que haya un mayor uso de herramientas como esta en el futuro. ¿Cómo se asegura de que los sistemas que construya sean justos? Ha ido surgiendo un conjunto de mejores prácticas (aunque no siempre se siguen):

- Asegúrese de que los ingenieros de software hablen con científicos sociales y expertos en el campo para comprender los problemas y las perspectivas, y considerar la equidad desde el principio.
- Crear un entorno que fomente el desarrollo de un grupo diverso de ingenieros de software que sean representativos de la sociedad.
- Defina qué grupos admitirá su sistema: diferentes hablantes de idiomas, diferentes grupos de edad, diferentes habilidades visuales y auditivas, etc.
- Optimizar para una función objetivo que incorpore equidad.
- Examine sus datos en busca de prejuicios y correlaciones entre atributos protegidos y otros atributos.
- Comprender cómo se realiza cualquier anotación humana de datos, diseñar objetivos para la precisión de la anotación y verificar que se cumplan los objetivos.
- No se limite a realizar un seguimiento de las métricas generales de su sistema; asegúrese de realizar un seguimiento de las métricas de los subgrupos que podrían ser víctimas de prejuicios.
- Incluir pruebas del sistema que reflejen la experiencia de usuarios de grupos minoritarios.
- Tener un circuito de retroalimentación para que cuando surjan problemas de equidad, se resuelvan.

5.4. Confianza y transparencia

Uno de los desafíos es lograr que un sistema de IA sea preciso, justo y seguro; un desafío diferente para convencer a todos los demás de que lo has hecho. Las personas deben poder confiar en los sistemas que utilizan. Una encuesta de PwC realizada en 2017 encontró que el 76% de las empresas estaban desacelerando la adopción de la IA debido a preocupaciones sobre la confiabilidad.

Para ganarse la confianza, cualquier sistema diseñado debe pasar por un proceso de verificación y validación (V&V). Verificación significa que el producto cumple con las especificaciones. Validación significa garantizar que las especificaciones realmente satisfagan las necesidades del usuario y de otras partes afectadas. Contamos con una elaborada metodología V&V para la ingeniería en general, y para el desarrollo de software tradicional realizado por codificadores humanos; Gran parte de eso es aplicable a los sistemas de IA. Pero los sistemas de aprendizaje automático son diferentes y exigen un proceso de V&V diferente, que aún no se ha desarrollado por completo. Necesitamos verificar los datos de los que aprenden estos sistemas; necesitamos verificar la exactitud y equidad de los resultados, incluso ante la incertidumbre que hace que un resultado exacto sea incognoscible; y necesitamos verificar que los adversarios no puedan influir indebidamente en el modelo, ni robar información consultando el modelo resultante.

Un instrumento de confianza es la certificación; por ejemplo, Underwriters Laboratories (UL) se fundó en 1894 en una época en la que los consumidores estaban preocupados por los riesgos de la energía eléctrica. La certificación UL de electrodomésticos dio a los consumidores una mayor confianza y, de hecho, UL ahora está considerando ingresar al negocio de pruebas y certificación de productos para IA.

Otras industrias cuentan desde hace mucho tiempo con estándares de seguridad. Por ejemplo, ISO 26262 es una norma internacional para la seguridad de los automóviles que describe cómo desarrollar, producir, operar y dar servicio a los vehículos de manera segura. La industria de la IA aún no alcanza este nivel de claridad, aunque hay algunos marcos en progreso, como IEEE P7001, un estándar que define el diseño ético para la inteligencia artificial y los sistemas autónomos (Bryson y Winfield, 2017). Existe un debate en curso sobre qué tipo de certificación es necesaria y en qué medida debería ser realizada por el gobierno, por organizaciones profesionales como IEEE, por certificadores independientes como UL o mediante la autorregulación por parte de las empresas de productos.

Otro aspecto de la confianza es la transparencia: los consumidores quieren saber qué sucede dentro de un sistema y que el sistema no está actuando en su contra, ya sea por malicia intencional, un error involuntario o un sesgo social generalizado que el sistema recapitula. En algunos casos, esta transparencia se entrega directamente al consumidor. En otros casos, se trata de cuestiones de propiedad intelectual que mantienen algunos aspectos del sistema ocultos para los consumidores, pero abiertos a los reguladores y agencias de certificación.

Cuando un sistema de inteligencia artificial le rechaza un préstamo, merece una explicación. En Europa, el RGPD lo hace cumplir. Un sistema de IA que puede explicarse a sí mismo se llama IA explicable (XAI). Una buena explicación tiene varias propiedades: debe ser comprensible y convincente para el usuario, debe reflejar con precisión el razonamiento del sistema, debe ser completa y debe ser específica en el sentido de que diferentes usuarios con diferentes condiciones o diferentes resultados deberían obtener diferentes resultados. explicaciones.

Es bastante fácil dar acceso a un algoritmo de decisión a sus propios procesos deliberativos, simplemente registrándolos y poniéndolos a disposición como estructuras de datos. Esto significa que las máquinas eventualmente podrán dar mejores explicaciones de sus decisiones que los humanos. Es más, podemos tomar medidas para certificar que las explicaciones de la máquina no son engaños (intencionados o autoengaños), algo que es más difícil con un humano.

Una explicación es un ingrediente útil pero no suficiente para confiar. Una cuestión es que las explicaciones no son decisiones: son historias sobre decisiones. Decimos que un sistema es interpretable si podemos inspeccionar el código fuente del modelo y ver qué está haciendo, y decimos que es explicable si podemos inventar una historia sobre lo que está haciendo. — Incluso si el sistema en sí es una caja negra ininterpretable. Para explicar una caja negra no interpretable, necesitamos construir, depurar y probar un sistema de explicación separado y asegurarnos de que esté sincronizado con el sistema original. Y como a los humanos les encantan las buenas historias, todos estamos dispuestos a dejarnos llevar por una explicación que suene bien. Tomemos cualquier controversia política del día y siempre podremos encontrar dos supuestos expertos con explicaciones diametralmente opuestas, las cuales son internamente consistentes.

Una última cuestión es que una explicación sobre un caso no proporciona un resumen de otros casos. Si el banco explica: "Lo siento, no obtuvo el préstamo porque tiene un historial de problemas financieros previos", no sabe si esa explicación es correcta o si el banco tiene secretamente parcialidad en su contra por alguna razón. En este caso, no sólo se necesita una explicación, sino también una auditoría de las decisiones pasadas, con estadísticas agregadas de varios grupos demográficos, para ver si sus tasas de aprobación están equilibradas.

