Newsletter trimestral

CÁTEDRA

iDANAE

INTELIGENCIA · DATOS · ANÁLISIS · ESTRATEGIA

4T19

Ética e Inteligencia Artificial

UNIVERSIDAD POLITÉCNICA DE MADRID

ManagementSolutions
Making things happen

Análisis de metatendencias

La Cátedra iDanae (inteligencia, datos, análisis y estrategia) en Big Data y Analytics, que surge en el marco de la colaboración de la Universidad Politécnica de Madrid (UPM) y Management Solutions, tiene el objetivo de promover la generación de conocimiento, su difusión y la transferencia de tecnología y el fomento de la I+D+i en el área de Analytics.

En este contexto, una de las líneas de trabajo que desarrolla la Cátedra es el análisis de las metatendencias en el ámbito de Analytics. Una metatendencia se puede definir como un ámbito de interés en un campo de conocimiento determinado que requerirá de inversión y desarrollo por parte de gobiernos, empresas y la sociedad en general en un futuro próximo[1].

En este informe trimestral se ha abordado la ética en el desarrollo de proyectos de inteligencia artificial.

[1]Para más información, véase iDanae, 2019.

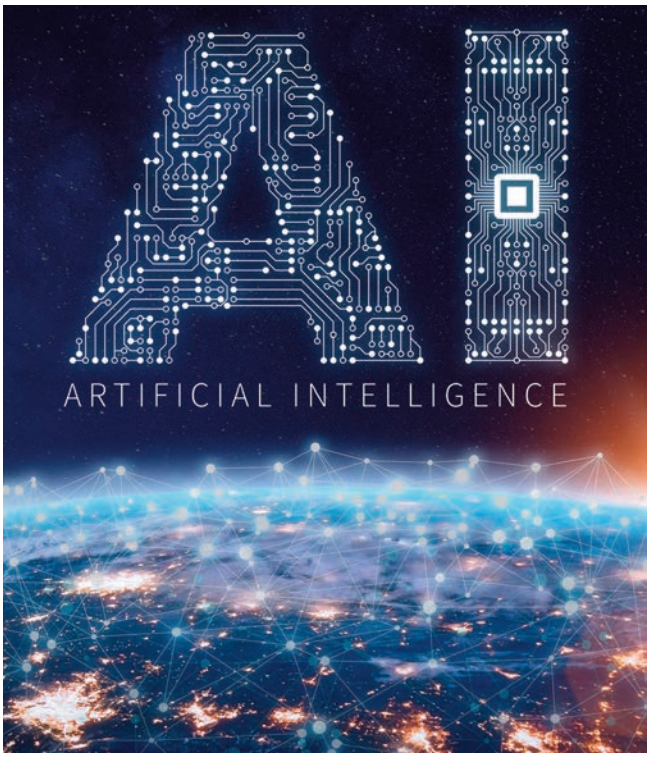
Introducción

El desarrollo de la inteligencia artificial (IA) está suponiendo una revolución en todos los sectores, especialmente en los que han transformado digitalmente sus procesos o se basan de forma nativa en procesos digitales. Los sistemas que incorporan IA van a interactuar con las personas en la toma de decisiones (e incluso sustituirlas en algunas ocasiones). Resulta, por tanto, necesario que estas decisiones se tomen en consonancia con las normas éticas establecidas en la sociedad. Por ejemplo, en Estados Unidos se utiliza desde 1998 un programa llamado COMPAS[2] que se encarga de predecir la probabilidad de que un criminal vuelva a reincidir en un futuro, empleando un algoritmo en el que se evalúan 137 parámetros del sujeto. Este sistema, que se utiliza como apoyo, revisa el historial de más de un millón de convictos, y consigue un porcentaje de acierto del 65%[3]. Este tipo de herramientas han sido objeto de revisión por distintos organismos. Por ejemplo, el NCSC[4] ha incorporado en su Guía a los Tribunales de Justicia[5] un principio que versa sobre el uso de este tipo de herramientas de análisis. Si se considera la utilización de un software como COMPAS para tomar decisiones que afectarán de forma significativa a la vida de las personas, se debe disponer de un marco ético que impida la existencia de discriminaciones.

El debate, aún abierto, sobre cómo deben incorporarse al uso de la inteligencia artificial unos principios éticos que garanticen la consonancia con los criterios humanos, está adquiriendo cada vez mayor relevancia. Esto es resultado tanto de su mayor implementación como de la aparición de errores de juicio relevantes derivados del uso de estos sistemas de decisión. Resulta, por tanto, necesario explicitar una jerarquía ética.

Este debate está presente en organismos reguladores, cristalizándose en propuestas normativas; en el ámbito empresarial, dando lugar a cambios organizativos y de gobierno; en el académico, resultando en marcos teóricos y estudios académicos; y, por último, en la sociedad civil en su conjunto, a través de actividades de divulgación, concienciación, etc.

El término de Inteligencia Artificial se puede utilizar con distintos significados. Este informe se focaliza en la IA débil, que hace referencia a la capacidad de desarrollar algoritmos que realizan tareas concretas de forma muy superior a la humana, y son especialmente efectivos encontrando patrones y relaciones entre los datos que utilizan para aprender, pero sin capacidad de incluir en su análisis criterios humanos que no se incluyan explícitamente[6]. En este caso, donde existe capacidad analítica pero desligada de criterios típicamente humanos como la moral, la sabiduría o la compasión, es donde el tratamiento ético se convierte en relevante en todo lo



relacionado con el dato: desde su obtención hasta su uso, pudiendo incluir el almacenamiento.

No se abordan en este documento otros conceptos, como el de IA fuerte (hace referencia a una máquina que tenga una inteligencia similar a la humana[7], incluyendo la capacidad de interpretar sentimientos y emociones), o el de Inteligencia superior (conocida como singularidad[8], una inteligencia superior a la humana, capaz de aprender por sí misma).

En los debates sobre este ámbito se plantea si tiene sentido hablar de ética con relación a la inteligencia artificial, bajo la hipótesis de que un algoritmo no tiene capacidad de pensar[9].

