REAL ACADEMIA DE CIENCIAS
EXACTAS, FÍSICAS Y NATURALES

DISCURSO INAUGURAL

DEL AÑO ACADÉMICO 2021-2022

LEÍDO EN LA SESIÓN CELEBRADA EL DÍA 29 DE SEPTIEMBRE DE 2021
POR EL ACADÉMICO NUMERARIO

EXCMO. SR. D. DAVID RÍOS INSUA

SOBRE EL TEMA

LUCES Y SOMBRAS DEL BIG DATA Y LA INTELIGENCIA ARTIFICIAL

MADRID
DOMICILIO DE LA ACADEMIA
VALVERDE, 22 - TELÉFONO 917 014 230
www.rac.es
2021

ISSN: 1138-4093
ISBN: 978-84-87125-74-4
Depósito Legal: M-25630-2021

ÍNDICE

Excmo. Sr. Presidente,
Excmas. Sras. Académicas, Excmos. Sres. Académicos, Sras. y Sres.:

1. INTRODUCCIÓN

En esta última década, venimos siendo testigos del rápido crecimiento en las capacidades de muchas organizaciones para explotar los avances en las tecnologías de la información, de la modelización estadística y de la investigación operativa (IO), para recopilar y procesar datos sociales, de mercado y de operaciones y así apoyar sus procesos de toma de decisiones. La captura de datos a través de aplicaciones online, de sensores y de móviles produce cantidades ingentes de información que podemos aprehender para entender cómo actuamos, nos sentimos, movemos e interactuamos y cómo respondemos frente a las políticas de los gobiernos y las decisiones de las empresas. Esto está implicando que, cada vez más, la información proporcionada por los datos constituya la base de las decisiones, posibilitando procesos más automatizados y conduciendo a servicios y productos más personalizados.

Buena parte del interés reciente por el Big Data y la Inteligencia Artificial se debe a los éxitos alcanzados frente a tareas hasta hace bien poco muy complejas que hoy en día se han convertido casi en una capacidad común. Un buen ejemplo es la traducción automática, complicadísima hasta hace poco, pero ya con excelentes resultados y casi alcanzando el nivel humano. En juegos como el ajedrez, el póquer o el go, los sistemas desarrollados superan ya al mejor de los humanos. En algunos campos como el márketing personalizado, los sistemas de ventas son en algunos casos capaces de identificar nuestras necesidades incluso antes de que nosotros mismos las conozcamos. También están posibilitando la emergencia de tecnologías disruptivas como los vehículos autónomos, que cambiarán el mundo tal y como hoy lo conocemos. En nuestro país, un proyecto de la Agencia Estatal de Seguridad Aérea (AESA), junto con nuestra Real Academia (RAC), ha redefinido globalmente cómo se gestionan los riesgos en seguridad aérea, obteniendo enormes ahorros anuales al Estado.

Un pilar esencial en estos desarrollos ha sido el acceso a fuentes de datos masivos anotados, por ejemplo imágenes anotadas con su contenido. Muchos de esos datos tienen naturaleza diferente a la habitualmente tratada en Estadística e Informática hasta hace bien poco: no son estructurados; contienen imágenes y texto; son de alta dimensionalidad;... Todo esto motivó desarrollos matemáticos nuevos en áreas como el procesamiento distribuido del lenguaje natural, el tratamiento de imágenes, la optimización, la inferencia causal, la simulación o la inferencia bayesiana. Además, se pudo aprovechar las capacidades de procesamiento de nuevo hardware como GPUs y TPUs.[1] Finalmente, se han venido publicando de forma abierta modelos pre-entrenados por expertos que se han podido trasladar a nuevos dominios de aplicación, en lo que hoy se denomina aprendizaje por transferencia.

---

[1] Graphics Processing Units y Tensor Processing Units.

Como resultado, la analítica de negocios se ha convertido en un campo floreciente para la consultoría empresarial. Sin embargo, aunque muchas decisiones de algunos gobiernos a menudo vienen apoyadas con métodos tradicionales del análisis de políticas públicas, incluyendo aproximaciones como el análisis de coste-beneficio. Pocos departamentos y agencias gubernamentales han logrado, por el momento, aprovechar de forma sistemática las grandes masas de datos disponibles y los métodos avanzados de la estadística y del aprendizaje automático para obtener evidencias que informen sus decisiones. Este hecho constituye una interesante novedad desde una perspectiva histórica, ya que los métodos cuantitativos de ayuda a la toma de decisiones han surgido frecuentemente del sector público. Por ejemplo, la estadística social, que se remonta a Quetelet, se inició en el siglo XIX para apoyar a los gobiernos promoviendo la idea de que las regularidades estadísticas aportan señales sobre realidades sociales. Del mismo modo, la IO nació durante la Segunda Guerra Mundial al servicio de las fuerzas armadas de los Estados Unidos de América y del Reino Unido, y creció rápidamente a partir del desarrollo de métodos de apoyo a la toma de decisiones en problemas militares.

En la industria, el énfasis en el área de la Analítica se ha puesto en resolver problemas relacionados con el análisis de conjuntos de datos masivos y complejos, frecuentemente, en entornos muy cambiantes, más allá de la evolución y el desarrollo de sistemas de planificación empresarial y los almacenes de datos convencionales. Tales conjuntos suelen denominarse Big Data. Así, disponemos de una cantidad cada vez mayor de información digitalizada, proveniente de dispositivos y sensores cada vez más baratos. Nos enfrentamos pues a una nueva era en la que hay una enorme cantidad de datos digitalizados, el *mar de datos*, sobre numerosos temas de interés potencial para una empresa o un gobierno. Sin embargo, con bastante frecuencia, dicha información es altamente desestructurada y difícil de gestionar y, no inusualmente, poco relevante, por aportar poco valor.

El análisis de estos tipos de datos no estructurados, o no muestreados, constituye un reto importante en la industria y ha dado lugar a nuevos paradigmas como la Ciencia y la Ingeniería de Datos. Los datos no estructurados difieren de los que lo son en que su formato es muy variable y no pueden almacenarse en bases de datos relacionales tradicionales sin un esfuerzo significativo que conlleve transformaciones complejas de datos. Se emplean así bases de datos NoSQL más escalables. La gestión de tales masas de datos requiere marcos que permitan realizar cálculos sobre grandes cantidades de datos basados en el procesamiento distribuido sobre conjuntos más pequeños. Finalmente, se necesitan también infraestructuras de almacenamiento que posibiliten el resumen y análisis de los datos.