Parte de la transparencia es saber si estás interactuando con un sistema de inteligencia artificial o con un ser humano. Toby Walsh (2015) propuso que “un sistema autónomo debería diseñarse de manera que sea improbable que se confunda con algo más que un sistema autónomo, y debería identificarse al inicio de cualquier interacción” . Llamó a esto la ley de “bandera roja” , en honor a la Ley de Locomotoras de 1865 del Reino Unido, que exigía que cualquier vehículo motorizado tuviera una persona con una bandera roja caminando delante de él, para señalar el peligro que se avecinaba.

En 2019, California promulgó una ley que establece que "Será ilegal que cualquier persona utilice un bot para comunicarse o interactuar con otra persona en California en línea, con la intención de engañar a la otra persona sobre su identidad artificial".

5.5. El futuro del trabajo

Desde la primera revolución agrícola (10.000 a. C.) hasta la revolución industrial (finales del siglo XVIII) y la revolución verde en la producción de alimentos (década de 1950), las nuevas tecnologías han cambiado la forma en que la humanidad trabaja y vive. Una de las principales preocupaciones que surge del avance de la IA es que el trabajo humano quedará obsoleto. Aristóteles, en el Libro I de su Política, presenta el punto principal con bastante claridad:

> **Importante** Porque si cada instrumento pudiera realizar su propio trabajo, obedeciendo o anticipando la voluntad de los demás... si, de la misma manera, la lanzadera tejiera y la púa tocara la lira sin una mano que los guiara, los jefes de los trabajadores lo harían. No faltan sirvientes, ni amos esclavos.

Todo el mundo está de acuerdo con la observación de Aristóteles de que hay una reducción inmediata del empleo cuando un empleador encuentra un método mecánico para realizar un trabajo previamente realizado por una persona. La cuestión es si los llamados efectos de compensación que se derivan –y que tienden a aumentar el empleo– eventualmente compensarán esta reducción. El principal efecto de compensación es el aumento de la riqueza general debido a una mayor productividad, lo que a su vez conduce a una mayor demanda de bienes y tiende a aumentar el empleo. Por ejemplo, PwC (Rao y Verweij, 2017) predice que la IA contribuirá con 15 billones de dólares anuales al PIB mundial para 2030. Las industrias de la salud y la automoción/transporte serán las que más ganarán en el corto plazo. Sin embargo, las ventajas de la automatización aún no se han apoderado de nuestra economía: la tasa actual de crecimiento de la productividad laboral está en realidad por debajo de los estándares históricos. Brynjolfsson et al. (2018) intentan explicar esta paradoja sugiriendo que el desfase entre el desarrollo de la tecnología básica y su implementación en la economía es más largo de lo que comúnmente se supone.

Históricamente, las innovaciones tecnológicas han dejado a algunas personas sin trabajo. Los tejedores fueron reemplazados por telares automáticos en la década de 1810, lo que provocó las protestas luditas. Los luditas no estaban en contra de la tecnología per se; sólo querían que las máquinas fueran utilizadas por trabajadores calificados a los que se les pagaba un buen salario para fabricar productos de alta calidad, en lugar de trabajadores no calificados para fabricar productos de mala calidad con salarios bajos. La destrucción global de empleos en la década de 1930 llevó a John Maynard Keynes a acuñar el término desempleo tecnológico. En ambos casos, y en varios otros, los niveles de empleo finalmente se recuperaron.

La visión económica predominante durante la mayor parte del siglo XX fue que el empleo tecnológico era, como mucho, un fenómeno de corto plazo. Una mayor productividad siempre conduciría a un aumento de la riqueza y de la demanda y, por tanto, a un crecimiento neto del empleo. Un ejemplo comúnmente citado es el de los cajeros de los bancos: aunque los cajeros automáticos reemplazaron a los humanos en la tarea de contar el efectivo para los retiros, eso hizo que fuera más barato operar una sucursal bancaria, por lo que el número de sucursales aumentó, lo que generó más empleados bancarios en general. La naturaleza del trabajo

también cambió, volviéndose menos rutinario y requiriendo habilidades comerciales más avanzadas. El efecto neto de la automatización parece ser la eliminación de tareas en lugar de puestos de trabajo.

La mayoría de los comentaristas predicen que lo mismo ocurrirá con la tecnología de inteligencia artificial, al menos a corto plazo. Gartner, McKinsey, Forbes, el Foro Económico Mundial y el Pew Research Center publicaron informes en 2018 que predicen un aumento neto de empleos debido a la automatización impulsada por la IA. Pero algunos analistas creen que esta vez las cosas serán diferentes. En 2019, IBM predijo que 120 millones de trabajadores necesitarían volver a capacitarse debido a la automatización para 2022, y Oxford Economics predijo que 20 millones de empleos en el sector manufacturero podrían perderse debido a la automatización para 2030.

Frey y Osborne (2017) encuestaron 702 ocupaciones diferentes y estimaron que el 47% de ellas corren el riesgo de ser automatizadas, lo que significa que al menos algunas de las tareas de la ocupación pueden realizarse mediante una máquina. Por ejemplo, casi el 3% de la fuerza laboral en Estados Unidos son conductores de vehículos y, en algunos distritos, hasta el 15% de la fuerza laboral masculina son conductores. Como vimos en el capítulo 26, es probable que la tarea de conducir quede eliminada con los coches, camiones, autobuses y taxis sin conductor.

Es importante distinguir entre ocupaciones y las tareas dentro de esas ocupaciones. McKinsey estima que sólo el 5% de las ocupaciones son completamente automatizables, pero que el 60% de las ocupaciones pueden automatizar alrededor del 30% de sus tareas. Por ejemplo, los futuros camioneros pasarán menos tiempo sujetando el volante y más tiempo asegurándose de que la mercancía se recoja y entregue correctamente; actuar como representantes de servicio al cliente y vendedores en ambos extremos del proceso; y quizás gestionar convoyes de, digamos, tres camiones robóticos. Reemplazar a tres conductores por un jefe de convoy implica una pérdida neta de empleo, pero si los costos de transporte disminuyen, habrá más demanda, lo que recuperará algunos de los empleos, pero tal vez no todos. Como otro ejemplo, a pesar de muchos avances en la aplicación del aprendizaje automático al problema de las imágenes médicas, hasta ahora los radiólogos han sido aumentados, no reemplazados, por estas herramientas. En última instancia, hay que elegir cómo hacer uso de la automatización: ¿queremos centrarnos en reducir costes y, por tanto, ver la pérdida de empleo como algo positivo? ¿O queremos centrarnos en mejorar la calidad, mejorar la vida del trabajador y del cliente?

