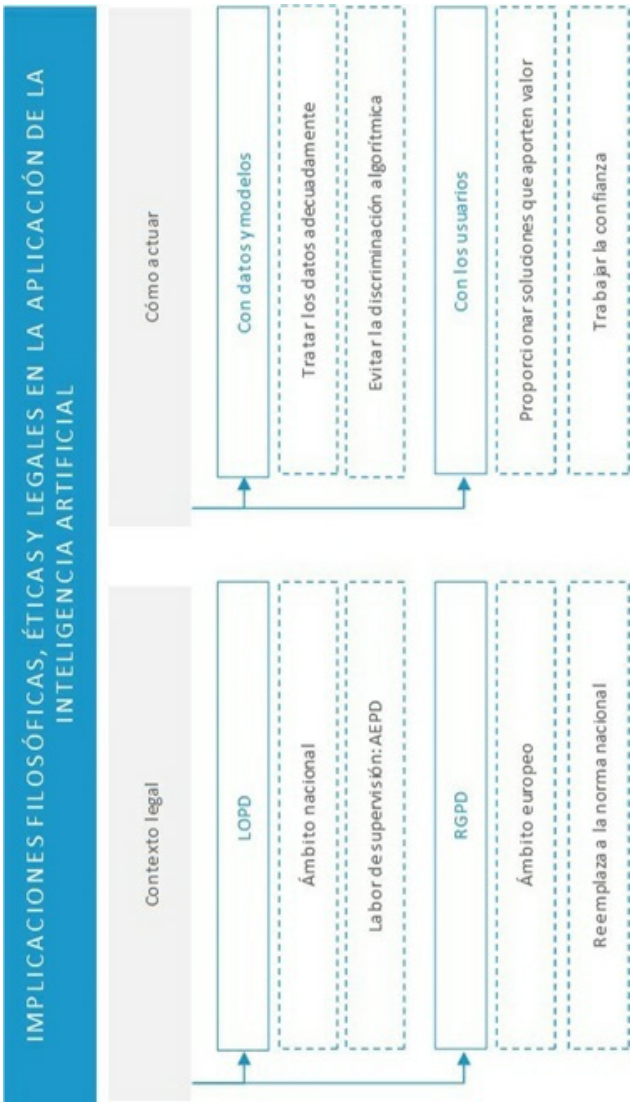
Investigación y Gestión de Proyectos en Inteligencia Artificial

Tema 12. Implicaciones filosóficas, éticas y legales en la aplicación de la inteligencia artificial

Índice

Esquema



12.1. Introducción y objetivos

La aplicación de la inteligencia artificial está unida no solo a retos técnicos o científicos sino también a **retos éticos y legales** propios de una disciplina que está llamada a cambiar notablemente la forma en la que interactuamos con los demás, los métodos de trabajo y toda la actividad económica. En esta unidad nos planteamos los **objetivos** siguientes:

- Identificar retos éticos y filosóficos en el empleo de la IA (inteligencia artificial).
- Conocer las peculiaridades legales del trabajo basado en datos.
- Identificar nuevas tendencias regulatorias que implican cambios a la hora de trabajar con datos.

12.2. Contexto legal aplicable a proyectos de inteligencia artificial

Los proyectos de inteligencia artificial están **basados en datos.** Necesitamos datos para entrenar y para validar los modelos. También necesitamos almacenar y analizar los datos que genera la interacción con el modelo, solo así seremos capaces de determinar si la puesta en práctica de la solución cumple con los objetivos previstos. La evaluación constante de los modelos es lo que permite su mejora y evolución.

Por tanto, los proyectos de inteligencia artificial se ven claramente influenciados por la legislación vigente en materia de protección de datos.

En España, la Agencia Española de Protección de Datos (AEPD), como autoridad de control independiente, tiene encargada la misión de velar por el cumplimiento de la normativa de protección de la información personal.

La agencia es un ente de derecho público, con personalidad jurídica propia y plena capacidad pública y privada, que actúa con plena independencia de las Administraciones públicas en el ejercicio de sus funciones. Se relaciona con el Gobierno a través del Ministerio de Justicia. La **Ley Orgánica de Protección de Datos de Carácter Personal (LOPD) 15/1999, de 13 de diciembre,** regula gran parte de las directrices que se deben tener en cuenta a la hora de manipular información personal.

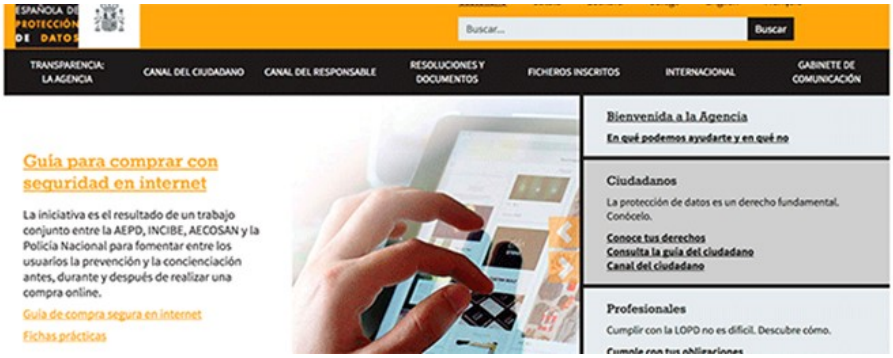


Figura 1. Página web de la AEPD. Fuente: https://www.agpd.es/portalwebAGPD/index-ides-idphp.php

Bajo el paraguas de la Unión Europea surge el Reglamento General de Protección de Datos (RGPD), con la aspiración de unificar los regímenes de todos los Estados miembros sobre la materia. Aunque entró en vigor el día 25 de mayo de 2016, su cumplimiento solo será obligatorio transcurridos dos años desde esta fecha.

A continuación, se desglosarán **algunas de las características más destacadas de estas normativas,** así como sus implicaciones para la tipología de los proyectos que nos ocupa.