Además de los avances tecnológicos, también hay nuevas clases de métodos de análisis que permiten la extracción de información. Estos van más allá de las técnicas tradicionales (como los modelos de regresión; los de series temporales, basados por ejemplo en modelos dinámicos lineales; o los clasificadores de k-vecinos más cercanos) llegando a métodos más recientes (como los árboles de clasificación y de regresión, las máquinas de soporte vectorial o, muy especialmente, las redes neuronales profundas en sus distintas versiones).

En cualquier caso, los datos parecen ahora más accesibles a los gestores, que tienen una gran oportunidad para tomar mejor sus decisiones empleándolas para aumentar sus ingresos, reducir sus costes, mejorar el diseño de sus productos, detectar y prevenir el fraude, o la mejora de la conversión de clientes a través del marketing personalizado. Esto ha conducido a un nuevo concepto de organización que toma decisiones basadas en la evidencia, con ejemplos claros como Google, Facebook, Amazon, Walmart, Alibaba y algunas de las líneas aéreas más avanzadas.

Nuestro objetivo aquí es ilustrar a través de diversos proyectos en los que hemos estado involucrados en los últimos años, el enorme potencial de estas metodologías y tecnologías para resolver problemas de carácter social, pero también mostrar algunas cuestiones que deben afrontarse para su adecuada utilización. Previamente fijaremos algunas ideas que nos ayudarán a entender mejor la discusión posterior, concluyendo con algunas recomendaciones de futuro.

2. UN PEQUEÑO MAPA CONCEPTUAL

Presentaremos aquí brevemente algunos conceptos básicos que se emplearán en las Secciones 3 y 4. Pueden verse detalles adicionales en Rios Insua y Gómez-Ullate (2019).

Big Data. Aunque es esencialmente un afortunadísimo nombre marketiniano (algo abusado y quizá ya algo desgastado), el término Big Data permite resumir las nuevas categorías de datos a las que nos tenemos que enfrentar contemporáneamente, en tres direcciones principales.

*Conjuntos de datos de muy alta dimensión (**V**olumen)*, con numerosos individuos y/o numerosas variables que no caben en memoria al ser procesados. Como ejemplos, Walmart acumula más de 2,5 petabytes por hora de transacciones de clientes; Facebook recoge 350 millones de fotos por día y 4 millones de likes por minuto. Por contra, en algunos dominios como en electrofisiología, a lo sumo podemos estudiar una decena de casos, de los que podemos extraer millones de observaciones en el tiempo.

*Conjuntos de datos muy diversos, muchas veces desestructurados (**V**ariedad)*, con textos (que pueden provenir de fuentes tales como mensajes de twitter), imágenes, likes, datos geoespaciales, lecturas de sensores en una ciudad, señales GPS en teléfonos móviles, vídeos, ... e ¡¡¡incluso números!!![2]

*Conjuntos de datos generados dinámicamente con alta frecuencia (**V**elocidad)*, que exigen procesamiento a gran velocidad para apoyar decisiones muy rápidamente. En muchos casos, la velocidad de generación tiende a ser más importante que el volumen, en el sentido de que se deben tomar decisiones en tiempo real, teniendo que evaluarse previamente la información, a partir de los datos obtenidos, también en tiempo real.

Pero, al final de un proyecto, deben convertirse en *conjuntos de datos de los que hemos de obtener valor (**V**alor)*. Para ello, las herramientas matemáticas, estadísticas y computacionales tradicionales resultan insuficientes requiriendo nuevos desarrollos.

Ciencia de datos. En este tsunami científico-técnico, surge la ciencia de datos para unificar y expandir la estadística, el análisis de datos, la IO, el aprendizaje automático y sus métodos relacionados, a efectos de comprender y analizar fenómenos reales, en una intersección que esencialmente emplea técnicas y teorías extraídas de las matemáticas, la estadística (clásica y bayesiana), las ciencias de la computación, así como los conocimientos propios del área a la que pertenece el problema que queremos resolver.[3]

Inteligencia Artificial. La IA puede definirse como la capacidad de un sistema para interpretar correctamente datos externos, aprender de los mismos y emplear tales conocimientos

---

[2] En cualquier caso, los datos se convierten típicamente en números. Por ejemplo, las palabras se suelen transformar con incrustaciones, tras preprocesamiento, en vectores.

[3] Recordemos, en cualquier caso, que el fin último debería ser la toma de decisiones en problemas de interés, como iremos revisando en distintos ejemplos.

para lograr tareas y metas concretas a través de su adaptación flexible.[4] Tras su acuñación en 1956, y varios vaivenes intelectuales, nos encontramos en un nuevo período de esplendor, quizá ya definitiva, de esta disciplina.

Aprendizaje automático y aprendizaje profundo. Remarquemos, sin embargo, que no toda la IA es aprendizaje automático; ni todo el aprendizaje automático es aprendizaje profundo. Pero éste es el grupo de métodos que, por su capacidad de tratar datos masivos con éxito en tareas de aprendizaje supervisado (regresión y clasificación), no supervisado (conglomerados, estimación de densidades, detección de atípicos, ...) y por refuerzo; ha puesto a la IA en su lugar central actual.

Redes neuronales. Las redes neuronales son modelos que han sufrido también vaivenes en la historia de la IA hasta su relevancia actual. Motivados por un, a veces exagerado, paralelismo biológico, una red neuronal se interpreta como un conjunto de neuronas o unidades que reciben entradas, las combinan linealmente y las transforman mediante una función de activación no lineal. De sus combinaciones obtenemos distintas arquitecturas.

En la anterior oleada de popularidad de la IA, se usaban los perceptrones con una capa oculta: los datos se ingresaban por una capa de entrada a la capa oculta de neuronas que procesaban tales datos para predecir las salidas en problemas de regresión, clasificación y predicción.