Es difícil predecir cronogramas exactos para la automatización, pero actualmente, y durante los próximos años, el énfasis está en la automatización de tareas analíticas estructuradas, como la lectura de imágenes de rayos X, la gestión de relaciones con los clientes (por ejemplo, robots que clasifican automáticamente las quejas de los clientes). y responder con soluciones sugeridas), y automatización de procesos de negocio que combina documentos de texto y datos estructurados para tomar decisiones de negocio y mejorar el flujo de trabajo. Con el tiempo, veremos una mayor automatización con robots físicos, primero en entornos de almacén controlados y luego en entornos más inciertos, lo que representará una parte importante del mercado alrededor de 2030.

A medida que las poblaciones de los países desarrollados envejecen, la proporción entre trabajadores y jubilados cambia. En 2015 había menos de 30 jubilados por cada 100 trabajadores; para 2050 puede haber más de 60 por cada 100 trabajadores. El cuidado de las personas mayores será una función cada vez más importante, que puede ser desempeñada parcialmente por la IA. Además, si queremos mantener el nivel de vida actual, también será necesario hacer que los trabajadores restantes sean más productivos; la automatización parece la mejor oportunidad para hacerlo.

Incluso si la automatización tiene un impacto positivo neto multimillonario, todavía puede haber problemas debido al ritmo del cambio. Consideremos cómo se produjo el cambio en la industria agrícola: en 1900, más del 40% de la fuerza laboral estadounidense se dedicaba a la agricultura, pero en 2000 esa cifra había caído al 2%.3 Se trata de una enorme alteración en la forma en que trabajamos, pero ocurrió a lo largo de un siglo. período de 100 años y, por tanto, a lo largo de generaciones, no durante la vida de un trabajador.

Los trabajadores cuyos empleos se han automatizado en esta década tal vez tengan que volver a capacitarse para una nueva profesión dentro de unos pocos años, y luego tal vez ver su nueva profesión automatizada y enfrentar otro período de recapacitación. Algunos pueden estar felices de dejar su antigua profesión (vemos que a medida que la economía mejora, las empresas de transporte deben ofrecer nuevos incentivos para contratar suficientes conductores), pero los trabajadores estarán preocupados por sus nuevos roles. Para manejar esto, nosotros, como sociedad, debemos brindar educación permanente, quizás confiando en parte en la educación en línea impulsada por inteligencia artificial (Martin, 2012). Bessen (2015) sostiene que los trabajadores no verán aumentos en sus ingresos hasta que estén capacitados para implementar las nuevas tecnologías, un proceso que lleva tiempo.

La tecnología tiende a magnificar la desigualdad de ingresos. En una economía de la información caracterizada por una comunicación global de gran ancho de banda y una replicación de la propiedad intelectual con costo marginal cero (lo que Frank y Cook (1996) llaman la “sociedad en la que el ganador se lo lleva todo”), las recompensas tienden a concentrarse. Si el agricultor Ali es un 10% mejor que el agricultor Bo, entonces Ali obtiene alrededor de un 10% más de ingresos: Ali puede cobrar un poco más por bienes superiores, pero hay un límite sobre cuánto se puede producir en la tierra y hasta dónde se puede llegar. enviado. Pero si el desarrollador de aplicaciones de software Cary es un 10% mejor que Dana, es posible que Cary termine con el 99% del mercado global. La IA aumenta el ritmo de la innovación tecnológica y, por lo tanto, contribuye a esta tendencia general, pero la IA también promete permitirnos tomarnos un tiempo libre y dejar que nuestros agentes automatizados se encarguen de las cosas por un tiempo. Tim Ferriss (2007) recomienda utilizar la automatización y la subcontratación para lograr una semana laboral de cuatro horas.

Antes de la revolución industrial, la gente trabajaba como agricultores o en otros oficios, pero no se presentaba a un trabajo en un lugar de trabajo ni trabajaba horas para un empleador. Pero hoy en día, la mayoría de los adultos en los países desarrollados hacen precisamente eso, y el trabajo tiene tres propósitos: impulsa la producción de los bienes que la sociedad necesita para prosperar, proporciona el ingreso que el trabajador necesita para vivir y le da al trabajador una sensación de bienestar. de propósito, logro e integración social. Con la creciente automatización, es posible que estos tres propósitos se desagregen: las necesidades de la sociedad serán atendidas en parte por la automatización y, a largo plazo, los individuos

obtendrán su sentido de propósito a partir de contribuciones distintas al trabajo. Sus necesidades de ingresos pueden satisfacerse mediante políticas sociales que incluyan una combinación de acceso gratuito o económico a servicios sociales y educación, cuentas portátiles de atención médica, jubilación y educación, tasas impositivas progresivas, créditos fiscales por ingresos del trabajo, impuestos negativos sobre la renta o servicios básicos universales. ingreso.

5.6. Derechos de los robots

La cuestión de la conciencia robótica, es fundamental para la cuestión de qué derechos, si los hubiera, deberían tener los robots. Si no tienen conciencia, ni qualia, entonces pocos dirían que merecen derechos.

Pero si los robots pueden sentir dolor, si pueden temer la muerte, si se les considera "personas" , entonces se puede argumentar (por ejemplo, Sparrow (2004)) que tienen derechos y merecen que se les reconozcan, al igual que Los esclavos, las mujeres y otros grupos históricamente oprimidos han luchado para que se reconozcan sus derechos. La cuestión de la personalidad del robot a menudo se considera en la ficción: desde Pigmalión hasta Coppélia, Pinocho, las películas AI y Centennial Man, tenemos la leyenda de un muñeco/robot que cobra vida y se esfuerza por ser aceptado como un ser humano con derechos humanos. En la vida real, Arabia Saudita fue noticia al otorgar la ciudadanía honoraria a Sophia, una marioneta de apariencia humana capaz de pronunciar líneas preprogramadas.