[2] COMPAS, 2015.
[3] Sam Corbett-Davies, 2016.
[4] National Center for State Courts (NCSC, 2011).
[5] Ibídem.
[6] Coppin, 2004.
[7] Searle, 1980.
[8] Chalmers, 2010.
[9] Una cuestión diferente, fuera del objeto de este artículo, es si los ordenadores pueden pensar, si pueden tener consciencia y, en función de la respuesta a estas preguntas, si deben tener una ética.

Por tanto, ¿qué significa exactamente ética de la inteligencia artificial? ¿Es la que deben practicar desde sus valores los sistemas inteligentes (algoritmos, robots, etc.) o es la que el ser humano debe aplicar al desarrollar y utilizar sistemas inteligentes? ¿Quién es el responsable cuando una máquina se equivoca o cuando alguien toma una decisión errónea basada en un sistema inteligente?

Este informe se centra en la IA débil y en la ética con consecuencias jurídicas actuales o potenciales. En particular se analiza el proceso de desarrollo de un proyecto de IA para poder comprender las responsabilidades a lo largo del mismo, considerando la relación que esta responsabilidad puede tener con las consecuencias legales. Por ejemplo, ante un sistema automático que ayuda en el proceso de admisión de un préstamo, en el asesoramiento sobre la contratación de un fondo de pensiones, o en la decisión sobre la aptitud de una persona para un puesto de trabajo, surge la siguiente pregunta: ¿quién es el responsable ante una decisión errónea? Podrían plantearse distintas responsabilidades: i) la persona que hace caso de la recomendación del sistema; ii) la persona que decidió adquirir y usar dicho sistema; iii) la persona que comercializa el sistema; iv) quien programó el algoritmo; v) la persona que decidió usar datos que podrían estar equivocados; etc.

En algunos casos puede existir una consecuencia legal, incluso aunque el desarrollador no haya pensado en ello, pudiendo llegar a tener un impacto económico no calculado.

Estos sistemas están en uso en la actualidad. El desarrollo de la digitalización incrementa el número de acciones y, por tanto, de usuarios que pueden estar potencialmente afectados por decisiones incorrectas. Por eso es necesario dar respuesta a la pregunta: ¿quién es el responsable?

Para desarrollar una respuesta a esta pregunta se han de analizar las fases de desarrollo de un proyecto de IA y comprender los aspectos éticos que pueden surgir en cada una de ellas:

1. **Establecimiento del objetivo del proyecto**. En esta fase es necesario analizar si el objetivo es ético (lo que es común a todo proyecto, no solo en el ámbito de la inteligencia artificial).

2. **Obtención de los datos**. En esta fase es necesario analizar el proceso de obtención de los datos, incluyendo la propiedad de los datos, la privacidad, la recolección y el uso de la información.

3. **Construcción del modelo**. En esta fase nuevamente se han de analizar varios aspectos:

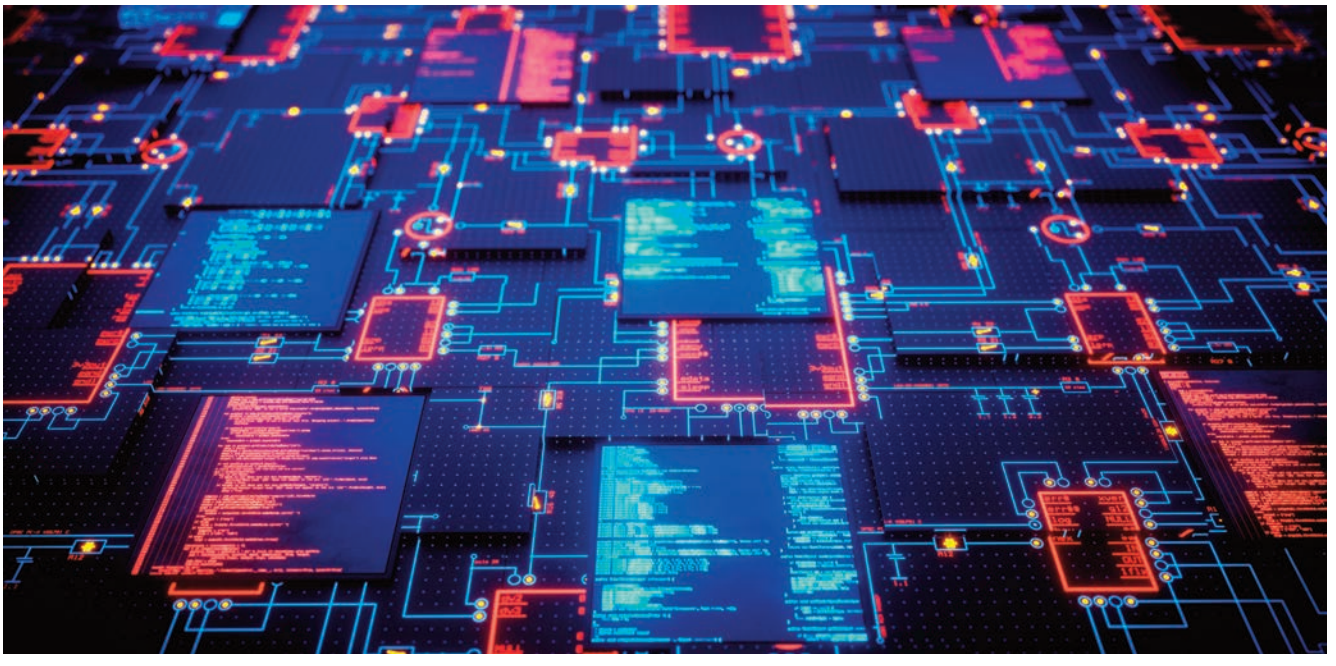
    a. Los posibles sesgos del modelo.

    b. Los errores del modelo.

    c. La transparencia e interpretabilidad del modelo generado.

4. **Responsabilidad en la implantación y uso de los modelos**. Finalmente se ha de analizar la responsabilidad en la implantación y la comunicación de los modelos de cara a su posterior uso.