Protección de datos de carácter personal

Se considera un dato de carácter personal **cualquier información concerniente a personas físicas identificadas o identificables.** Por tanto, la legislación no solo contempla como datos de carácter personal aquellos casos en los que es posible identificar unívocamente al individuo en base a un atributo exclusivo como el documento nacional de identidad. También contempla la posibilidad de que dicha identificación pueda producirse en el futuro vía cruce de datos de distintas tipologías.

> Por ejemplo, supongamos que tenemos almacenada en un fichero o base de datos información sobre los clientes de una determinada compañía. Supongamos ahora que contamos con campos poco relevantes y genéricos. En esta situación, la compañía decide añadir la edad de cada usuario a la información existente. Si lo datos anteriores realmente son

irrelevantes (siempre referido a cuestiones de identificación personal), la edad por sí sola no podría servir para identificar a un individuo, puesto que, ¿cuántas personas hay de 20 años?, ¿y de 35? Las opciones son muy diversas. En estas condiciones, la compañía decide incorporar a la base de datos el género. Ahora las opciones se reducen, evidentemente, hay más personas que tienen 35 años a personas que tienen 35 años y son mujeres (por ejemplo). No obstante, parece poco creíble pensar que solo existe una única persona que tenga 35 años y sea mujer. Animados por lo útil que es insertar este tipo de atributos en el estudio, el responsable del proyecto decide ahora incorporar el código postal del usuario. En esta situación, y si la compañía dispone de especialistas en seguridad y normativa de protección de datos, se levantará una señal de alarma. Con la edad, género y código postal sí que existe una mínima posibilidad de que se pueda identificar de forma unívoca a una persona (especialmente en grupos de edades menos frecuente). En diversos entornos de la materia, a la combinación de los atributos edad, género y código postal se la conoce con el nombre de triada.

Antes de continuar, queremos dejar patente que no se prohíbe el almacenamiento de información personal. Si así fuese, ninguna empresa o institución pública podría funcionar. La **misión de la ley es establecer distintos niveles obligatorios de seguridad y control** en función de la tipología de la información. Incluso dentro de la información que puede ser considerada como información personal existen distintos niveles de criticidad. No es lo mismo almacenar edad, género y código postal, que almacenar el número de la cuenta bancaria o, más crítico todavía, el número completo de la tarjeta de crédito con su fecha de caducidad y código de seguridad.

Las distintas iniciativas regulatorias que van surgiendo se encaminan a posibilitar a los usuarios finales o clientes un mayor control sobre los datos que ceden a las

compañías y el uso que estas hacen de estos datos. En su artículo 4, la LOPD dice:

> «1. Los datos de carácter personal solo se podrán recoger para su tratamiento, así como someterlos a dicho tratamiento, cuando sean adecuados, pertinentes y no excesivos en relación con el ámbito y las finalidades determinadas, explícitas y legítimas para las que se hayan obtenido.
>
> »2. Los datos de carácter personal objeto de tratamiento no podrán usarse para finalidades incompatibles con aquellas para las que los datos hubieran sido recogidos. No se considerará incompatible el tratamiento posterior de estos con fines históricos, estadísticos o científicos».

Es por tanto preciso **que el propietario de los datos autorice el almacenamiento y tratamiento de su información personal.** Así mismo, los responsables del repositorio de datos deben garantizar a los usuarios los métodos adecuados para la rectificación, modificación, acceso y cancelación de la información suministrada.

Confianza

Dentro de ciertos sectores sociales existe una preocupación relevante por el uso que determinadas compañías están haciendo con sus datos, y quizá esta preocupación esté justificada si nos basamos en ciertos titulares periodísticos que se han publicado en los últimos años. En este tipo de cuestiones entraremos más en detalle en el próximo apartado. Ahora intentaremos explicar cómo se puede establecer una relación de confianza entre el usuario y la organización para que este ceda sus datos.

> Supongamos que un paciente estándar de 50 años está en la revisión anual con su médico. El hospital está interesado en crear un banco de datos y compartirlos con personal externo. Dado que es obligatorio otorgar permiso explícito para ceder datos de carácter personal, el facultativo nos pode delante un formulario donde simplemente se indica el texto: «autorizo la cesión de mis datos…».

Sin disponer de más información, ¿qué haría el lector? Cambiemos la situación yendo a un caso algo más extremo.

> Supongamos que ese mismo paciente ha sido diagnosticado recientemente de una enfermedad que tiene una incidencia muy baja, pero, desgraciadamente, sus efectos en el enfermo son importantes. Textualmente, lo que nos comenta ahora el médico es:

> «Nos gustaría que nos cediese su datos personales y clínicos para incorporarlos a nuestro banco de datos. Lo que pretendemos con este proyecto es compartir información con especialistas de todo el país e investigadores muy destacados en el área. Dado que su enfermedad es extremadamente rara, consideramos que sería muy útil compartir sus datos con la comunidad científica. Seguro que así podremos recibir recomendaciones que nos ayudaría a personalizar el tratamiento a sus características. Además, no compartiremos sus datos personales más íntimos, toda la información estará anonimizada. Ante cualquier cuestión, seremos nosotros quienes nos comunicaremos con usted como venimos haciendo hasta ahora».

¿Qué haría el lector en esta hipotética situación?

Es probable que las respuestas ante ambos casos de uso sean muy distintas. Y sería normal, puesto que se trata de dos casos muy diferentes. En la primera situación, el paciente no aprecia el retorno de valor que le produce la cesión de sus datos. Además, al recibir poca información sobre el proyecto es difícil generar una relación de confianza. En el segundo caso, el beneficio que el paciente percibe como consecuencia de la cesión de datos es evidente y relevante. La comunidad podría aportar un tratamiento que retarde la enfermedad o incluso la cure. Pero para que

esto sea una realidad es preciso que especialistas compartan información sobre qué tratamientos han funcionado, cuáles no, cuáles han sido los efectos secundarios, qué peculiaridades tenían los pacientes… Además, el facultativo ha trabajado la confianza al solicitar la cesión de los datos, ha explicado la utilidad de la cesión de los datos y ha establecido unas garantías que invitan al paciente a estar tranquilo sobre quién podrá contactarle. ¿Cuántas veces hemos recibido llamadas publicitarias ofreciendo productos sin ningún tipo de interés para nosotros y sin saber cómo habían accedido a nuestros datos?