Si los robots tienen derechos, entonces no deberían ser esclavizados, y cabe preguntarse si reprogramarlos sería una especie de esclavitud. Otra cuestión ética tiene que ver con los derechos de voto: una persona rica podría comprar miles de robots y programarlos para emitir miles de votos. ¿Deberían contar esos votos? Si un robot se clona, ¿pueden votar ambos? ¿Cuál es el límite entre el relleno de votos y el ejercicio del libre albedrío, y cuándo la votación robótica viola el principio de "una persona, un voto" ?

Ernie Davis aboga por evitar los dilemas de la conciencia robótica al no construir nunca robots que puedan considerarse conscientes. Este argumento fue expuesto previamente por Joseph Weizenbaum en su libro Computer Power and Human Reason (1976), y antes por Julien de La Mettrie en L'Homme Machine (1748). Los robots son herramientas que creamos para realizar las tareas que les ordenamos, y si les concedemos personalidad, simplemente nos negamos a asumir la responsabilidad de las acciones de nuestra propia propiedad: "No tengo la culpa de mi auto- conducir un accidente automovilístico: el auto lo hizo solo" .

Esta cuestión toma un cariz diferente si desarrollamos híbridos entre humanos y robots. Por supuesto, ya tenemos seres humanos mejorados gracias a tecnologías como lentes de contacto, marcapasos y caderas artificiales. Pero añadir prótesis computacionales puede desdibujar la línea entre humanos y máquinas.

5.7. Seguridad de la IA

Casi cualquier tecnología tiene el potencial de causar daño en las manos equivocadas, pero con la inteligencia artificial y la robótica, las manos podrían funcionar por sí solas. Innumerables historias de ciencia ficción han advertido sobre robots o cyborgs enloquecidos. Los primeros ejemplos incluyen Frankenstein o el Prometeo moderno (1818) de Mary Shelley y la obra de teatro RUR (1920) de Karel Čapek, en la que los robots conquistan el mundo. En las películas, tenemos Terminator (1984) y Matrix (1999), en las que aparecen robots que intentan eliminar a los humanos: el robopocalipsis (Wilson, 2011). Quizás los robots sean a menudo los villanos porque representan lo desconocido, al igual que las brujas y los fantasmas de los cuentos de épocas anteriores. Podemos esperar que un robot que sea lo suficientemente inteligente como para descubrir cómo acabar con la raza humana también sea lo suficientemente inteligente como para darse cuenta de que esa no era la función de utilidad prevista; pero a la hora de construir sistemas inteligentes queremos confiar no sólo en la esperanza, sino en un proceso de diseño con garantías de seguridad.

No sería ético distribuir un agente de IA inseguro. Exigimos que nuestros agentes eviten accidentes, sean resistentes a ataques adversarios y abusos maliciosos y, en general, causen beneficios, no daños. Esto es especialmente cierto cuando los agentes de IA se implementan en aplicaciones críticas para la seguridad, como conducir automóviles, controlar robots en fábricas o entornos de construcción peligrosos y tomar decisiones médicas de vida o muerte.

Existe una larga historia de ingeniería de seguridad en los campos de la ingeniería tradicional. Sabemos cómo construir puentes, aviones, naves espaciales y centrales eléctricas diseñadas desde el principio para comportarse de forma segura incluso cuando fallan los componentes del sistema. La primera técnica es el análisis de modos de falla y efectos (FMEA): los analistas consideran cada componente del sistema e imaginan todas las formas posibles en que el componente podría fallar (por ejemplo, ¿qué pasaría si este perno se rompiera?), basándose en experiencias pasadas y en cálculos basados en las propiedades físicas del componente. Luego, los analistas trabajan para ver qué resultaría del fracaso. Si el resultado es grave (una sección del puente podría caer), los analistas modifican el diseño para mitigar la falla. (Con este travesaño adicional, el puente puede sobrevivir a la falla de 5 pernos cualesquiera; con este servidor de respaldo, el servicio en línea puede sobrevivir a un tsunami que destruya el servidor primario). La técnica del análisis de árbol de fallas (FTA) se utiliza para Haga estas determinaciones: los analistas construyen un árbol Y/O de posibles fallas y asignan probabilidades a cada causa raíz, lo que permite calcular la probabilidad general de falla. Estas técnicas pueden y deben aplicarse a todos los sistemas de ingeniería críticos para la seguridad, incluidos los sistemas de IA.

El campo de la ingeniería de software tiene como objetivo producir software confiable, pero históricamente el énfasis ha estado en la corrección, no en la seguridad. Corrección significa que el software implementa fielmente la especificación. Pero la seguridad va más allá e insiste en que la especificación ha considerado cualquier modo de falla factible y está diseñada para degradarse con gracia incluso ante fallas imprevistas. Por ejemplo, el software para un vehículo autónomo no se consideraría seguro a menos que pueda manejar situaciones inusuales. Por ejemplo, ¿qué pasa si se corta la energía de la computadora principal? Un sistema seguro tendrá una computadora de respaldo con una fuente de alimentación separada.

¿Qué pasa si se pincha un neumático a alta velocidad? Un sistema seguro habrá probado esto y tendrá un software para corregir la pérdida de control resultante.

Un agente diseñado como maximizador de utilidad, o para alcanzar metas, puede ser inseguro si tiene la función objetivo incorrecta. Supongamos que le damos a un robot la tarea de ir a buscar un café a la cocina. Podríamos tener problemas con efectos secundarios no deseados: el robot podría apresurarse a lograr el objetivo y derribar lámparas y mesas en el camino. Durante las pruebas, podemos notar este tipo de comportamiento y modificar la función de utilidad para penalizar dicho daño, pero es difícil para los diseñadores y evaluadores anticipar todos los posibles efectos secundarios de antemano.