Con el objetivo de aportar una visión general sobre este contexto, en este artículo se revisan los cuatro ámbitos descritos bajo la perspectiva de la ética con consideraciones legales (sin tratar otros posibles aspectos de la ética), y se finaliza con una visión sobre las iniciativas y los marcos normativos que tratan de abordarlos.



Ética en el desarrollo de proyectos de inteligencia artificial

Objetivo del proyecto

La determinación del objetivo del proyecto en el que se aplicará inteligencia artificial para su consecución es uno de los puntos en los que surgen dilemas éticos, en especial en aquellos aspectos en los que este sistema sustituye al juicio humano. Las preocupaciones a este respecto no surgen solo por los posibles errores que puedan cometer estos sistemas, que se comentarán más adelante, sino por la reducción de la capacidad de decisión del ser humano en ámbitos donde no existe una forma clara en términos éticos de resolver el problema. Un caso típico es la generación de sistemas de inteligencia artificial aplicados al diagnóstico de enfermedades o al análisis de interacciones entre medicamentos, donde se ha de decidir si el sistema sustituye el diagnóstico o la prescripción del médico, o si solo se utiliza como una guía de apoyo, y la responsabilidad se mantiene en el facultativo. Otro ejemplo conocido es la posible priorización de víctimas que debería llevar a cabo el sistema de conducción de un coche autónomo[10].

Obtención de los datos

La ética de los datos hace referencia a la rama de la ética encargada de evaluar los problemas morales relacionados con los datos, los algoritmos y las prácticas relacionadas[11]. Esta parte de la ética incluye el análisis de las dimensiones morales de la información[12]: la privacidad, la anonimización, la transparencia, la veracidad y la responsabilidad en relación a los procesos relacionados con los datos (captura, transformación, análisis y uso)[13]. Se centra por tanto en los problemas éticos que plantea la recopilación y el análisis de grandes conjuntos de datos y en cuestiones que van desde el uso de grandes datos en la investigación biomédica y las ciencias sociales[14], hasta la publicidad[15], entre otros.

Una cuestión de interés se basa en el análisis sobre la necesidad de desarrollar una ética de los datos, que incluya si los actuales conceptos de privacidad[16], de trato justo y no discriminatorio, trazabilidad (con normativas como GDPR), así como el resto de posibles normas aplicables, como son las relativas al honor o el derecho a la imagen, son suficientes. Por ejemplo, Strava hizo públicos ciertos datos de localización de carreras, lo cual reveló la localización de bases militares en Iraq y Afganistán debido a los entrenamientos realizados[17]. La propia GDPR[18] recoge en su artículo 22 lo siguiente: 'todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar'. Por tanto, la normativa actual no solo recoge elementos relativos meramente a la recolección y almacenamiento de datos, sino que también alude a su modelización y uso. Asimismo, existen datos que, aun siendo de carácter privado o íntimo, pueden estar disponibles a efectos de modelización. ¿Resulta ético utilizar esos datos para tomar decisiones que podrían variar el trato dado a una persona? Por ejemplo, en línea con lo anterior, el Consejo de Europa emitió una recomendación que incluye la propuesta de prohibir que las aseguradoras pudieran calcular riesgos futuros de sus clientes basados en test genéticos[19].

Finalmente, conviene señalar tres aspectos relacionados con los datos que pueden confundirse entre ellos pero que poseen matices distintos: la propiedad, la recolección y el uso de los datos.

- **Propiedad de los datos:** de manera independiente a dónde se almacenen, los derechos de propiedad de los datos pueden pertenecer a distintos agentes. La normativa GDPR se posiciona a favor de los usuarios en lugar de las empresas que hacen uso de los datos.

- **Recolección:** se refiere a las formas de recolección de los datos, de manera independiente de la propiedad o publicidad que le den los usuarios. En este sentido una pregunta habitual es la siguiente: ¿es ético que los servicios recolecten datos si el usuario no lo desea, aunque estos sean públicos? Los ejemplos típicos hacen referencia a los datos de navegación o información compartida por redes sociales.

- **Uso de los datos:** se refiere a los posibles usos derivados de la explotación de datos, lo que incluye el posible valor generado para las empresas y los gobiernos. Las preguntas que surgen son similares a las del punto anterior. Por ejemplo, ¿es necesario detallar el uso que se dará a un conjunto de datos? Un ejemplo es el experimento que se llevó a cabo por el servicio de Salud Pública de los Estados Unidos en colaboración con la Tuskegee University en

[10] Jean-François Bonnefon, 2016.
[11] Floridi, L., y Taddeo, M. (2016).
[12] Algunos autores hablan de las 5 C's: consentimiento, claridad, consistencia y veracidad, control y transparencia, y consecuencias (véase DJ Patil, 2018).
[13] Íbidem.
[14] Mittelstadt BD, 2016.
[15] Hildebrandt, 2008.
[16] Incluido el concepto de medición de la privacidad en la red, entendido como la observación de websites y servicios para detectar, caracterizar y cuantificar comportamientos impactados por la privacidad (Steven, E., y Narayanan, A. 2016).
[17] DJ Patil, 2018.
[18] GDPR, 2018.
[19] European Council, 2016.

Alabama sobre la sífilis no tratada, donde a los hombres afroamericanos se les comunicó solamente que recibían atención médica gratuita[20]. Para mitigar este tipo de situaciones, la Comisión Europea[21] indica que se ha de informar al participante en una investigación sobre el uso concreto que se dará a sus datos. Asimismo, se plantea si el valor generado por los datos debe repercutirse a los usuarios o puede retenerse por las empresas y agentes que los utilizan. Por ejemplo, el gobernador del estado de California propuso en su discurso inaugural una ley que obligue a las empresas cuyo modelo de negocio se base en la recolección y uso de datos a pagar un dividendo a los usuarios que han generado dichos datos[22].