Lo que se ha pretendido explicar en los párrafos anteriores se puede resumir en los siguientes puntos:

- Los algoritmos de inteligencia artificial (por ejemplo, los algoritmos de aprendizaje automático) emplean datos como materia prima.
- Muchas veces estos datos incluyen información personal.
- Para manipular información personal (aunque esté anonimizada) el usuario debe haber emitido su autorización.
- Para que el usuario ceda sus datos personales es preciso trabajar un entorno de confianza. Este entorno de confianza se basa en los siguientes puntos:
  - El usuario tiene claro a quién (a nivel de institución) y para qué va a acceder los datos.
  - El usuario tiene claro qué se va a hacer con sus datos.
  - El usuario percibe un retorno de valor evidente que le invita a permitir la cesión de datos.
  - El usuario es informado de que sus datos van a ser tratados de forma adecuada y

segura.

Establecer mecanismos que cultiven e implementen esta relación de confianza entre el usuario o cliente y la empresa o la institución es uno de los grandes retos a los que se enfrenta la industria hoy en día.

Anonimización de datos personales

El tratamiento, análisis y explotación de grandes volúmenes de datos puede producir infinitud de beneficios a la sociedad, pero es preciso compatibilizar dicho almacenamiento y manipulación con el respeto a la protección de datos personales.

El mecanismo principal que permite alinear la generación de beneficios tangibles con el respeto a la privacidad es la **anonimización de la información.** Según la RAE, anonimizar es «expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad». Por otro lado, y según la AEPD, «un proceso de anonimización es aquel que elimina o reduce al mínimo los riesgos de identificación de los datos anonimizados manteniendo la veracidad de los resultados tras el tratamiento de estos».

Para garantizar su efectividad, **este proceso de anonimización debe ser irreversible.** Es decir, no debe existir posibilidad de recuperar el dato adicional empleando solo los datos anonimizados. Es necesario precisar que, debido a los avances tecnológicos y algorítmicos, no es posible garantizar de forma absoluta y completa la anonimización de la información. El objetivo realizable es proporcionar las mayores garantías posibles de cara a asegurar el respeto a la privacidad de las personas.

Según la AEPD, el proceso de anonimización debe atender a los siguientes **principios:**

Ideas clave

- **Principio proactivo.** La privacidad se debe garantizar de forma proactiva y no de forma reactiva y una vez que se haya producido alguna fuga de información.
- **Principio de veracidad por defecto.** Se debe considerar la granularidad o grado de detalle final que deben tener los datos anonimizados. Esto lleva a que, en ocasiones, se exija la eliminación de ciertos datos para garantizar la anonimización del conjunto.
- **Principio de privacidad objetiva.** Siempre existirá un error residual de riesgo de reidentificación que deberá ser aceptable en función de la información anonimizada, conocido por el usuario y asumido por el responsable del fichero.
- **Principio de plena funcionalidad.** El proceso de anonimización debe garantizar la utilidad de los datos anonimizados en base a los objetivos inicialmente establecidos.
- **Principio de privacidad en el ciclo de vida de la información.** El respeto a la privacidad de los usuarios debe garantizarse durante todo el proceso de anonimización. Por ejemplo, realizando el proceso de anonimización en los sistemas preparados y autorizados a almacenar la información sin anonimizar.
- **Principio de información y formación.** Todo el personal con acceso a datos anonimizados o no deben estar correctamente formados e informados acerca de sus obligaciones.

Técnicamente hablando, existen varios **procedimientos para asegurar la anonimización de la información.** Algunos de los más destacados son:

- **Desnaturalizar:** consistente en transformar la naturaleza del dato. Por ejemplo, en lugar de representar la edad (edad = 42), podemos indicar el rango de edad al que pertenece en base a alguna división previamente establecida (edad = 5 donde 5

hace referencia al intervalo [40, 50]).

- **Cifrar:** consiste en hacer ilegible un mensaje concreto con base en la aplicación de un algoritmo que precisa de un conjunto de claves. En este caso, el descifrado de la información es posible siempre que se disponga del algoritmo de las claves necesarias.

- ***Tokenizar:*** se reemplaza el valor a anonimizar por un valor distinto (token) que no suele respetar la naturaleza del dato. Por ejemplo, se *tokeniza* el DNI 04345566D cambiándolo por el valor YID884S3VVQW4ZZY1. Se puede observar cómo no se respeta el formato estándar de un DNI. Para que el proceso mantenga la coherencia, al mismo valor le debe siempre corresponder el mismo *token*. La reversibilidad es posible siempre que se disponga del *token* correspondiente a cada valor.

- **Funciones *hash*:** es un método parecido a la *tokenización*. En este caso se aplica una función matemática al valor a anonimizar. Dicho valor reemplaza al valor original. En este caso, y por la propia naturaleza del proceso, se garantiza la irreversibilidad del proceso. No obstante, esta técnica podría, ocasionalmente, generar el mismo valor hash para distintos valores de entrada.

- **Disociar:** eliminar parte de la información para evitar la identificación personal. Por ejemplo, pongamos que disponemos de la siguiente información sobre un paciente: fecha de la consulta, hora de la consulta, código postal, edad, sexo y síntomas. Con esa información encima de la mesa existiría un riesgo de identificar al paciente, ya que solo la fecha y hora de la consulta proporcionan bastante información. La disociación consistiría en eliminar ciertos campos quedándonos solo (por ejemplo), con código postal, sexo y síntomas. De esta forma se reduce el riesgo de identificación.