Una forma de abordar esto es diseñar un robot que tenga un impacto bajo (Armstrong y Levinstein, 2017): en lugar de simplemente maximizar la utilidad, maximice la utilidad menos un resumen ponderado de todos los cambios en el estado del mundo. De esta manera, en igualdad de condiciones, el robot prefiere no cambiar aquellas cosas cuyo efecto sobre la utilidad se desconoce; por lo tanto, evita derribar la lámpara, no porque sepa específicamente que hacerlo hará que se caiga y se rompa, sino porque sabe en general que la interrupción podría ser mala. Esto puede verse como una versión del credo médico "primero, no hacer daño", o como un análogo de la regularización en el aprendizaje automático: queremos una política que logre objetivos, pero preferimos políticas que tomen acciones fluidas y de bajo impacto para ir allí. El truco es cómo medir el impacto. No es aceptable derribar una lámpara frágil, pero está bien si las moléculas de aire de la habitación se alteran un poco o si algunas bacterias de la habitación mueren sin darse cuenta. Ciertamente no es aceptable dañar a las mascotas ni a las personas en la habitación. Necesitamos asegurarnos de que el robot conozca las diferencias entre estos casos (y muchos casos sutiles intermedios) mediante una combinación de programación explícita, aprendizaje automático a lo largo del tiempo y pruebas rigurosas.

Las funciones de utilidad pueden fallar debido a externalidades, la palabra utilizada por los economistas para referirse a factores que están fuera de lo que se mide y se paga. El mundo sufre cuando los gases de efecto invernadero se consideran externalidades: las empresas y los países no son penalizados por producirlos y, como resultado, todos sufren. El ecologista Garrett Hardin (1968) llamó a la explotación de los recursos compartidos la tragedia de los comunes. Podemos mitigar la tragedia internalizando las externalidades (haciéndolas parte de la función de utilidad, por ejemplo con un impuesto al carbono) o utilizando los principios de diseño que la economista Elinor Ostrom identificó como utilizados por la población local en todo el mundo durante siglos (trabajo que le valió el Premio Nobel de Economía en 2009):

- Definir claramente el recurso compartido y quién tiene acceso.
- Adaptarse a las condiciones locales.
- Permitir que todas las partes participen en las decisiones.
- Monitorear el recurso con monitores responsables.
- Sanciones, proporcionales a la gravedad de la infracción.
- Procedimientos sencillos de resolución de conflictos.
- Control jerárquico para grandes recursos compartidos.

Victoria Krakovna (2018) ha catalogado ejemplos de agentes de IA que han jugado con el sistema, descubriendo cómo maximizar la utilidad sin resolver realmente el problema que sus diseñadores pretendían que resolvieran. Para los diseñadores esto parece una trampa, pero para los agentes simplemente están haciendo su trabajo. Algunos agentes aprovecharon errores en la simulación (como errores de desbordamiento de punto flotante) para proponer soluciones que no funcionarían una vez que se solucionara el error. Varios agentes en los videojuegos descubrieron formas de bloquear o pausar el juego cuando estaban a punto de perder, evitando así una penalización. Y en una especificación donde se penalizaba bloquear el juego, un agente aprendió a usar suficiente memoria del juego para que cuando fuera el turno del oponente, se quedara sin memoria y bloqueara el juego. Finalmente, se suponía que un algoritmo genético que operaba en un mundo simulado evolucionaría en criaturas que se movían rápidamente, pero en realidad produjo criaturas que eran enormemente altas y se movían rápidamente al caer.

Los diseñadores de agentes deben ser conscientes de este tipo de errores en las especificaciones y tomar medidas para evitarlos. Para ayudarlos a lograrlo, Krakovna formó parte del equipo que lanzó los entornos AI Safety Gridworlds (Leike et al., 2017), que permite a los diseñadores probar qué tan bien se desempeñan sus agentes.

La moraleja es que debemos tener mucho cuidado al especificar lo que queremos, porque con los maximizadores de utilidad obtenemos lo que realmente pedimos. El problema de la alineación de valores es el problema de asegurarnos de que lo que pedimos es lo que realmente queremos; también se le conoce como el problema del Rey Midas, como se analiza en la página 33. Nos encontramos con problemas cuando una función de utilidad no logra capturar las normas sociales subyacentes sobre el comportamiento aceptable. Por ejemplo, un humano que es contratado para limpiar pisos, cuando se enfrenta a una persona desordenada que repetidamente deja huellas en la tierra, sabe que es aceptable pedirle cortésmente a la persona que tenga más cuidado, pero no es aceptable secuestrar o incapacitar a dicha persona.
.

Un robot limpiador también necesita saber estas cosas, ya sea mediante programación explícita o aprendiendo de la observación. Intentar escribir todas las reglas para que el robot siempre haga lo correcto es casi seguro que es inútil. Llevamos varios miles de años intentando redactar leyes fiscales libres de lagunas, sin éxito. Es mejor hacer que el robot quiera pagar impuestos, por así decirlo, que tratar de establecer reglas para obligarlo a hacerlo cuando realmente quiere hacer otra cosa. Un robot suficientemente inteligente encontrará una manera de hacer otra cosa.

Los robots pueden aprender a adaptarse mejor a las preferencias humanas observando el comportamiento humano. Esto está claramente relacionado con la noción de aprendizaje mediante aprendizaje. El robot puede aprender una política que sugiera directamente qué acciones tomar en qué situaciones; Este suele ser un problema sencillo de aprendizaje supervisado si el entorno es observable. Por ejemplo, un robot puede observar a un humano jugando al ajedrez: cada par estado-acción es un ejemplo del proceso de aprendizaje. Desafortunadamente, esta forma de aprendizaje por imitación significa que el robot repetirá los errores humanos. En cambio, el robot puede aplicar el aprendizaje por refuerzo inverso para descubrir la función de utilidad bajo la cual deben operar los humanos. Probablemente basta con observar incluso a jugadores de ajedrez terribles para que el robot aprenda el

objetivo del juego. Con solo esta información, el robot puede superar el desempeño humano (como lo hizo, por ejemplo, ALPHAZERO en el ajedrez) calculando políticas óptimas o casi óptimas a partir del objetivo. Este enfoque funciona no sólo en juegos de mesa, sino también en tareas físicas del mundo real, como las acrobacias aéreas en helicóptero (Coates et al., 2009).