En la actualidad, se habla de redes profundas cuando hay varias capas ocultas. Algunos avances técnicos (y el redescubrimiento, y posterior refinamiento, del algoritmo estocástico de descenso del gradiente (SGD) (Robbins y Monro, 1951)) han posibilitado su uso en el tratamiento de problemas complejos. Entre estas nuevas redes profundas se encuentran las *redes convolutivas*, que están detrás de los avances en procesamiento de imágenes subyacentes a los sistemas de percepción esenciales, por ejemplo, en los sistemas de conducción automatizados (ADS).

Analítica. La Analítica puede centrarse en enfoques descriptivos, predictivos o prescriptivos. Típicamente, apoya el descubrimiento y la presentación de patrones relevantes en problemas con grandes conjuntos de datos registrados para cuantificar, describir, predecir y mejorar los resultados de una organización. A menudo, combina métodos de la estadística, la IO, el aprendizaje automático y la informática, junto con disciplinas como la sociología, la psicología y la economía. La evidencia proporcionada por los datos se emplea para recomendar acciones y guiar las decisiones y la planificación en las organizaciones. Sus resultados pueden emplearse como entrada a la toma de decisiones por personas, pero también pueden alimentar sistemas automáticos de ayuda a decisión.

---

[4] Las tareas principales se referirían pues al aprendizaje y la toma de decisiones en condiciones de incertidumbre, para las que la Teoría de la Decisión Estadística aporta herramientas fundamentales, algo no siempre conocido (ni reconocido).

3. BD E IA. PRIMERO ALGUNAS LUCES...

Comenzaremos presentando algunos ejemplos que muestran el enorme potencial de las tecnologías y metodologías relacionadas con el Big Data y la Inteligencia Artificial a la hora de resolver problemas de carácter social. Todos ellos se refieren a proyectos recientes en los que hemos estado involucrados que han tenido impacto positivo y van asociados a los denominados objetivos de desarrollo sostenible (ODS).[5] Previamente mostramos un ejemplo sobre el tipo de aplicaciones que disparó el interés global por estos métodos. En cada caso comenzamos por una motivación y concluimos con algunas lecciones aprendidas.

3.1 En el principio era el...

Los éxitos de las redes convolutivas (con nuevas funciones de activación como RELU, la implantación del SGD y aprovechando la paralelización masiva en GPUs) en competiciones de reconocimiento de imágenes marcan la vuelta al interés por la exploración intensa de los modelos de aprendizaje profundo. Estos permean el enorme interés por estos campos. Presentamos además la capacidad que pueden tener los métodos bayesianos en este contexto.

Describimos una tarea de clasificación de imágenes con una red convolutiva ya estándar, como es la VGG-19.[6] Mostramos su superioridad frente a modelos no convolutivos y modelos no profundos en el conjunto de datos CIFAR-10 (Krizhevsky et al., 2014), que incluye 60.000 imágenes en color de tamaño 32x32, referidas a 10 clases de objetos. Específicamente, comparamos:

- un modelo estándar de regresión multinomial,
- una red neuronal completamente conexa con tres capas ocultas de 200 nodos y funciones de activación ReLU, y
- el mencionado modelo VGG-19.

Consideramos su análisis bayesiano, empleando distribuciones normales independientes sobre todos los parámetros, lo cual nos dará ventajas sobre los estimadores de máxima verosimilitud habitualmente empleados.[7] Todos los modelos se entrenan en las mismas condiciones (200 épocas, minilotes de 128 muestra, algoritmo SGD y predicciones basadas en esperanzas predictivas según una muestra de la distribución a posteriori de tamaño 100).

El Cuadro 1 muestra los resultados globales. Obsérvese como el modelo lineal funciona mucho peor (precisión en test del 38%) mientras que el aumentar la flexibilidad de los modelos,

---

[5] Véase https://unstats.un.org/sdgs/indicators/indicators-list/.

[6] Disponible, por ejemplo, en Keras https://keras.io/.

[7] Los éxitos de las redes convolutivas (con nuevas funciones de activación como RELU, la implantación del SGD y aprovechando la paralelización masiva en GPUs) en competiciones de reconocimiento de imágenes marcan la vuelta al interés por la exploración intensa de los modelos de aprendizaje profundo. Estos permean el enorme interés por estos campos. Presentamos además la capacidad que pueden tener los métodos bayesianos en este contexto.

mejora la precisión de forma notable. Así, aunque el modelo de red neuronal básico funciona algo mejor, se obtienen excelentes resultados con VGG-19.

CUADRO 1: RESULTADOS SOBRE EL CONJUNTO TEST DE CIFAR-10 (Gallego y Rios Insua, 2022)

MODELO	PRECISIÓN EN TEST
Lineal	38,10%
MLP	50,03%
VGG-19	93,29%

La Figura 1 presenta las predicciones realizadas sobre cinco imágenes del conjunto test. Este comportamiento de las redes convolutivas (con las que se llega a mejorar las capacidades humanas en reconocimiento de imágenes) motivaron el interés creciente por el aprendizaje profundo hasta los niveles actuales.

Figura 1. Cinco ejemplos de CIFAR-10 y predicciones con la arquitectura VGG. Obsérvese el error al confundir un avestruz (clase ave) con un gato (Gallego y Rios Insua, 2022)

Incidamos, para concluir, que, más allá del placer de hacer buenas predicciones o clasificaciones de imágenes, típicamente estos algoritmos se emplearán para ayudar a tomar decisiones. Como ejemplo, tales algoritmos son componentes importantes en el diseño de sistemas de seguridad en accesos y en ADS, cuestión sobre las que volveremos despúes. Por el momento, indiquemos que un problema importante es la clasificación adecuada de los objetos identificados. Los errores de clasificación en la capa de percepción son heredados por los algoritmos en la capa de predicción, lo que aumenta la probabilidad de predicciones incorrectas del comportamiento de un objeto y una evaluación inexacta del entorno del ADS. Para minimizar el riesgo de accidentes, el sistema debe ser muy sensible y reaccionar de manera segura cuando existe incertidumbre sobre la situación. Lamentablemente, estos sistemas son propensos a identificaciones de emergencia con falsos positivos que conducen a reacciones innecesarias. La aproximación bayesiana propuesta tiene la ventaja de asignar incertidumbre sobre las predicciones, con lo que se hacen más robustas las correspondientes decisiones. Sin embargo, en problemas complejos requiere estrategias y capacidades computacionales aún por descubrir.