Construcción del Modelo

Posibles sesgos del modelo

Existen múltiples sesgos[23] en la construcción y el uso de un sistema de inteligencia artificial. Entre otros: (i) la existencia de datos de entrenamiento incompletos (debido a que no se dispone de atributos o características de la población relevantes para el modelo) o no equilibrados (debido a que los atributos o la variable predicha no esté convenientemente representada), o bien (ii) posibles sesgos que pueden provenir del uso de ciertas técnicas utilizadas en el proceso de entrenamiento del modelo. Si no se acomete ningún tratamiento, el resultado del modelo podría estar sesgado, arrojando falsos positivos o falsos negativos que afectan a la posterior toma de decisiones. Esto puede darse en distintos tipos de algoritmos, como los modelos de reconocimiento facial, algoritmos para la captación dentro de la función de Recursos Humanos, o la evaluación del desempeño. Uno de los problemas derivados de los sesgos es que su tratamiento puede ser incompatible con la obtención de un modelo que optimice el poder predictivo dada una misma muestra. Es por esto por lo que un tratamiento adecuado de los sesgos del modelo resulta fundamental para evitar, por un lado, que se pierda capacidad predictiva, y, por otro lado, que se realicen discriminaciones innecesarias o que vulneren los derechos de la persona.

Un ejemplo que sirve para ilustrar el aspecto (i) de este problema es el modelo desarrollado por Amazon para contratar a nuevos empleados. Se descubrió que el modelo estaba sesgado, penalizando las candidaturas realizadas por mujeres. Esto ocurrió debido a que se entrenó al modelo con los currículos recibidos por Amazon en los diez años anteriores, en los que la mayoría de los perfiles eran de hombres, haciendo que el modelo interpretase que ser mujer era una característica desfavorable a la hora de buscar un perfil técnico[24]. Este es un error surgido por unos datos de entrenamiento no equilibrados, en este caso por contener un sesgo histórico.

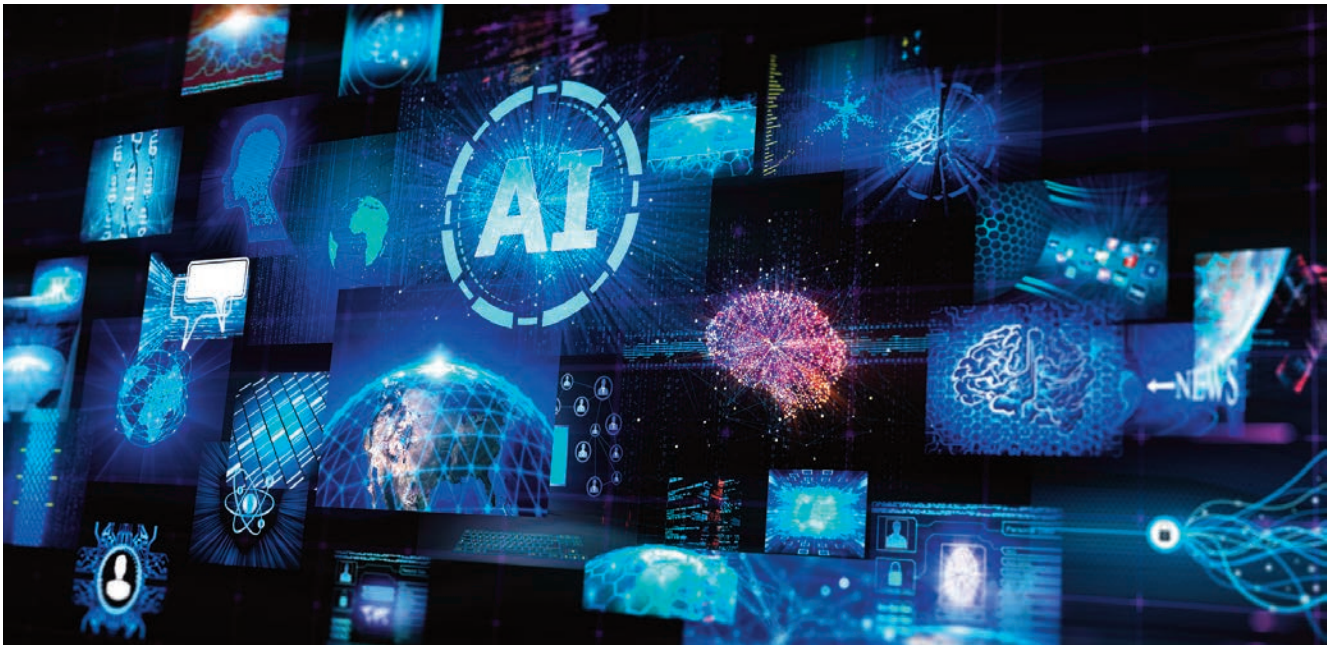
Un ejemplo del aspecto (ii) podría surgir tras la sentencia del Tribunal de Justicia de la Unión Europea[25] en el que se establece que la diferenciación de las primas de los seguros para los hombres y para las mujeres, exclusivamente por razones de género, es incompatible con la Carta de los Derechos Fundamentales de la UE. Sin embargo, en el caso de los seguros de automóvil se observa que estadísticamente las mujeres tienen menos accidentes que los hombres. Un modelo de machine learning podría acabar discriminando entre hombres y mujeres, pese a que en los datos de entrenamiento no se especifique el sexo del cliente, utilizando variables que se encuentren correlacionadas con el sexo, lo que resultaría en un resultado sesgado a través del tratamiento de dichos datos.

Errores del modelo

Los modelos de inteligencia artificial tienen de forma inherente un porcentaje de error. Al extender este uso a decisiones de elevada importancia, por ejemplo, como soporte en decisiones judiciales, de seguridad o salud, y donde no es posible corregir a posteriori el error, ¿es ético utilizar estos modelos, aun sabiendo que cometen errores?