Reglamento General de Protección de Datos (RGPD)

Aunque diseñada en el año 2016, el RGPD de la Unión Europea pasó a ser de

obligado cumplimiento en el año 2018. La aplicación de esta normativa supone un gran reto para muchas compañías en las que el dato y la analítica sobre el mismo constituye un eje esencial de su actividad. El incumplimiento de la normativa implica cuantiosas multas para los infractores.

El RGPD exige **medidas adicionales que garanticen la transparencia en el tratamiento de datos personales.** Se pretende así empoderar al ciudadano para que tome las decisiones más adecuadas. Las organizaciones deben informar al usuario sobre el tipo de perfilado o modelización que realizan en base a su dato personal facilitándole la denegación del permiso para realizar dicho tratamiento si así es solicitado.

Especialmente en el entorno financiero y asegurador, una de las medidas que más impacto causará es el **derecho a explicación.** Esta medida obliga, por ejemplo, a las entidades financieras a explicar las razones por las que un crédito ha sido denegado. De esta forma, el usuario puede actuar en consecuencia buscando alternativas para obtener una mejor clasificación en el futuro.

Además, los algoritmos implementados deben asegurar que se garantiza la no discriminación a la hora de tomar decisiones basadas en datos. No podrán tomarse decisiones basadas en criterios como la raza, la edad, el sexo, la religión, etc.

La Unión Europea ha habilitado el portal https://www.eugdpr.org/ donde se desglosa todo tipo de información sobre esta regulación.

Figura 2. Portal de la RGPD. Fuente: https://www.eugdpr.org/

Ideas clave

En esta unidad se propondrá al alumno la realización de un trabajo que versará, precisamente, sobre esta normativa.

12.3. Sesgos, legalidad y responsabilidad

Lo que hace unos años era una cuestión ética, evitar la discriminación a la hora de emplear datos y algoritmos en la toma de decisiones, ahora se ha convertido en ley. En un entorno tecnológico y científico con una evolución tan rápida e impredecible, el entorno regulatorio no siempre es capaz de llegar a tiempo para proteger los derechos de los ciudadanos, por lo que el especialista, científico o técnico debe mantener altos estándares éticos en el día a día de su trabajo.

Hay una serie de puntos para tener en cuenta en este sentido:

- Tener siempre presente la normativa de protección de datos.
- Asegurarse que los algoritmos que empleamos no conllevan la toma de decisiones implicando la discriminación de algún colectivo por edad, sexo, raza, religión o cualquier otro aspecto.
- Comprobar que los datos empleados no contienen sesgos que puedan llevar a tomar decisiones equivocadas.
- Interpretar los resultados de los modelos científicamente, evitando interpretaciones interesadas y no ajustadas a la realidad.
- Emplear los métodos de trabajo adecuados que garanticen la fiabilidad de los resultados.

Conceptos como inteligencia artificial y aprendizaje automático forman parten del vocabulario diario de comités de empresa y gobiernos de todo el mundo. Las decisiones que se toman con base en estos modelos afectan a millones de personas.

Por tanto, es responsabilidad de los expertos asegurar que el objetivo final es aportar beneficio a la sociedad en general o los clientes de la compañía en particular, respetando los derechos fundamentales de la ciudadanía.

Sesgos: el motivo por el que los algoritmos aprenden y su principal punto débil

Como ya vimos anteriormente, **los sesgos son imposibles de eliminar ya que es el mecanismo por el que los algoritmos aprenden.** Crear sesgos y hacer suposiciones globales despreciando los detalles concretos es la base del aprendizaje y, por tanto, las excepciones siempre van a estar ahí. Lo importante en este caso es minimizarlos y que estos sesgos no aprendidos no sean sesgos inducidos por el entrenamiento debido a una mala elección de los datos de entrenamiento.

Es por tanto muy importante tener en cuenta este tipo de cuestiones en entornos sensibles y con sesgos que pueden incurrir en discriminaciones de cualquier tipo. Si no se puede discriminar a un individuo por razones ideológicas, de sexo, étnicos, etc., los algoritmos no pueden hacerlo tampoco.

¿Cómo evitar este hecho?

- Seleccionar cuidadosamente los datos de entrenamiento.
- Validar los algoritmos no solo con los datos provenientes del primer mundo o de nuestra área de influencia sino de otras partes del mundo con rasgos, culturas o éticas diferentes.
- Mantener una vigilancia continua de las decisiones que estos toman, para intervenir lo antes posible si se detectan estos sesgos.
- Tener evaluadores humanos que confirmen las decisiones tomadas por los algoritmos o al menos que los usuarios que se vean afectados por ellos puedan acudir para que se revise su caso particular.

Este va a ser uno de los retos más importante de aquí a unos años cuando los algoritmos de IA sean los que vayan controlando cada vez más y más procesos en los que la vida de los ciudadanos se vea afectada.

En cuanto a la **responsabilidad de los errores de estos algoritmos,** pues es bastante complicado establecerla y es algo en lo que debemos esforzarnos como sociedad. Obviamente, la responsabilidad siempre debe ser de la organización que use estos algoritmos que deben probarlos correctamente. Pero nada está exento de errores al 100 %. Así que hay que establecer cuál es el margen de error admisible para cada tarea que se delegue a la IA. Al fin y al cabo, los humanos también nos equivocamos. La diferencia aquí es que la cadena de responsabilidad entre los humanos es trazable y queda más o menos clara. Cuando entra dentro un algoritmo de IA ya no está tan clara, ¿es de los diseñadores del algoritmo?, ¿es de quien seleccionó los datos de entrenamiento? ¿Es de quien no revisó correctamente dicha decisión? Se entra en un terreno desde nuestro punto de vista pantanoso que tendremos que ir valorando con el tiempo y con leyes nuevas que probablemente lleguen tarde para algunos casos. Este es un reto importante a tener en cuenta en el futuro.