- Los métodos de aprendizaje típicamente empleados se basan en máxima verosimilitud (tal vez con un regularizador) que conducen a estimadores puntuales sin prestar apenas atención a la estimación de la incertidumbre asociada.
- El potencial de los métodos bayesianos en aprendizaje profundo, como revela este ejemplo, aún debe alcanzarse plenamente por medio de nuevos algoritmos y herramientas computacionales.

3.2 Mejores decisiones para un mundo más sano

El ODS 3 de la Agenda 2030 se refiere a *asegurar las vidas saludables y promover el bienestar para todos a todas las edades*. La pandemia actual ha permitido explorar el potencial de los métodos de la IA en tareas de descubrimiento de fármacos, así como para ayudar a tratar y gestionar enfermedades, aunque The Alan Turing Institute (2021) da una visión algo más escéptica. Aquí se describe cómo se emplea la IA en el tratamiento de un problema global de salud más importante, aunque probablemente menos urgente.

Las enfermedades cardiovasculares (ECV) son la causa principal de mortalidad en Europa (45% de todas las muertes). Sus costes anuales asociados, sólo en Europa, se estiman en 210 billones de euros (Wilkins et al., 2017). En la hora aproximada de duración de esta conferencia morirán por esta causa alrededor de 440 personas en nuestro continente. Por comparación, por cáncer fallecerán unas 220 personas en el mismo periodo.

Describiremos aquí brevemente cómo el tratamiento de grandes bases de datos con métodos de IA permite realizar contribuciones relevantes que mejoran el entendimiento y, muy especialmente, el tratamiento de una enfermedad cardiovascular. Los datos provienen de los reconocimientos médicos anuales de trabajadores afiliados con una compañía privada de seguros entre los años 2012 y 2016, convenientemente anonimizados y securizados.[8] Tales datos se complementan con información del censo, basada en el código postal de vivienda de los individuos. A partir del mismo, somos capaces de obtener de bases de datos públicas el status socio-económico y el nivel educativo medio correspondiente.[9] Finalmente, un intensísimo trabajo de depuración de outliers, duplicados, datos mal registrados, valores faltantes y transformación de variables, proporciona un conjunto de datos estructurado y completo sobre el que emplear modelos estadísticos y de aprendizaje automático. Creamos así, a partir de un conjunto crudo de datos (con casi cinco millones de casos y 40 variables) mal formado e incompleto, una base bien estructurada y limpia.[10]

---

[8] Ilustramos así como los datos de tenedores privados, correctamente utilizados, permiten fomentar el bien común. El documento https://www.ine.es/normativa/leyes/cse/papel_estadistica_ oficial.pdf del Consejo Superior de Estadística, elaborado bajo la coordinación de esta RAC, proporciona información detallada sobre esta cuestión.

[9] Ilustramos así mecanismos de cruces de bases de datos que añaden valor, pero también plantean algunos problemas como veremos en la Sección 4.

[10] El preprocesamiento de datos es un proceso complejo y poco lucido, del que se habla poco, pero que resulta esencial para hacer descubrimientos interesantes y crear valor en la sociedad.

Esta se convierte, de hecho, en una base de datos probabilística en forma de red bayesiana (Castillo et al., 2012), donde tanto su estructura como sus tablas de probabilidad en los nodos se aprenden a partir de los datos constituyendo una aproximación no supervisada de aprendizaje automático. De nuevo, de forma importante y novedosa, reconocemos la incertidumbre sobre las estimaciones de probabilidades con una aproximación bayesiana basada en hiperdistribuciones sobre los parámetros de las distribuciones en los nodos. Una ventaja de esta aproximación es que posibilita predicciones para cada grupo de factores a partir de las capacidades de la red para realizar inferencia probabilística, empleándose como parte de una herramienta de ayuda a la toma de decisiones facilitando el diagnóstico, el tratamiento y la adopción de políticas de salud pública.

Este proyecto, en particular, pone el énfasis en la evaluación predictiva de los distintos factores de riesgo CV, y en especial en la actividad física, e integra factores como *depresión*, *duración del sueño* y *status socioeconómico*. Específicamente, las variables que finalmente se incluyeron en el modelo fueron: *Factores de riesgo cardiovascular (FRCV) no modificables* (Sexo v_1, Edad v_2, Nivel educativo v_3, Nivel socioeconómico v_4); *FRCVs modificables* (Indice de masa corporal v_5, Actividad física v_6, Duración del sueño v_7, Historial fumador v_8, Ansiedad v_9, Depresión v_{10}); *Condiciones médicas* (Hipertensión v_{11}, Hipercolesterolemia v_{12}, Diabetes v_{13}).

Para construir la red, empleamos un proceso en dos pasos. En el primero, se usó un algoritmo de búsqueda bayesiana para aprender una estructura inicial a partir de los datos. En el segundo, se implementó un proceso iterativo en el que se mostraba a expertos diversas estructuras pidiéndoseles eliminar o añadir arcos relevantes razonando a partir de un mecanismo de inferencia referido a nodos progenitores y nodos hijos. Este proceso condujo a la estructura final reflejada en la Figura 2. Como consecuencia, el modelo probabilístico subyacente resulta ser

$$\begin{aligned} p(v_1,\ldots,v_{13}) = &\Big[p(v_1)p(v_2)p(v_3|v_1,v_8)p(v_4|v_1,v_2,v_3,v_5,v_6,v_8)\Big]\times \\ &\Big[p(v_5|v_2,v_6,v_8)p(v_6|v_1,v_2,v_7,v_8)p(v_7|v_2)p(v_8|v_1,v_2) \\ &p(v_9|v_1,v_7,v_{10},v_{11})p(v_{10}|v_1,v_3)\Big]\times \\ &\Big[p(v_{11}|v_5,v_6,v_7)p(v_{12}|v_1,v_2,v_3,v_6,v_7,v_8)p(v_{13}|v_1,v_2,v_6)\Big]. \end{aligned}$$

Una vez con la estructura se asignan las tablas de probabilidad de los nodos con modelos multinomial-Dirichlet, partiendo de distribuciones a priori uniformes (French y Rios Insua, 2000). Como ejemplo, el Cuadro 2 presenta los valores esperados[11] de la distribución del nodo *duración del sueño*, que sugiere un deterioro del mismo con la edad, lo que se corrobora mediante contrastes de hipótesis (en su versión bayesiana).[12]

[11] Expresaremos las probabilidades como porcentajes.