Por un lado, es necesario analizar la pertinencia de utilizar estos algoritmos, aun cuando mejoren la ratio de acierto frente a los sistemas tradicionales, dado que pueden esconder tasas de error mayores para determinadas subpoblaciones, con el consiguiente trato injusto para ellas. Por otro lado, el renunciar a su uso puede generar un coste de oportunidad. En este contexto, aunque existen fórmulas para realizar revisiones por ambos sistemas, la decisión final arrojará falsos positivos y negativos en cualquier caso. Un ejemplo paradigmático de esta disyuntiva se encuentra en el campo de la salud: por un lado, un falso positivo puede provocar un tratamiento innecesario, lo que acarrea un proceso doloroso para el paciente y un gasto para el sistema; por otro, un falso negativo podría conllevar retrasar un tratamiento necesario y reducir su tasa de éxito. Además, la aplicación de este tipo de modelos, que requieren grandes cantidades de datos para poder aprender y encontrar patrones, resulta problemática en situaciones en las que dichos datos escasean, como por ejemplo en el caso de las enfermedades raras. No obstante, el análisis de los errores cometidos por los algoritmos y la pertinencia de su implementación debe también realizarse teniendo en consideración si dichos errores reducen el error cometido por los humanos en el desarrollo de la misma tarea: es importante considerar si hay que exigir que el modelo tenga un porcentaje de acierto del 100% o si es suficiente con que se reduzcan los errores cometidos por las personas, aunque ello conlleve el asumir la existencia de errores.

[20] Brandt, 1978
[21] European Commision, 2018
[22] Newsom, 2019.
[23] Se entiende por sesgo un peso desproporcionado a favor o en contra de una cosa, persona o grupo en comparación con otra, generalmente de una manera que se considera injusta.
[24] Dastin, 2018.
[25] European Commission, 2012.



Transparencia y trazabilidad

La mayor complejidad de muchos de los modelos de inteligencia artificial, así como su uso, a veces autónomo o a veces como soporte en la toma de decisiones, pone de manifiesto la importancia de entender cómo funciona el modelo y para qué debe usarse. Una de las principales razones es la propia necesidad de interpretar y supervisar cómo funciona el modelo para determinar si este es ético (en particular para modelos no replicables), no solo en términos del resultado, sino también en términos de cómo se llega a esta conclusión[26]. Por otro lado, la transparencia sobre el funcionamiento del modelo puede llegar a determinar las posibles limitaciones de su construcción o implementación y, por tanto, también en su uso.

Sin embargo, el derecho de acceso a la información sobre los modelos está sujeto a ciertos límites legales, lo cual puede ser un impedimento para lograr la interpretabilidad o la transparencia. Esto responde a distintos motivos: preservar el secreto profesional o la propiedad intelectual, no poner en riesgo la seguridad nacional, garantizar la confidencialidad requerida en ciertos procesos como los judiciales o sanitarios, o no entorpecer la investigación de delitos. En este contexto, existen algunas iniciativas basadas en la posibilidad de restringir el acceso a la información y que esta sea responsabilidad de organismos reguladores específicos en los casos en los que se justifique su relevancia. En este caso son fundamentales tanto el grado de entendimiento que puede llegar a considerarse como razonable por parte de los agentes que lo utilicen, así como los instrumentos que un organismo regulador pueda utilizar para llevar a cabo su responsabilidad de supervisión.

Responsabilidad en la implantación y uso de los modelos

A medida que aumenta el uso de la IA existen más situaciones en las que es necesario determinar la responsabilidad de las decisiones tomadas. Se han de incorporar los valores éticos en los desarrollos tecnológicos en inteligencia artificial, y se ha de tratar cómo se enfoca la respuesta de estos desarrollos ante cuestiones que tienen un impacto ético[27].

Esto a su vez es un impulsor para que esta responsabilidad se fije explícitamente tanto en los marcos jurídicos como en los acuerdos comerciales. La IA puede cometer errores que causen daños, o puede ser utilizada de forma maliciosa pese a su correcto funcionamiento, siendo necesario en ambos casos el adecuado establecimiento de la responsabilidad. Esta necesidad es especialmente relevante en las situaciones en las que el impacto de la aplicación de la IA tiene una gran repercusión, como por ejemplo cuando se aplica en el sector sanitario, en el financiero o el judicial.

Un ejemplo que ilustra lo anterior es el accidente de tráfico que produjo la primera víctima mortal involucrando un vehículo autónomo[28]. Surgió como consecuencia un debate sobre quién debería responsabilizarse del accidente: si el constructor del coche, la empresa que había desarrollado e instalado el sistema de conducción autónoma o la persona que supervisaba el funcionamiento del sistema y que debería haber frenado.

Conforme a lo anterior, el origen de la responsabilidad puede radicar en cualquiera de las fases descritas en un proyecto de desarrollo de un modelo de inteligencia artificial, e incluso en algunos casos la responsabilidad puede derivarse de decisiones de varias de estas etapas.

A pesar de que este tipo de retos éticos no son nuevos y por tanto no son consecuencia de la aplicación de algoritmos en la toma de decisiones, la necesidad en algunos casos de explicitar en la programación qué decisiones se tomarían en casos extremos, así como la aplicación a gran escala de un solo criterio en sustitución de los comportamientos individuales derivados del libre albedrío, determina un nuevo aspecto de análisis.

[26]Para más información, véase iDanae, 2019.
[27]Dignum, 2018.
[28]Laris, 2018.