Implicaciones de la inteligencia artificial

La explosión de la inteligencia artificial ha marcado o marcará el inicio de una nueva revolución de dimensiones equivalentes a la Revolución Industrial. Como en toda gran transformación, y a pesar de las nuevas posibilidades que deslumbran, surgen también dudas y temores. Por ejemplo:

- ¿Perderán las personas su trabajo siendo reemplazadas por máquinas?

Ideas clave

- ¿En qué trabajarán los humanos? ¿Cuánto tiempo?
- ¿Cómo se transformará nuestro tiempo de ocio?
- ¿Qué sucederá con nuestro derecho a la privacidad?
- ¿Cómo interactuaremos con los sistemas basados en inteligencia artificial? ¿Quién liderará la relación?
- ¿Podrán las máquinas sentir? ¿Cómo actuaremos entonces?
- ¿Dominarán los robots a la raza humana?

A la mayoría de estas preguntas no es posible responderla en estos momentos. Sin embargo, el debate sobre las implicaciones laborales de la robótica y la inteligencia artificial ya ha empezado.

En primer lugar, la gran mayoría de flujos de trabajo serán automatizados y no requerirán intervención humana. Esto permitirá mejorar la eficiencia de los procesos y, quizá, reducir los costes.

Los **algoritmos se convertirán en los principales agentes en ciertos escenarios.** Por ejemplo, muchas de las operaciones bursátiles que se llevan a cabo hoy en día son ejecutadas de forma automática mediante algoritmos que evalúan la evolución del mercado e incluso el contexto informativo analizando la información que circula sobre el contexto económico y social.

Los robots no solo han transformado las fábricas del siglo XXI, también están introduciéndose en el terreno sanitario e incluso militar.

El reto que tenemos por delante es **convertir los riesgos en oportunidades reales.**

Igual que sucedió con la Revolución Industrial, muchos trabajos desaparecerán

porque ya no serán necesarios al poder realizarse de forma automática. Sin embargo, surgirán otros muchos nuevos trabajos y, por supuesto, los métodos de trabajo cambiarán radicalmente. La **educación adquiere una relevancia especial** para dirigir la transformación global de habilidades y conocimientos que permita la inmersión en el nuevo ecosistema. Algunos autores sugieren gravar mediante nuevos impuestos ciertos productos basados en inteligencia artificial, como los robots industriales para que el Estado reciba, al menos temporalmente, fuentes adicionales de financiación que permitan ayudar a aquellas personas que se queden sin trabajo debido a la automatización de sus posiciones y, además, pueda impulsar un nuevo sistema educativo que promueva la creatividad y genere el ecosistema humano que este nuevo paradigma necesita. A esto se le ha denominado que los **robots paguen impuestos.** Otros autores comentan que se debe crear una renta básica que permita la subsistencia mínima individual. En cualquier caso, es un reto muy importante que debemos abordar en los próximos años.

La inteligencia artificial redefinirá la economía del futuro impulsando la productividad y generando una oleada de productos personalizados basados en las necesidades del cliente. Este impacto se extenderá por todos los ámbitos de actividad.

Regulando la inteligencia artificial

En los últimos años, destacados y muy relevantes personajes públicos han planteado la necesidad de regular de forma más exigente el ámbito de la inteligencia artificial de cara a evitar un mal uso de esta.

Entre estos actores destaca especialmente la organización Future of Life (https://futureoflife.org/). Esta fundación propone medidas para mitigar los riesgos de un mal uso no solo de la inteligencia artificial, sino también de la biotecnología o de las armas nucleares. Además, contemplan acciones para mitigar el cambio climático.

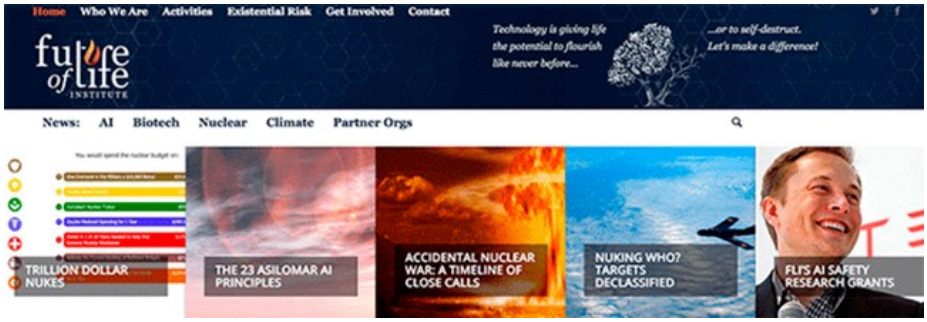


Figura 3. Página web de Future of Life. Fuente: https://futureoflife.org/

La lista de personalidades detrás de este proyecto es absolutamente espectacular. Entre sus fundadores podemos encontrar destacados científicos procedentes del MIT, DeepMind, Universidad de Santa Cruz o Boston, además de a Jaan Tallinn, cofundador de Skype. La labor de la institución está apoyada públicamente por figuras como Alan Alda, Erik Brynjolfsson (director del MIT Center for Digital Business), Morgan Freeman, el conocidísimo Stephen Hawking, Elon Musk (fundador de Tesla), Stuart Russell (autor del libro de referencia en esta asignatura) y un largo etcétera.

Una de las grandes preocupaciones de este grupo es la **aplicación de la inteligencia artificial con fines militares.** Atributos como la empatía, la justicia, la responsabilidad, la compasión, etc., son de momento atributos puramente humanos que las máquinas no tienen ocasión de contemplar.