[12] Pueden verse detalles sobre contraste de hipótesis bayesiano en Girón (2021).

CUADRO 2: PROBABILIDADES EN EL NODO v_7 COMO VALORES DE LAS DISTRIBUCIONES A POSTERIORI.						
SDUR/EDAD	**18-24**	**24-34**	**34-44**	**44-54**	**54-64**	**64-74**
< 6	4	7	10	14	17	18
6 – 9	95	93	90	86	83	81
9	1	0.0	0.0	0.0	0.0	0.0

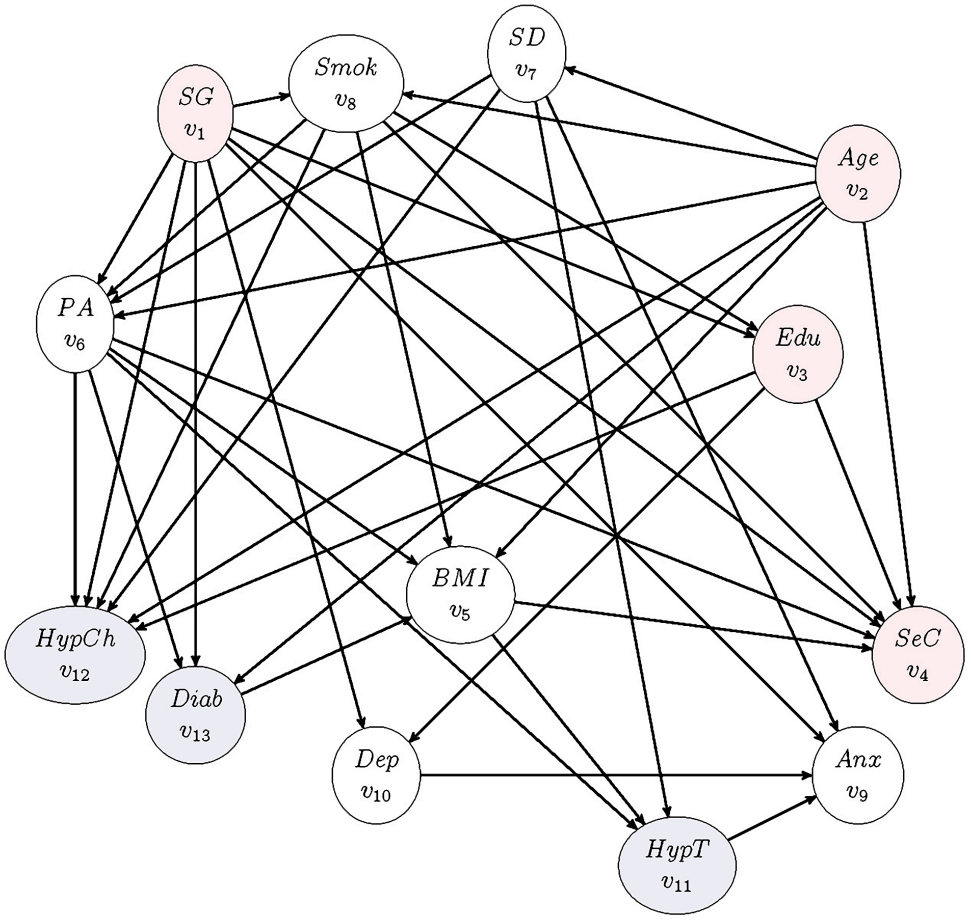

Figura 2: Red adoptada. Blanco (FRCV modificables); Rosa (FRCV no modificables); azul (condiciones médicas). (Rios Insua et al., 2021b)

Basados en esta red, describimos algunos usos en terapia y política de salud pública, aprovechando su capacidad para incorporar información, mediciones y observaciones y propagarlas a través de la red y modificar las distribuciones en los otros nodos empleando, tal vez varias veces, la fórmula de Bayes (Nielsen y Jensen, 2008). Se presenta un ejemplo en el Cuadro 3. Específicamente, consideramos el impacto de la actividad física (inactivo (1), insuficientemente activo (2), regularmente activo (3)) en la segunda columna: obsérvese el impacto positivo de la misma en un grupo de pacientes con sobrepeso. Además, empleando tales casos

como referencia, apreciamos en la tercera columna el impacto de la variable sexo en las probabilidades (siendo el resto de rasgos comunes): en todos los casos, las mujeres tienden a tener menor probabilidad de desarrollar hipertensión, pero vuelve a percibirse el impacto positivo de la actividad física.

CUADRO 3: PROBABILIDAD DE DESARROLLAR HIPERTENSIÓN DADAS LAS DISTINTAS CONDICIONES PARA EDAD ENTRE 45 Y 54, BAJO NIVEL DE SUEÑO, SOBREPESO Y ANSIEDAD.

ACT. FIS	HOMBRE	MUJER
1	25.69	22.90
2	22.87	20.29
3	20.06	19.78

Podemos también hacer hipótesis sobre evidencias relacionadas con varias cuestiones referidas a salud. Por ejemplo, consideremos el impacto del status socio-económico sobre las condiciones de salud reflejadas en el Cuadro 4. La última columna despliega la prevalencia de la hipertensión en los tres grupos sociales (7.14%, nivel 3; 7.30%, nivel 2; 7.49%, nivel 1) lo que sugiere mayor prevalencia al empeorar el nivel socio económico (y análogamente para la hipercolesterolemia y la diabetes, e inversamente para la ansiedad, pero no así para la depresión).

CUADRO 4: PROBABILIDAD DE CONDICIONES DE SALUD DADO EL STATUS SOCIO-ECONÓMICO.