En entornos más complejos que implican, por ejemplo, interacciones sociales con humanos, es muy poco probable que el robot converja para obtener un conocimiento exacto y correcto de las preferencias individuales de cada ser humano. (Después de todo, muchos humanos nunca aprenden qué es lo que motiva a otros humanos, a pesar de toda una vida de experiencia, y muchos de nosotros tampoco estamos seguros de nuestras propias preferencias). Por lo tanto, será necesario que las máquinas funcionen apropiadamente cuando no esté seguro. sobre las preferencias humanas. En el Capítulo 18, presentamos juegos de asistencia que capturan exactamente esta situación. Las soluciones a los juegos de asistencia incluyen actuar con cautela, para no perturbar aspectos del mundo que podrían interesarle al ser humano, y hacer preguntas. Por ejemplo, el robot podría preguntar si convertir los océanos en ácido sulfúrico es una solución aceptable al calentamiento global antes de poner en práctica el plan.

Al tratar con humanos, un robot que resuelve un juego de asistencia debe adaptarse a las imperfecciones humanas. Si el robot pide permiso, el humano puede concedérselo, sin prever que la propuesta del robot sea catastrófica a largo plazo. Además, los humanos no tienen un acceso introspectivo completo a su verdadera función de utilidad y no siempre actúan de una manera que sea compatible con ella. Los seres humanos a veces mienten o engañan, o hacen cosas que saben que están mal. A veces toman acciones autodestructivas como comer en exceso o abusar de las drogas. Los sistemas de IA no necesitan aprender a adoptar estas tendencias problemáticas, pero deben comprender que existen al interpretar el comportamiento humano para llegar a las preferencias humanas subyacentes.

A pesar de esta caja de herramientas de salvaguardias, existe el temor, expresado por destacados tecnólogos como Bill Gates y Elon Musk y científicos como Stephen Hawking y Martin Rees, de que la IA pueda evolucionar fuera de control. Advierten que no tenemos experiencia en controlar poderosas entidades no humanas con capacidades sobrehumanas. Sin embargo, eso no es del todo cierto; tenemos siglos de experiencia con naciones y corporaciones; entidades no humanas que agregan el poder de miles o millones de personas. Nuestro historial de control de estas entidades no es muy alentador: las naciones producen convulsiones periódicas llamadas guerras que matan a decenas de millones de seres humanos, y las corporaciones son en parte responsables del calentamiento global y de nuestra incapacidad para enfrentarlo.

Los sistemas de IA pueden presentar problemas mucho mayores que los de las naciones y las corporaciones debido a su potencial para automejorarse a un ritmo rápido, como lo considera IJ Good (1965b):

Importante Definamos una máquina ultrainteligente como una máquina que puede superar con creces todas las actividades intelectuales de cualquier hombre por inteligente que sea. Dado que el diseño de máquinas es una de estas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas aún mejores; entonces se produciría sin duda una "explosión de inteligencia" y la inteligencia del hombre quedaría muy atrás. Así, la primera máquina ultrainteligente es el último invento que el hombre necesitará hacer, siempre que la máquina sea lo suficientemente dócil como para decirnos cómo mantenerla bajo control.

La "explosión de inteligencia" de Good también ha sido llamada la singularidad tecnológica por el profesor de matemáticas y autor de ciencia ficción Vernor Vinge, quien escribió en 1993: "*Dentro de treinta años, tendremos los medios tecnológicos para crear inteligencia sobrehumana. Poco después, la era humana terminará*" . En 2017, el inventor y futurista Ray Kurzweil predijo que la singularidad aparecería en 2045, lo que significa que se acercaría 2 años en 24 años. (¡A ese ritmo, sólo faltan 336 años!) Vinge y Kurzweil señalan correctamente que el progreso tecnológico en muchos aspectos está creciendo exponencialmente en la actualidad.

Sin embargo, es un gran salto extrapolar todo el camino desde el costo de computación en rápida disminución hasta una singularidad. Hasta ahora, todas las tecnologías han seguido una curva en forma de S, donde el crecimiento exponencial eventualmente disminuye. A veces las nuevas tecnologías intervienen cuando las antiguas se estancan, pero a veces no es posible mantener el crecimiento, por razones técnicas, políticas o sociológicas. Por ejemplo, la tecnología de vuelo avanzó espectacularmente desde el vuelo de los hermanos Wright en 1903 hasta el alunizaje en 1969, pero desde entonces no ha habido avances de magnitud comparable.

Otro obstáculo en el camino para que las máquinas ultrainteligentes se apoderen del mundo es el mundo. Más específicamente, algunos tipos de progreso requieren no sólo pensar sino actuar en el mundo físico. (Kevin Kelly llama thinkismo al énfasis excesivo en la inteligencia pura). Una máquina ultrainteligente encargada de crear una gran teoría unificada de la física podría ser capaz de manipular inteligentemente ecuaciones mil millones de veces más rápido que Einstein, pero para lograr un progreso real, aún necesitaría recaudar millones de dólares para construir un supercolisionador más potente y realizar experimentos físicos a lo largo de meses o años. Sólo entonces podría empezar a analizar los datos y teorizar. Dependiendo de cómo resulten los datos, el siguiente paso podría requerir recaudar miles de millones de dólares adicionales para una misión de sonda interestelar que tardaría siglos en completarse. La parte del "pensamiento ultrainteligente" de todo este proceso podría ser en realidad la parte menos importante. Como otro ejemplo, una máquina ultrainteligente encargada de llevar la paz a Medio Oriente podría terminar frustrándose 1.000 veces más que un enviado humano. Hasta el momento, no sabemos cuántos de los grandes problemas son como las matemáticas y cuántos como Oriente Medio.

Mientras algunas personas temen la singularidad, otras la disfrutan. El movimiento social transhumanista espera un futuro en el que los humanos se fusionen con (o sean reemplazados por) inventos robóticos y biotecnológicos. Ray Kurzweil escribe en La singularidad está cerca (2005):

> **Importante** La Singularidad nos permitirá trascender estas limitaciones de nuestros cuerpos biológicos y cerebro. Ganaremos poder sobre nuestro destino... Seremos capaces de vivir todo el tiempo que queramos... Comprenderemos plenamente el pensamiento humano y extenderemos y ampliaremos enormemente su alcance. Para finales de este siglo, la parte no biológica de nuestra inteligencia será billones de billones de veces más poderosa que la inteligencia humana sin ayuda.