Future of Life también se plantea romper con muchos de los mitos ligados a la inteligencia artificial:

Ideas clave

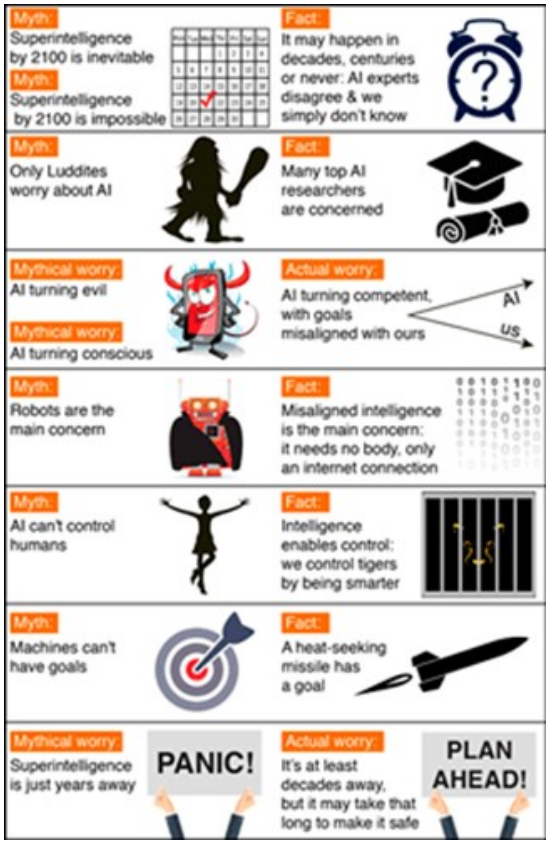


Figura 4. Mitos asociados a la inteligencia artificial según Future of Life. Fuente: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

La imagen anterior podría resumirse en los siguientes puntos:

- No es preciso definir qué es lo que seremos capaces de conseguir y cuándo.
- Muchos y grandes expertos están muy preocupados sobre el uso que se podría hacer de la inteligencia artificial.
- En ocasiones la inteligencia artificial puede producir soluciones no alienadas con los objetivos adecuados.
- No es solo cuestión de preocuparse por los robots, los algoritmos autónomos son también una cuestión que hay que vigilar.

Ideas clave

- Ciertas soluciones de inteligencia artificial podrían condicionar y controlar el comportamiento humano.
- Las máquinas pueden plantearse un objetivo de forma predefinida.
- Es mejor prevenir que curar. Actuemos antes de que sea tarde estableciendo las bases que permitan el desarrollo de una inteligencia artificial segura y alineada con las necesidades humanas.

Hay que destacar que este grupo no está en contra de la inteligencia artificial. De hecho, mucho de sus colaboradores forman parte de la élite investigadora de la materia. Lo que se pide no es abandonar el desarrollo de la inteligencia artificial, sino **establecer unas bases regulatorias adecuadas** que permitan asegurar que su desarrollo se produce de forma segura.

12.4. Seguridad y tolerancia ante ataques

Con el uso cada vez mayor de la inteligencia artificial y de la compartición de datos, las posibilidades de ataques en este tipo de sistemas están aumentando considerablemente. El uso intensivo de datos en los sistemas de IA hace que estos mejoren y se puedan aplicar a múltiples entornos que antes no eran posibles. Pero nos surge una duda, ¿cómo queda la seguridad y la privacidad de mis datos al adoptarla?

Ahora mismo muchas compañías están utilizando sistemas y modelos que van aprendiendo sobre la marcha y estos modelos no solo son usados por una compañía. Muchas veces, distintas compañías hacen uso de los mismos modelos debido a que la mayoría de estos modelos se enfocan al uso del *software* como servicio. De forma que, de una u otra forma, tus datos forman parte de una red que usa más gente. Por esta razón, ¿están seguros esos datos?

Por otro lado, con la creciente digitalización cada vez habrá más datos conectados a Internet y, por tanto, más puertas abiertas que pueden ser aprovechadas por los ciberatacantes. Por eso, es muy importante que estos datos estén bien protegidos; sobre todo, si son datos sensibles.

También comienzan a verse ataques a los propios algoritmos de IA. Como todo sistema informático, estos sistemas pueden tener agujeros de seguridad. Pero también podemos ir más allá. Los sistemas de IA y de *machine learning* como sabemos tienen sesgos y los atacantes pueden utilizar esos sesgos a su favor. Por ejemplo, falseando los datos de entrada, aprovechando de alguna vulnerabilidad del algoritmo para que este falle en su beneficio. Por ejemplo, para conseguir que una IA descarte o acepte un préstamo en un banco.

Los sistemas de reconocimiento de texto y de lenguaje natural, por ejemplo, han

avanzado enormemente en los últimos años, pero los programas de IA que analizan el texto pueden ser engañados por frases cuidadosamente elaboradas. Una frase que parece sencilla para un humano puede engañar a un algoritmo de IA. Esto puede implicar, por ejemplo, que un sistema que filtre noticias falsas sea engañado colando una noticia falsa sutilmente camuflada o que rechace o clasifique correctamente un posible candidato a un empleo. A este tipo de ejemplos que se le conoce con el nombre de adversarial *examples.*

Podemos definir, por tanto, un **adversarial example** como un caso con pequeñas perturbaciones intencionales de características que hacen que un modelo de aprendizaje de una máquina haga una predicción falsa.

Algunos ejemplos de este tipo de adversarial examples pueden ser:

- Un auto que se conduce a sí mismo choca con otro coche porque ignora una señal modificada para que el algoritmo lo identificara erróneamente.
- Armas diseñadas para engañar los sistemas de escaneo de un aeropuerto.
- Engañamos a un sistema recomendador para que recomiende nuestros productos cuando los usuarios buscan el de la competencia.

En la siguiente figura podemos ver un ejemplo de cómo se ha conseguido engañar a AlexNet, un conocido clasificador de imágenes. En las imágenes de la izquierda el algoritmo los reconoce, pero al aplicar el ruido de la imagen central, el algoritmo clasifica todas las imágenes como avestruces.