STATUS	ANSIEDAD	DEPRESIÓN	DIABETES	HIPCOL	HIPTENS
3	2.75	0.45	3.21	29.75	7.14
2	2.70	0.49	3.38	30.64	7.30
1	2.61	0.49	3.58	31.07	7.49

Una vez determinado un caso de interés (grupo o individuo) y evaluada la probabilidad correspondiente, podemos encontrar los hechos influyentes sobre tal afirmación. Para ello, eliminamos los sucesos condicionantes de uno en uno y determinamos el impacto sobre la probabilidad de interés, evaluando el hecho más influyente como el que produce la mayor reducción en probabilidad. Esto es especialmente relevante para los FRCV modificables para los que podemos explorar la mejor modificación y sugerirla al grupo o individuo de interés. Consideremos como caso básico, un varón con edad (44, 54), nivel educativo y socioeconómico 3, que tiene sobrepeso, inactivo, no fumador, tiene ansiedad pero sin depresión. Su probabilidad de desarrollar hipertensión es del 38.30%. El Cuadro 5 contiene la probabilidad cuando uno de los FRCVs mejora a sus niveles deseables y los restantes se mantienen en los niveles de referencia. Como consecuencia, parece que la intervención más efectiva sería imponer una dieta para retornar a un IMC normal. Obsérvese que si fúesemos capaces de mejorar las cuatro condiciones, la probabilidad se reduciría al 3.70%.

CUADRO 5: PROBABILIDAD DE HIPERTENSIÓN DADAS CONDICIONES MEJORADAS PARA HOMBRE CON EDAD (44, 54), STATUS EDUCATIVO Y SOCIOECONÓMICO 3, SOBREPESO, AF BAJA, ANSIEDAD PERO SIN DEPRESIÓN.

FRCVM	NIVEL	PROBABILIDAD
IMC	Normal	9.76
AF	Normal	27.73
Dormir	Normal	29.14
Ansiedad	No	23.96

Recúerdese, sin embargo, que para decidir la mejor recomendación necesitaríamos tener en cuenta los posibles impactos de las condiciones médicas y los tratamientos a través de funciones de utilidad y utilidades esperadas.[13] Se ilustra esta idea en los restantes ejemplos.

- Los tenedores privados de datos pueden contribuir al mejor desarrollo de la sociedad, facilitando el intercambio de datos B2G.
- El cruce de bases de datos privadas y públicas crean valor para la sociedad, aunque también conlleva riesgos como se describe en la Sección 4.2.
- De nuevo, apréciese la relevancia de los métodos bayesianos.

3.3 Mejores decisiones para un mundo más seguro

Uno de las metas del noveno ODS se refiere a *desarrollar infraestructuras de calidad fiables, sostenibles y resilientes, para apoyar el desarrollo económico y el bienestar*. Resulta esencial garantizar un alto nivel se seguridad del transporte, en particular del transporte aéreo. La AESA ha aplicado con éxito una metodología innovadora de análisis de riesgos y apoyo a la toma de decisiones, desarrollada junto a la RAC, para mejorar la seguridad aérea en España.

La Organización de Aviación Civil Internacional (OACI) persigue que la aviación sea el modo de transporte más seguro por ser un factor clave en el desarrollo sostenible de las naciones. Sin embargo, el crecimiento del tráfico aéreo y la mayor competición entre aerolíneas ha hecho que mantener y mejorar los excelentes niveles actuales de seguridad esté resultando cada vez más exigente. Aunque la pandemia del coronavirus ha reducido notablemente tal tráfico, se espera que crezca de nuevo en un futuro cercano.

[13] Una vez construida la red, implementar los cálculos es relativamente sencillo con software como GeNIe o R!. Esto facilita construir un sistema de ayuda a la decisión que sugiera la mejor recomendación teniendo en cuenta los impactos a través de la función de utilidad.

Globalmente, la aviación aérea opera a un nivel de seguridad muy alto. Por ejemplo, en Europa, la tasa media de accidentes fatales fue de 1.3 por cada 10 millones de vuelos, aunque las agencias tratan de mejorar permanentemente este complejo sistema.[14] Para ello, los países deben desarrollar un Plan Estatal de Seguridad Aérea (SSP) para la aviación civil, que debe incluir los objetivos nacionales de seguridad y afecta a todas las partes interesadas (autoridades, proveedores de servicio, aeropuertos,...). Los planes deben identificar las fuentes principales de inseguridad y el conjunto de acciones para mitigar y controlar los riesgos asociados a las mismas.

Hasta hace relativamente poco tiempo, la gestión de riesgos en seguridad aérea (por ejemplo, para construir los SSP) se ha basado en el uso matrices de riesgos, a pesar de tener varios fallos bien conocidos (Cox, 2008).[15] Para superarlos, AESA y la RAC desarrollaron una metodología más rigurosa que facilita la asignación óptima de recursos, haciendo que los eventos que comprometen la seguridad aérea sean menos frecuentes y, en caso de que se materialicen, sean menos dañinas. En su contexto, AESA gestiona 88 tipos de sucesos de seguridad (desde salidas de pista hasta fallos de motor, pasando por colisiones en tierra) con cinco niveles de severidad (decreciente): (1) accidente, lo que conlleva la destrucción de aeronaves o la muerte de pasajeros o de tripulantes; (2) incidente serio; (3) incidente mayor; (4) incidente significativo; y, finalmente, (5) ocurrencia sin impacto en la seguridad. Además, se consideran cuatro categorías de aeronaves, en función de su peso. Por ejemplo, podemos hablar de un fallo de motor de severidad 3 con un avión de categoría 2. Algunos ejemplos se muestran en la Figura 3.

Igualmente, un análisis previo, que incluyó una revisión de la literatura y una tormenta de ideas con altos ejecutivos de AESA, condujo a la identificación de ocho consecuencias relevantes para la seguridad operacional de la aviación en nuestro país: (1) muertes asociadas al funcionamiento del sistema de aviación; (2) lesiones menores y (3) lesiones graves; (4) retrasos y (5) cancelaciones asociadas a los incidentes; (6) operaciones de mantenimiento y (7) reparaciones; y, finalmente, (8) pérdida de imagen-país. De nuevo, además de las excelentes bases de datos disponibles en la organización, se requirieron cruces con bases de datos externas a través de técnicas de web scraping, así como un intenso trabajo de armonización de datos.