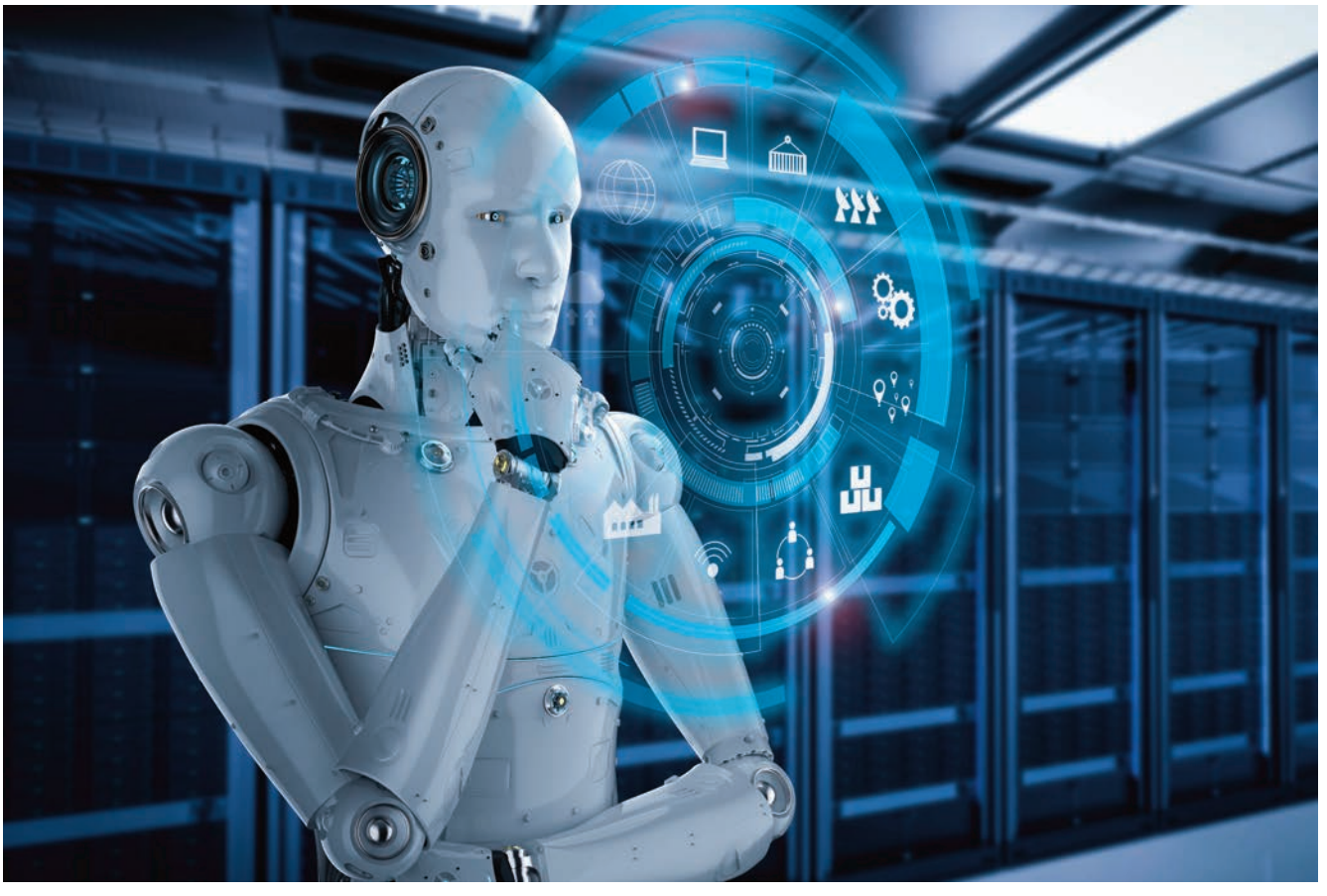
Marcos normativos

Por todo lo anterior, los distintos reguladores, así como la sociedad civil y los distintos agentes involucrados, se plantean establecer unos principios éticos en sus distintos niveles de actuación que regulen tanto la forma en la que se desarrollan estos algoritmos como su utilización.

En este sentido, en la actualidad conviven distintos marcos y propuestas. Desde las más básicas, donde estos principios se tratan como una extensión de la privacidad o como un conjunto de criterios de trato justo, a otros basados en principios que deben cumplir todos estos algoritmos.

Iniciativas públicas y privadas

En el ámbito de la ética en la inteligencia artificial, el mundo empresarial y los gobiernos han llevado a cabo diversas iniciativas que abogan por regular algunas cuestiones, o al menos establecer ciertos principios, que alimenten regulaciones relacionadas con el uso de la IA en la toma de decisiones. En el primer caso, las empresas están optando por el refuerzo de las funciones de validación, auditoría y control del riesgo de modelo, poniendo especial énfasis en la interpretabilidad y explicabilidad a lo largo de todo el ciclo de vida del modelo (desarrollo, implementación y uso) y los datos utilizados (origen, almacenamiento y uso), así como la propia percepción de los usuarios. Este mismo sentido es el que se da en materia de gobierno corporativo, que se cristaliza en el refuerzo de la función de control a través de mayores peticiones de información, así como en una mayor responsabilidad de la alta dirección y, por último, en la inclusión en los comités donde se tratan cuestiones de modelos de figuras independientes. Todo ello permite que tanto las decisiones directamente automáticas como las que se toman utilizando algoritmos de machine learning tengan un mayor grado de conocimiento y responsabilidad en la entidad.



Ya desde 1993, la Association for Computing Machinery[29] hizo público un código ético y de conducta profesional dirigido a cualquier persona que use las tecnologías de la computación de forma que tenga un impacto (profesionales, instructores, estudiantes, etc.). En este código ético se incorporan principios éticos generales (honestidad, confiabilidad, no discriminación, privacidad, etc.), así como un conjunto de principios y responsabilidades de los profesionales de la computación. Aunque a nivel internacional aún no hay un consenso o estándar establecido, existen diversas iniciativas tanto públicas como privadas.

Tres de las iniciativas privadas actuales más destacadas son (i) el Future of Life Institute, conjuntamente con los participantes de la conferencia de Asilomar[30], desarrolló en enero de 2017 los Principios de Asilomar[31], donde se establecen unas directrices para el desarrollo de una IA ética, abordando los problemas habituales en la investigación (el objetivo con el que se desarrolla, quién lo financia, cooperación entre grupos de investigación y evitar la competitividad entre equipos), los problemas éticos (seguridad, transparencia, responsabilidad, privacidad, control humano, respeto por la sociedad y trazabilidad, entre otros) y los problemas a largo plazo (limitaciones, posibles riesgos, capacidades, importancia y limitación de la automejora); (ii) el IEEE[32] ha desarrollado una iniciativa sobre la ética de los sistemas autónomos e inteligentes; o (iii) el consorcio Partnership on AI[33] tiene como propósito establecer buenas prácticas y educar al público general sobre la IA.