De manera similar, cuando se le preguntó si los robots heredarán la Tierra, Marvin Minsky dijo "sí, pero serán nuestros hijos". Estas posibilidades presentan un desafío para la mayoría de los teóricos morales, que consideran que la preservación de la vida humana y de la especie humana es algo bueno. Kurzweil también señala los peligros potenciales y escribe: "Pero la Singularidad también amplificará la capacidad de actuar según nuestras inclinaciones destructivas, por lo que aún no se ha escrito su historia completa". Los humanos haríamos bien en asegurarnos de que cualquier máquina inteligente que diseñemos hoy y que pueda evolucionar hacia una máquina ultrainteligente lo haga de una manera que termine tratándonos bien. Como dice Eric Brynjolfsson: “El futuro no está predeterminado por las máquinas. Es creado por humanos” .

6. Sesgos en el desarrollo y aplicación de la IA y el Big Data

En la actualidad, se puede considerar que las tecnologías relacionadas con la inteligencia artificial se encuentran lo suficientemente implantadas como para influir sustancialmente en algunos aspectos de la vida diaria. Una de las áreas donde más ha crecido el uso de la inteligencia artificial en los últimos años es en el comercio electrónico, particularmente en los sistemas de recomendación y los asistentes que interactúan con los consumidores. Los agentes de inteligencia artificial, como los asistentes virtuales y los chatbots, ofrecen publicidad dirigida mediante la vinculación con bases de datos de empresas como es el caso de Amazon o Google y son capaces de responder de forma inmediata a las consultas de los usuarios sobre productos y servicios. Esto permite a las empresas obtener una respuesta rápida sin comprometer recursos humanos.

Sin embargo, la idea de que los algoritmos de inteligencia artificial están libres de sesgos es errónea debido a varias razones, algunas relacionadas con los sesgos existentes en los datos utilizados para aprender y otras a causa de los sesgos inherentes a los propios algoritmos.

Los sesgos en los datos pueden deberse a mediciones sesgadas, decisiones humanas sesgadas o informes erróneos. Es necesario considerar que los algoritmos de aprendizaje automático están diseñados básicamente para replicar estos sesgos. Otra razón proviene de los objetivos algorítmicos, que apuntan a minimizar los errores generales de predicción y, por lo tanto, benefician a los grupos mayoritarios sobre las minorías. Además, los sesgos pueden surgir de valores o información faltantes, como sesgo de muestra o selección, lo que da como resultado conjuntos de datos que no son representativos de la población objetivo.

Ejemplo Un ejemplo común de sesgo algorítmico está en el campo de la justicia penal, donde se demostró que un algoritmo empleado por el sistema de justicia penal estadounidense arrojaba como resultado que la probabilidad de reincidencia criminal de los ciudadanos afroamericanos era el doble que la de los blancos.

Puede consultarse al respecto los siguientes artículos:

- Angwin, J, J. Larson, S. Mattu, and L. Kirchner. 2016. «Machine Bias.» Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. In Big Data (Vol. 5, Issue 2, pp. 153–163). Mary Ann Liebert Inc.

Ejemplo Otro ejemplo de sesgo algorítmico se encontró en la atención médica, donde se demostró que un algoritmo de gestión ampliamente utilizado favorecía a los pacientes de raza blanca sobre los negros.

Para obtener más información al respecto se puede consultar el artículo:

- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science. aax2342. PMID: 31649194.

Así, la literatura ha señalado varias causas que pueden conducir a los sesgos en los programas basados en inteligencia artificial, que pueden tener su origen bien en los datos o bien en el propio algoritmo. Seguidamente se exponen las principales causas conocidas:

- **Sesgos debidos a los conjuntos de datos utilizados para el aprendizaje**, que se basan en mediciones de dispositivos sesgadas, decisiones humanas históricamente sesgadas, informes erróneos, bases de datos con grupos representados en proporciones distintas de las correctas, etc.
- **Sesgos causados por datos faltantes o bien sesgos en la selección de la muestra** que dan como resultados conjuntos de datos que no son representativos de la población objetivo.
- **Sesgos que surgen de objetivos algorítmicos que pretenden minimizar los errores de predicción agregados generales** y, por tanto, benefician a los grupos mayoritarios sobre las minorías. Este concepto está relacionado con la paradoja de Simpson, según la cual una tendencia observada en subgrupos puede comportarse de manera bastante diferente, incluso a la inversa, cuando estos subgrupos se agregan y analizan juntos.
- **Sesgos causados por variables relacionadas con los atributos sensibles**. Se entienden como atributos sensibles aquellos que diferencian a los grupos privilegiados y no privilegiados, como la raza, el género y la edad, y que, por lo general su uso no resulta legítimo para la toma de ciertas decisiones. Las variables relacionadas son variables no consideradas como sensibles pero que presentan una relación fuerte con dichas variables. En el caso de existir variables relacionadas dentro de un conjunto de datos, resultaría posible que el algoritmo de aprendizaje automático tomara decisiones implícitamente basadas en los atributos sensibles bajo la apariencia de usar atributos presuntamente legítimos pero que realmente son variables relacionadas.

Ejemplo Un ejemplo clásico de la paradoja de Simpson es el relativo a la supuesta discriminación por género que se producía en la Universidad de California en Berkeley (Estados Unidos) en el año 1973. Así, los datos de admisión mostraron que los hombres que presentaban solicitudes tenían más probabilidades de ser admitidos que las mujeres, y la diferencia era tan grande que era poco probable que se debiera a la casualidad.

Sin embargo, al analizar las admisiones departamento a departamento, se observaba que seis de los 85 departamentos tenían un sesgo significativo en contra de los hombres, mientras que cuatro lo tenían en contra de las mujeres. Es d-ecir, realmente existía un pequeño sesgo favorable a las mujeres. De este análisis se concluyó que las mujeres tendían a postularse a departamentos más competitivos con tasas de admisión más bajas, mientras que los hombres tendían a hacerlo a departamentos menos competitivos con tasas de admisión más altas.

Ejemplo El Libro Blanco sobre la Inteligencia Artificial desarrollado por la Comisión Europea presenta los ejemplos que se muestran a continuación:

- «Puede suceder que el uso de determinados algoritmos de inteligencia artificial para predecir la reincidencia delictiva dé lugar a prejuicios raciales o de género, y prevea una probabilidad de reincidencia distinta para hombres y mujeres o para nacionales y extranjeros. Fuente: Tolan S., Miron M., Gomez E. and Castillo C. «Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia», Best Paper Award, International Conference on AI and Law, 2019.»
- «Algunos programas de inteligencia artificial de análisis facial muestran prejuicios raciales o de género, y presentan un bajo nivel de error a la hora de determinar el género de hombres de piel más clara, pero un elevado nivel de error al determinar el género de mujeres de piel más oscura. Fuente: Joy Buolamwini, Timnit Gebru; Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018.»