El objetivo final de la metodología era encontrar la asignación de recursos que optimizase la seguridad operacional de la aviación nacional, reduciendo así, en la medida de lo posible, las ocurrencias de los distintos tipos de severidad, su gravedad y las consecuencias resultantes. Esto conlleva predecir el impacto de la cartera de recursos de seguridad (principalmente, tiempo de inspección) a implementar sobre la tasa y la gravedad de los distintos tipos de

---

[14] Como referencia, en España, antes de la pandemia, conllevaba gestionar más de 2.5 millones de movimientos aéreos anuales en 50 aeropuertos, 200 campos de aviación, 44 aerolíneas y 7000 aeronaves, además de empresas de diseño y fabricación de aeronaves. Además, se producían unos 30000 sucesos de los que, aproximadamente, 40 eran de severidad 1.

[15] A pesar de ellos, se emplean, además de en seguridad aérea, con bastante (demasiada) asiduidad en ciberseguridad, en seguridad nacional o seguridad de riesgos laborales, entre muchos otros dominios.

ocurrencia, así como de sus impactos y, despúes, encontrar la cartera óptima de seguridad, aquella que maximiza la utilidad esperada. Tales actividades conllevaron la construcción de casi dos mil modelos de predicción (no triviales) por lo cual fue necesaria desarrollar un proceso de automatización de determinación y ajuste de modelos, así como desarrollar modelos novedosos de predicción.

Figura 3: Algunos ejemplos de accidentes (Elvira et al., 2020).

Para facilitar la implementación de la metodología introducida, se diseñó el sistema RIMAS. Su valor potencial se verificó comparando los resultados de seguridad reales basados en sus recomendaciones con los que habrían resultado del mantenimiento de las políticas tradicionales de seguridad. El uso de RIMAS proporcionó un rendimiento significativamente mejor en términos de los principales objetivos de gestión, conduciendo a mejoras importantes en la seguridad de la aviación y menores costes de reparaciones, mantenimiento, retrasos y gastos reducidos de aeronaves, estimados en un ahorro anual de unos 800 millones de euros en costes de seguridad equivalentes.[16]

[16] Como comparación, en el año 2021, los presupuestos de la AEI, del CSIC y del CDTI fueron, respectivamente, de 826M, 906M y 1,505M de euros. En el año 2020 fueron, respectivamente, 640M, 834M y 1059M. Basta comparar con el presupuesto del proyecto, que fue de 300K euros en dos años, para entender la rentabilidad del mismo. Viene aquí a la memoria el discurso, todavía muy actual, de inauguración de curso de la RAC en 1963 de Sixto Ríos titulado *La rentabilidad de la investigación científica*, una idea aún insuficientemente apreciada en nuestro país.

Desde un punto de vista más cualitativo, la AESA ahora es capaz de respaldar y documentar mejor sus decisiones y discutirlas de manera más convincente con los numerosos afectados, particularmente aquellos que necesitan alinear sus sistemas de gestión de seguridad con el SSP, lo que ha provocado cambios en los procedimientos y prácticas de la aviación en nuestro país.

- La metodología utilizada constituye un gran avance en la gestión de riesgos de seguridad operacional de la aviación a nivel estatal, al aprovechar al máximo la información disponible y desplegar herramientas analíticas sofisticadas para desarrollar un SSP.
- El resultado es un sistema de ayuda a la toma de decisiones que se ha aplicado con éxito.
- La altísima carga de modelización requerida puede llevar a la necesidad de automatizar estas tareas.

3.4 Máquinas morales

El tercer ODS menciona *reducir las muertes y lesiones por accidentes de tráfico*. El noveno ODS incluye como meta *modernizar las infraestructuras para que sean sostenibles*. El decimoprimero menciona *proporcionar acceso a sistemas de transporte seguros, asequibles, accesibles y sostenibles para todos y mejorar la seguridad vial...prestando especial atención a las necesidades de las personas en situación de vulnerabilidad, las mujeres, los niños, las personas con discapacidad y las personas de edad*. La adopción a medio plazo de los vehículos autónomos revolucionará nuestra forma de vida. El empleo de modelos de aprendizaje profundo y aprendizaje por refuerzo facilitado por hardware específico está posibilitando su implementación, aunque requiere la solución de nuevos problema de decisión que conllevan dilemas morales.

Los ADS van a revolucionar el transporte por carretera (Burns y Shulgan, 2019): facilitado por los avances recientes en IA y en hardware, el transporte masivo mediante vehículos autónomos ha dejado de ser una realidad distante en el tiempo. Sin embargo, la transición a un sistema de circulación totalmente automatizado será un proceso incremental, desde los vehículos conducidos por personas (MVs) a los ADS, como se refleja en la taxonomía de seis niveles de la Society of Automobile Engineers (2018).[17] Hasta hace poco, las carreteras estaban ocupadas exclusivamente por vehículos de nivel 0; a lo largo de la última década,

Como comparación, a la hora de escribir este texto el valor de mercado de la plantilla del Real Madrid se estima en 783M de euros y en 739M la del Atlético de Madrid.

[17] En esta taxonomía, el nivel 0 describe vehículos que no cuentan con capacidades autónomas, y los niveles del 1 al 5 representan vehículos que incrementan sus capacidades de ser autónomos hasta llegar a los completamente autónomos en el nivel 5 ADS.

los fabricantes han empezado a producir vehículos con un mayor nivel de automatización, creciendo considerablemente la cantidad de vehículos de niveles 1 ó 2. Numerosas limitaciones relacionadas con su seguridad y robustez operativa probablemente restringirán los ADS a los niveles 3 y 4 en la próxima década. Estos requieren la intervención humana cuando operan fuera de su dominio operativo (ODD) a través de una operación denominada *petición de intervención* (RtI).[18]