Por su parte, los gobiernos y la sociedad civil están planteando diversas iniciativas que se soportan en cuestiones como la aprobación antes de la implementación, la auditoría por parte de organismos externos o la limitación del uso de modelos en determinados ámbitos.

Más concretamente, se están desarrollando iniciativas gubernamentales por parte de i) Estados Unidos[34], donde se abarcan todo tipo de impactos de la IA en la sociedad: en la innovación, en la industria y para los trabajadores; ii) Reino Unido[35], donde se define un marco de Data Ethics y se desarrolla un buen gobierno de las tecnologías basadas en el uso de datos; o iii) la Comisión Europea[36], donde se definen unos principios en el tratamiento de la información de las propuestas de investigación que incluyen el uso de datos personales, y se establecen unos principios a seguir para tener una IA ética basados en el respeto a los derechos humanos y unas palancas de vinculación a procesos. Otros organismos también han realizado análisis, como el Foro Económico Mundial[37], que ha establecido una serie de problemas complementarios, como son la posible generación de desempleo, los sesgos anteriormente mencionados o posibles problemas de ciberseguridad. Resulta de especial interés la propuesta de una futura Agencia Europea de Robótica e Inteligencia Artificial[38] que se encargaría de detectar las áreas de trabajo con potenciales riesgos, como pueden ser asuntos relacionados con la salud o de interés público.

[29] ACM, 2018.
[30] En Asilomar se han realizado diversas conferencias sobre estándares éticos de diversas disciplinas científicas, como por ejemplo sobre el diseño de genes y organismos vivos en 1975 o sobre los futuros riesgos de la IA en 2009.
[31] Asilomar, 2017.
[32] El IEEE es una organización técnica profesional dedicada al avance de la tecnología en beneficio de la humanidad. Véase IEEE, 2019.
[33] Este consorcio está compuesto por un grupo multidisciplinar de investigadores y académicos, y por empresas como Apple, Amazon, Accenture, Baidu, Facebook, Google, IBM, Intel, McKinsey, Microsoft, Nvidia, PayPal, Salesforce, Sony, Samsung o Unicef, entre otros. Véase Partnership on AI, 2016.
[34] USA Government, 2019.
[35] UK Government, 2018, y UK Government, 2019.
[36] European Comission, 2018, y European Comission, 2019.
[37] Bossman, 2016.
[38] European Parliament, 2016.

Conclusiones

En este informe se ha analizado la ética con consideraciones jurídicas actuales o potenciales. La evolución de la IA conlleva la necesidad de plantear y acometer problemas éticos que surgen debido a su desarrollo y uso, de cara a encontrar posibles soluciones. Esta necesidad no está cubierta, en parte por la dificultad de esta materia, y en parte por la velocidad de desarrollo de la IA. No obstante, al tratarse de sistemas que van a tener una influencia en la vida y forma de actuar de las personas, es necesario que las decisiones que se tomen a partir de las salidas de estos algoritmos se hagan en consonancia con las normas éticas establecidas en la sociedad. La digitalización de todos los sistemas y la posibilidad de capturar datos de todos los procesos va a potenciar de forma exponencial el efecto de las decisiones basadas en sistemas de IA.

Por otra parte, es importante resaltar que la interpretabilidad de los modelos obtenidos con IA tiene un impacto en los casos de usos. Por tanto, la interpretabilidad de los modelos y los posibles casos de usos está directamente relacionada.

Entre los dilemas más importantes actualmente se encuentran los relacionados con el objetivo de la IA, el uso de datos, la generación de sesgos y errores, así como los problemas que surgen a la hora de entender el proceso de decisión y la responsabilidad de dicha decisión. Es por tanto necesario establecer unos principios éticos y leyes que regulen la forma en la que se desarrollan y utilizan estos algoritmos. La responsabilidad de este desarrollo debe ser compartida tanto por los individuos y las empresas como por gobiernos y reguladores.

Bibliografía

ACM. (2018). Code of Ethics and Professional Conduct. Association for Computing Machinery Council. Retrieved from https://www.acm.org/code-of-ethics

Asilomar. (2017). Future of Life. Retrieved from Future of Life: https://futureoflife.org/ai-principles/

Bossman, J. (2016). Foro Económico Mundial. Retrieved from Foro Económico Mundial: https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

Brandt, A. M. (1978). Racism and Research: The Case of the Tuskegee Syphilis Study. The Hastings Center Report, 21-29.

Chalmers, D. (2010). The singularity: a philosophical analysis. Journal of Consciousness Studies .

COMPAS. (2015). Practitioner's Guide to COMPAS Core. Retrieved from Practitioner's Guide to COMPAS Core: http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Coppin, B. (2004). Artificial Intelligence Illuminated .

Dastin, J. (2018). Reuters. Retrieved from Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. Ethics Inf Technol 20, 1–3. doi:10.1007/s10676-018-9450-z

DJ Patil, H. M. (2018). Ethics and Data Science. O'Reilly Media, Inc.

European Comission. (2012). EU rules on gender-neutral pricing in insurance industry enter into force. European Comission.

European Comission. (2018). Ethics and data protection. European Comission.

European Comission. (2019). Policy and investment recommedations for trustworthy AI. European Comission.

European Council. (2016). Council of Europe. Retrieved from Council of Europe: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016806b2c5f

European Parliament. (2016). Directorate-General for Internal Policies. Retrieved from Directorate-General for Internal Policies: https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU%282016%29571379_EN.pdf

Floridi, L., y Taddeo, M. (2016). What is data ethics? Phil. Trans. R. Soc. A374: 20160360. Retrieved from http://dx.doi.org/10.1098/rsta.2016.0360

GDPR. (2018). GDPR. Retrieved from GDPR: http://www.privacy-regulation.eu/es/22.htm

Hildebrandt, M. (2008). Profiling and the identity of the European citizen. Springer.

iDanae. (2019). iDanae. Retrieved from iDanae: https://blogs.upm.es/catedra-idanae/newsletter-trimestral-3t19/

IEEE. (2019). IEEE. Retrieved from IEEE: https://ethicsinaction.ieee.org/

Jean-François Bonnefon, A. S. (2016). The Social Dilemma of Autonomous Vehicles. Science.