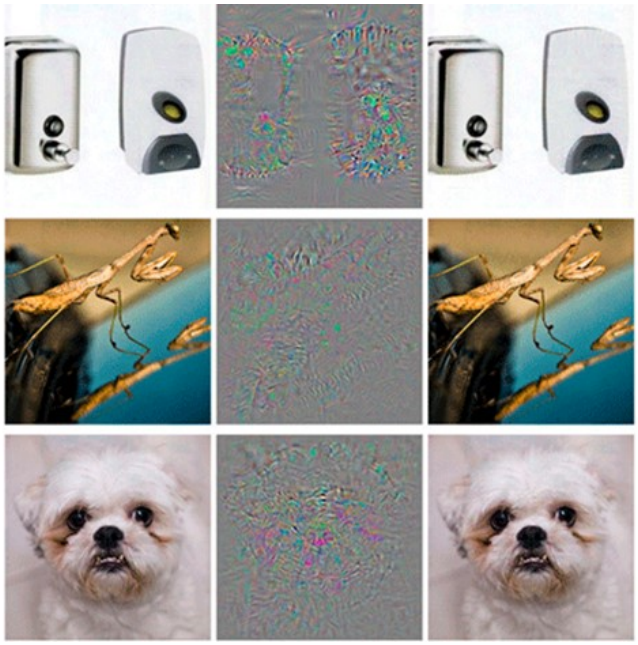


Figura 5. Algunas imágenes adversarias que engañan a Alexnet. Fuente: https://christophm.github.io/interpretable-ml-book/adversarial.html

¿Cómo podemos defendernos de estos ataques? Es complicado, pero debemos ser conscientes de ello para intentar remediarlo. Podemos mitigar el riesgo con las **siguientes medidas:**

- **Conocer a tu adversario.** Conocer cuáles son las motivaciones que un atacante tiene para hacerlo.
- **Ser proactivo,** esto es, intentar engañar constantemente a los sistemas con tus propios ejemplos de adversario para intentar detectar los posibles errores.
- **Protegerte a ti mismo** con reentrenamientos activos con adversarios o utilizar varios clasificadores y decidir en mayoría.

12.5. Explicabilidad de algoritmos

Como hemos visto en el apartado anterior, es posible engañar a los algoritmos de *machine learning* entregando ejemplos preparados para que estos se equivoquen. Si podemos explicar el modelo que se genera en el proceso de aprendizaje es más fácil evitar o predecir estos errores. Pero esto no siempre es posible, debido a que no todos los algoritmos son explicables o fácilmente explicables.

Los modelos que generan los algoritmos de *machine learning* podemos clasificarlos en función de su explicabilidad: en modelos de caja negra y modelos de caja blanca.

Los **modelos de caja negra** son aquellos en los que conseguir entender el modelo para poder analizarlo es muy complejo o casi imposible. Potencialmente todos pueden ser de caja negra si el número de parámetros de configuración es muy alto. Pero, en general, son propensos a ser algoritmos que generan modelos de caja negra las redes de neuronas, los *random forest,* el razonamiento basado en casos con medidas de distancia compleja, etc.

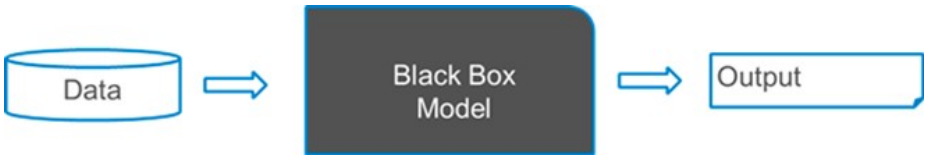


Figura 6. Modelo de caja negra.

Los **modelos de caja blanca,** por el contrario, son más sencillos de explicar y, por tanto, de analizar. Los árboles de decisión simples son algoritmos que generan modelos de cajas blancas. También las redes bayesianas o cualquier sistema que utilice lógica de predicados, con encadenamiento hacia delante o hacia atrás.

La **explicabilidad** nos da una serie de ventajas que debemos tener en cuenta como,

por ejemplo:

- **Confiabilidad.Es muy importante poder confiar en las decisiones de un algoritmo,** sobre todo si este está a cargo de algo importante. Por ejemplo, está conduciendo un vehículo por ti, tomando decisiones de compra en bolsa o manejando una central nuclear. Pero también para otros temas menores es importante saber qué decisión está tomando el algoritmo debido a que este puede incurrir en, como hemos visto durante este tema, sesgos, discriminaciones etc. Saber cómo llega a esa decisión puede ayudarnos a prevenir estos malos comportamientos de los algoritmos.

- **Adquirir nuevo conocimiento.** Los algoritmos a veces **son capaces de resolver problemas o descubrir nuevas soluciones a problemas que antes no se conocían.** Pero estos problemas muchas veces no pueden ser analizados correctamente debido a que no sabemos cómo el algoritmo ha llegado a esa conclusión. Por lo tanto, perdemos los detalles de ese nuevo conocimiento adquirido.

- **Detección de fallos**. Si el modelo tiene fallos y conocemos el modelo, podremos predecirlos, mitigarlos o reentrenarlo. Hasta ahora, solo podemos saber si un algoritmo de caja negra tiene fallos probándolo exhaustivamente. Pero siempre puede haber casos que no se han contemplado en los que el algoritmo falle. Durante el tema hemos visto varios de ellos.

Pero eso no significa que dejemos de usar algoritmos de caja negra. Hay ciertos entornos donde no hay problema por usar algoritmos de caja negra debido a que la necesidad de verificación del modelo no es crítica. Pensemos en algoritmos que controlan la calidad de imagen en un videojuego (DLSS), el algoritmo de YouTube o el recomendador de Netflix, o en algoritmos que ayuden a los científicos de la NASA a encontrar nuevos exoplanetas. Si hay algún error en estos algoritmos que clasifica un exoplaneta como planeta que no lo es, el error no es de vital importancia. Con el tiempo se descubrirá que no lo es, pero no tiene un coste de error alto y es mayor el beneficio de poder haber encontrado cientos de exoplanetas que sin la IA no

hubieran podido ser encontrados.

En resumen, hay que ser conscientes en qué dominios que el modelo no sea explicable no es un problema y que los beneficios de estos modelos compensen la falta de explicabilidad, y en qué dominios es recomendable usar algoritmos explicables, aunque obtengan peor rendimiento.