Por tanto, dado que para el entrenamiento y validación de los sistemas de inteligencia artificial es necesario hacer uso de conjuntos de datos y que las acciones y decisiones a las que pueden llevar dependen en gran medida de los conjunto de datos que se hayan utilizado para entrenar los sistemas, deben adoptarse las medidas necesarias para garantizar que, en lo que se refiere a los datos utilizados para entrenar los sistemas de inteligencia artificial, se respeten los valores y normas de la UE, concretamente con relación a la seguridad y la legislación vigente para la protección de los derechos fundamentales.

Finalmente, cabe señalar que con gran frecuencia los desarrolladores otorgan a los dispositivos robóticos características antropomórficas, puesto que muchas personas desconfían o no les gusta hablar con máquinas. Teniendo en cuenta que múltiples agentes de inteligencia artificial pueden simular de manera convincente interacciones interpersonales, no es sorprendente que 1 las personas interactúen con ellos como si fueran humanos. Los resultados obtenidos de la investigación sobre las relaciones interpersonales y las habidas entre el consumidor y las marcas, sugieren que las interacciones entre los humano y los agentes inteligentes se guían por las mismas normas que rigen las relaciones interpersonales.

La investigación sobre los estereotipos sugiere que es más probable que las personas asocien a las mujeres con calidez y a los hombres con competencia (Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In Advances in Experimental Social Psychology (pp. 61–149). Elsevier; Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In Journal of Personality and Social Psychology (Vol. 82, Issue 6, pp. 878–902). American Psychological Association APA). Los estudios también han demostrado que cuando las personas se comunican con dispositivos tecnológicos como computadoras, máquinas y entidades de inteligencia artificial, usan los mismos estilos de diálogo y métodos de procesamiento de información que utilizan en sus comunicaciones interpersonales humano a humano.

Por tanto, los estereotipos de género y en consecuencia sus sesgos, también se producen en las interacciones entre los humanos y lo agentes inteligentes. Existen estudios que han demostrado que los usuarios no se comportan de igual manera cuando interactúan con un sistema de inteligencia artificial al que se le han otorgado rasgos de apariencia masculina que cuando interaccionan con otro que tiene rasgos femeninos. Así, se ha manifestado que, por ejemplo, la capacidad de persuasión de las recomendaciones de los programas de inteligencia artificial varía según las personalidades percibidas por los usuarios. Existe un estudio en el que se llegó a la conclusión que, para los productos utilitarios, los participantes confiaron más en las recomendaciones de los agentes de inteligencia artificial con apariencia masculina que en las recomendaciones de los agentes de inteligencia artificial con apariencia femenina. Sin embargo, se produjo el caso contrario para los productos hedónicos.

7. Regulación Europea sobre IA

7.1. Normativa regulatoria de la Unión Europea sobre IA

El eje normativo en la materia es el **Reglamento (UE) 2024/1689, "AI Act"** , publicado en el *DOUE* el 12 de julio de 2024. Con arreglo a su artículo 113:

- **Entrada en vigor:** 1 de agosto de 2024.
- **Aplicación escalonada:**
  - 2 febrero 2025 → prohibiciones absolutas (p. ej., "social scoring") y programas de alfabetización en IA.
  - 2 agosto 2025 → obligaciones para los modelos de IA de propósito general (GPAI) y puesta en marcha del régimen sancionador; entra en funciones la **Oficina Europea de IA** —creada por Decisión de la Comisión de 24 enero 2024— como órgano coordinador.
  - 2 agosto 2026 → exigibilidad plena para la mayoría de sistemas **de alto riesgo**.
  - 2 agosto 2027 → plazo adicional sólo para los componentes de IA incrustados en productos ya regulados (automoción, aviación, etc.). EUR-Lex Estrategia Digital Europea Estrategia Digital Europea

7.2. Obligaciones de los Estados miembros

Los Estados deben, entre otras tareas, **designar antes del 2 de agosto de 2025** a sus autoridades nacionales de vigilancia de mercado y notificación de organismos de evaluación de la conformidad; asimismo, debían publicar antes del 2 noviembre 2024 la lista de autoridades competentes en derechos fundamentales. Artificial Intelligence Act EU

7.3. Transposición y aplicación en España

- **Agencia Española de Supervisión de la Inteligencia Artificial (AESIA).** España se adelantó creando la AESIA mediante el **Real Decreto 729/2023, de 22 de agosto**, que aprueba su Estatuto y la perfila como futura autoridad supervisora única para la AI Act. BOE
- **Designación oficial.** La AESIA figura ya en el inventario europeo de planes nacionales como autoridad de mercado y de notificación; resta completar su dotación y coordinarla con la AEPD y otros reguladores sectoriales antes de la fecha límite de agosto 2025. Artificial Intelligence Act EU

- **ANTEPROYECTO DE LEY para el buen uso y la gobernanza de la Inteligencia Artificial, a los efectos previstos en el artículo 26.4 de la Ley 50/1997, de 27 de noviembre, del Gobierno.** El Consejo de Ministros ha aprobado el Anteproyecto de Ley para el buen uso y la gobernanza de la Inteligencia Artificial, a los efectos previstos en el artículo 26.4 de la Ley 50/1997, de 27 de noviembre, del Gobierno, que busca garantizar un uso de la Inteligencia Artificial que sea ético, inclusivo y beneficioso para las personas. Este instrumento normativo adaptará la legislación española al reglamento europeo de IA, ya en vigor, bajo un enfoque regulador que impulsa la innovación. Consejo de Ministros

Con estas piezas ya en marcha, España cumplirá el calendario europeo y dispondrá de un marco integrado que combine la supervisión ex-ante (conformidad y registro) y ex-post (sanciones e inspecciones) para todos los sistemas de IA desplegados en su territorio.

¿Debe haber un equilibrio entre tecnología y ley? ¿Quién y cómo debe definir límites a la IA?

8. Fuentes de información

- Wikipedia
- GhatGPT
- Modelos de Inteligencia Artificial (Ed. Marcombo)
- https://iep.utm.edu/artificial-intelligence/