Laris, M. (2018). The Washington Post. Retrieved from The Washington Post: https://www.washingtonpost.com/transportation/2018/12/20/nine-months-after-deadly-crash-uber-is-testing-self-driving-cars-again-pittsburgh-starting-today/

Mittelstadt BD, F. L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Science and Engineering Ethics.

NCSC. (2011). NCSC. Retrieved from NCSC: https://www.ncsc.org/~/media/Microsites/Files/CSI/RNA%20Guide%20Final.ashx

Newsom, G. (2019). Office of Governor. Retrieved from Office of Governor: https://www.gov.ca.gov/2019/02/12/state-of-the-state-address/

Partnership on AI. (2016). Partnership on AI. Retrieved from Partnership on AI: https://www.partnershiponai.org/about/

Sam Corbett-Davies, E. P. (2016). The Washington Post. Retrieved from The Washington Post: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Searle, J. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences.

Steven, E., y Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis. SIGSAC.

UK Government. (2018). Data Ethics Framework. Department for Digital, Culture Media & Sports.

UK Government. (2019). Centre for data ethics and innovation. Retrieved from Centre for data ethics and innovation: https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation

USA Government. (2019). White House. Retrieved from White House: https://www.whitehouse.gov/ai/

Autores

Ernestina Menasalvas (UPM)

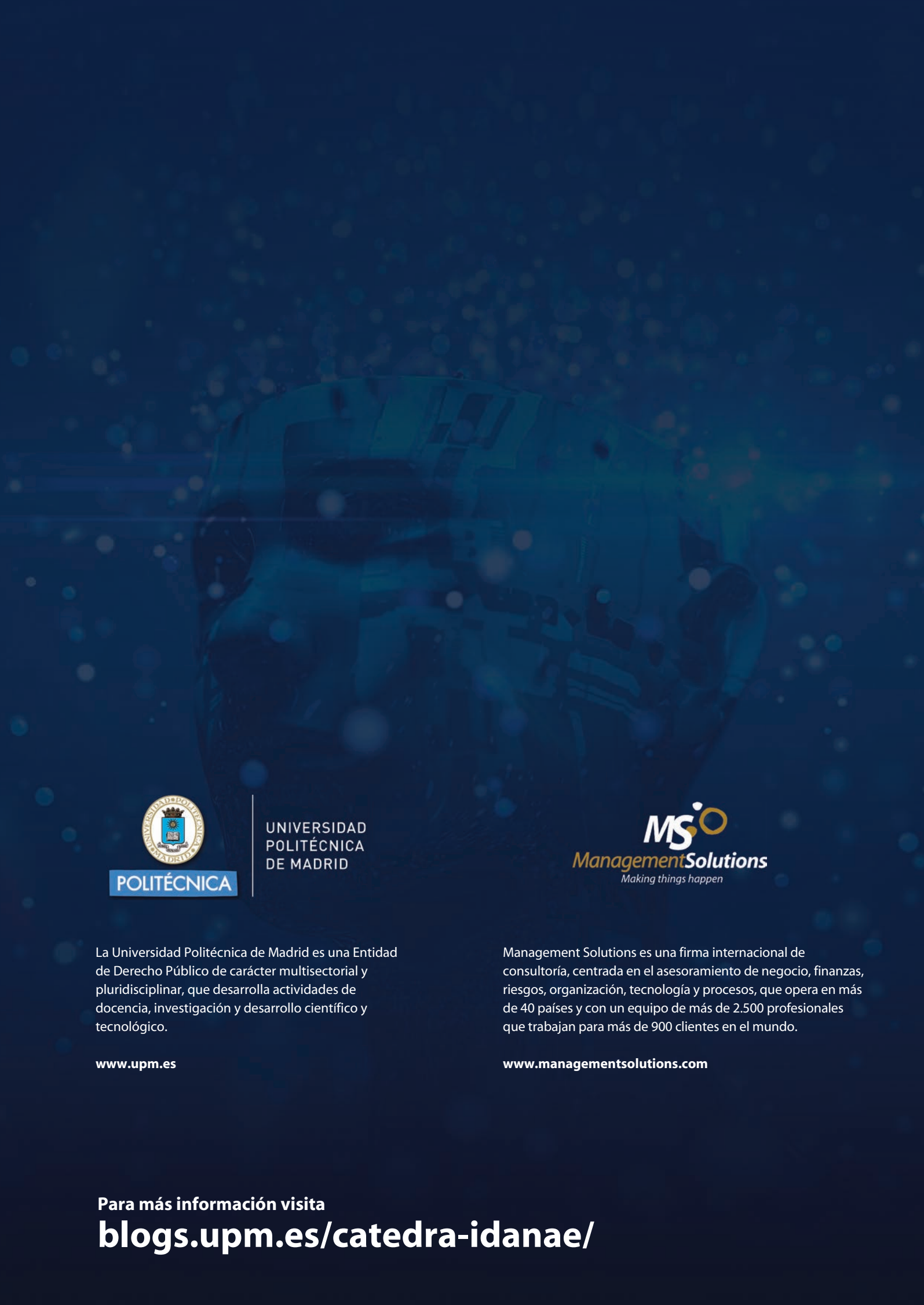
Alejandro Rodríguez (UPM)

Manuel Ángel Guzmán (Management Solutions)

Daniel Ramos (Management Solutions)

Carlos Alonso (Management Solutions)

David Castaño (Management Solutions)



UNIVERSIDAD POLITÉCNICA DE MADRID

La Universidad Politécnica de Madrid es una Entidad de Derecho Público de carácter multisectorial y pluridisciplinar, que desarrolla actividades de docencia, investigación y desarrollo científico y tecnológico.

www.upm.es

Management Solutions es una firma internacional de consultoría, centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos, que opera en más de 40 países y con un equipo de más de 2.500 profesionales que trabajan para más de 900 clientes en el mundo.

www.managementsolutions.com

Para más información visita

blogs.upm.es/catedra-idanae/