La aproximación, normalmente, a este tipo de dominios sensible es **tener soluciones híbridas entre algoritmos de caja negra y algoritmos de caja blanca.** Es decir, que parte del razonamiento esté inducido por algoritmos de caja blanca para que el cuerpo principal de la solución sea explicable y se use los algoritmos de caja negra en aquellas tareas donde la explicabilidad sea menos crítica. O también se busca intentar explicar los algoritmos de caja negra con otros métodos que ayuden a entenderlos.

12.6. Referencias bibliográficas

Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. *Boletín Oficial del Estado,* 298, de 14 de diciembre de 1999. Recuperado de https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

Orientaciones y garantías en los procedimientos de anonimización de datos personales

Agencia Española de Protección de Datos. (2016). Orientaciones y garantías en los procedimientos de anonimización de datos personales. *aepd.es.* Recuperado de: https://www.aepd.es/sites/default/files/2019-09/guia-orientaciones-procedimientos-anonimizacion.pdf

En el siguiente enlace podemos consultar las orientaciones y garantías en los procedimientos de anonimización de datos personales en la Agencia Española de Protección de Datos.

Código de buenas prácticas en protección de datos para proyectos Big Data

Sáiz, A. (coord.). (2017). Código de buenas prácticas en protección de datos para proyectos Big Data. *aepd.es.* Recuperado de https://www.aepd.es/sites/default/files/2019-09/guia-codigo-de-buenas-practicas-proyectos-de-big-data.pdf

En el siguiente enlace podemos consultar el código de buenas prácticas en protección de datos para proyectos *big data.*

Adversarial Examples

Molnar, C. (2020.). Adversarial Examples. En *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable (capítulo 6).* Recuperado de https://christophm.github.io/interpretable-ml-book/adversarial.html

Este autor nos explica lo que son los adversarial examples.

1. La Agencia Española de Protección de Datos:

A. Se limita a aconsejar al Gobierno sobre cómo debería legislarse en materia de protección de datos.

B. Tiene encargada la misión de velar por el cumplimiento de la normativa de protección de la información personal.

C. Dependiendo del Ministerio de Interior vela porque los datos de las Administraciones públicas se gestionen correctamente.

D. Depende directamente de la Unión Europea.

2. El Reglamento General de Protección de Datos:

A. Es de obligado cumplimiento desde el año 2016.

B. Se convierte en normativa de obligado cumplimiento en el año 2019.

C. Se convierte en normativa de obligado cumplimiento en el año 2018.

D. Se convierte en normativa de obligado cumplimiento en el año 2020.

3. En un contexto de anonimización y privacidad de la información, la triada hace referencia a:

A. Seguridad, consistencia y persistencia: elementos básicos de un repositorio que cumpla con la normativa oficial.

B. La combinación de los atributos edad, género y código postal y que, ocasionalmente, podrían permitir la identificación unívoca de una persona.

C. Las tres instituciones encargadas de velar por la privacidad de los usuarios, la AEPD, el Ministerio de Justicia y el Ministerio de Interior.

D. Ninguna de las anteriores.

4. A la hora de gestionar la relación con los usuarios o clientes:

A. Es preciso generar un entorno de confianza.

B. Los clientes deben percibir que reciben un beneficio por conceder permiso para trabajar con sus datos.

C. Se debe informar a los clientes de cuál es el objetivo que persigue el tratamiento de sus datos personales.

D. Todas las anteriores.

5. La anonimización de los datos personales:

A. Implica eliminar o reducir al mínimo los riesgos de identificación de los datos anonimizados.

B. Implica mantener la veracidad de los resultados tras el tratamiento de los datos.

C. Implica eliminar la referencia a la entidad.

D. Todas las anteriores.

6. Según la AEPD, el proceso de anonimización debe atender a los siguientes principios:

A. Principio proactivo, principio de veracidad, principio de privacidad objetiva, principio de plena funcionalidad y principio de privacidad.

B. Principio proactivo, principio de veracidad, principio de privacidad objetiva, principio de plena funcionalidad, principio de privacidad y principio de información y formación.

C. Principio activo, principio de veracidad, principio de privacidad objetiva, principio de plena funcionalidad, principio de privacidad y principio de información y formación.

D. Principio reactivo, principio de veracidad, principio de privacidad objetiva, principio de plena funcionalidad, principio de privacidad y principio de información y formación.

7. Una diferencia fundamental entre *tokenizar* y aplicar una función *hash* es:

A. No existe esa diferencia, son la misma cosa.

B. El proceso de *tokenización* es irreversible, lo contrario que si aplicamos una función *hash*.

C. La anonimización mediante función *hash* es irreversible y la *tokenización* no.

D. La *tokenización* tarda mucho más tiempo que aplicar una función *hash*.

8. ¿Qué podemos decir de las implicaciones de la inteligencia artificial?

A. Es posible determinar de antemano casi la totalidad de estas implicaciones, tanto para lo bueno como para lo malo.

B. El panorama es muy esperanzador puesto que solo se perciben beneficios.

C. Los problemas que se perciben a corto, medio y largo plazo superan a los beneficios.

D. La educación será esencial para conseguir que la sociedad se adapte a los nuevos puestos laborables.

9. ¿Qué son los *adversal examples?*

A. Ejemplos atípicos en los datos de entrenamiento.

B. Ejemplos manipulados sutilmente que hacen que los algoritmos de IA *fallen*.

C. Ejemplos que no son capaces de ser reconocidos por una IA pero que también engañarían a un humano.

D. Son los datos de entrañamiento que se utilizan en las Generative Adversarial Networks.

10. Un algoritmo de caja negra es aquel que:

A. Permite analizar fácilmente cómo han llegado a obtener una conclusión.

B. Es aquel en el que el análisis de cómo ha llegado a obtener una conclusión es muy complejo.

C. Son algoritmos que han aprendido sesgos y que pueden hacerlos tomar decisiones poco éticas.

D. No se usan ya que se prefiera usar algoritmos de caja blanca.