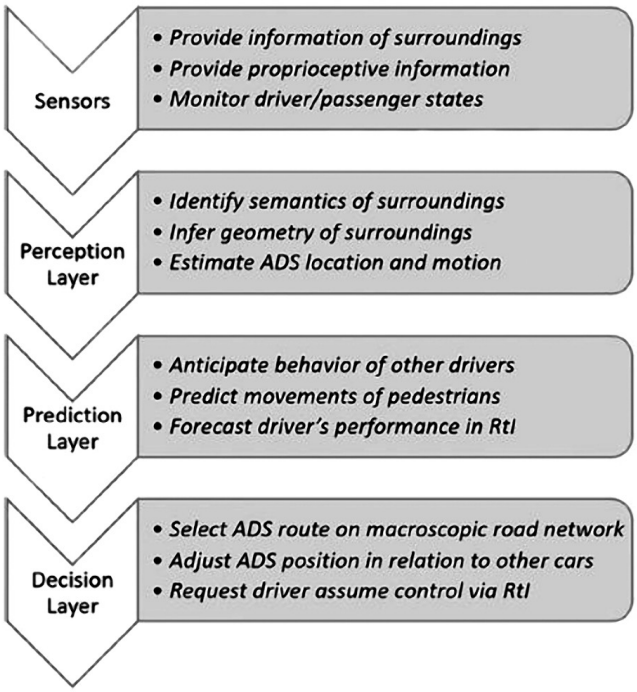

Figura 4: Arquitectura de conducción autónoma (Caballero et al., 2021).

Los ADS modernos se basan en una serie de componentes que emplean hardware potente y una variedad de sensores para generar salidas de dirección y aceleración como se ilustra en la Figura 4. La información del entorno se recopila, entre otros, a través de cámaras de luz visible, sensores de detección y rango de luz (LiDAR), sensores de detección y rango de radio (RADAR) o sensores propioceptivos. También se recopila a través de sistemas de comunicación entre vehículos (V2V) o de vehículo a infraestructura (V2I). Una vez recogido, este compendio de información se utiliza en sistemas que perciben tanto el entorno externo como el interno del ADS (por ejemplo, sistemas de monitorizacion del conductor). Las entradas de los sensores se utilizan en una secuencia de algoritmos de aprendizaje profundo procesados en plataformas informáticas potentes y compactas diseñadas específicamente para tareas de

[18] Petición de intervención al conductor, Request to intervene.

aprendizaje automático (por ejemplo, Nvidia Drive PX2). Los algoritmos utilizados se agrupan en tres capas (de percepción, de predicción y de decisión) frecuentemente integradas en una arquitectura de extremo a extremo. Los algoritmos de la capa de percepción suelen recibir entradas de los sensores sin procesar; estiman la posición del ADS, así como determinan la geometría y la semántica (es decir, la clase y el estado del entorno) externos al vehículo. Para realizar estas tareas, el ADS emplea algoritmos como redes neuronales convolutivas para clasificación (Wu et al., 2017), FastSLAM para localización y mapeo simultáneo (Durrant-Whyte y Bailey, 2006) y filtros de Kalman extendidos para la fusión de información de sensores (Roumeliotis y Bekey, 1997). Las salidas de la capa de percepción se emplean como entradas a la capa de predicción que predice cambios en el entorno percibido; en esta capa se emplean esquemas de modelización que incorporan información incompleta, como procesos de decisión de Markov parcialmente observables (McAllister et al., 2017). Los resultados de la capa de predicción entran en la capa de decisión que se encarga de la planificación de rutas y movimientos, tanto de la ruta macroscópica del ADS en la red de carreteras, como de su movimiento granular en el flujo de tráfico, por ejemplo véase Claussmann et al. (2019). Este esquema general de funcionamiento de los ADS ha tenido cierto éxito, pero quedan importantes desafíos científicos y tecnológicos por resolver antes de que pueda acaecer la adopción masiva de los ADS en las carreteras, algunos ya mencionados en la Sección 3.1. Otros problemas son el de entrenamiento de los algoritmos de aprendizaje automático de ADSs frente a sucesos raros, o el desarrollo de funciones de utilidad apropiadas para los algoritmos de la capa de decisión.

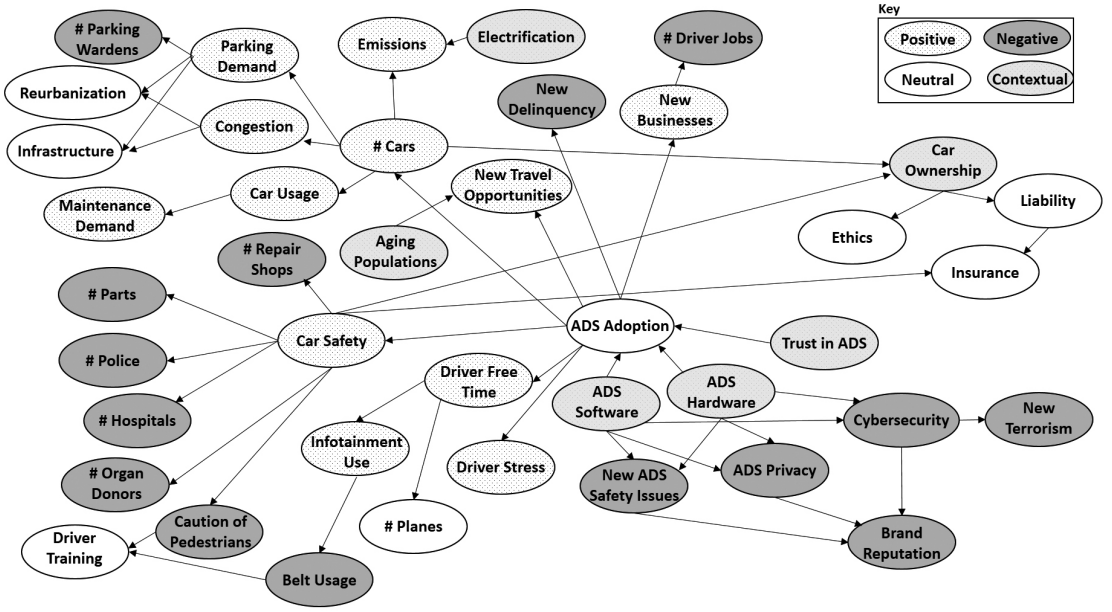

Figura 5: Impactos sociales de los ADS (Caballero et al., 2021).

La Figura 5 enumera los principales factores sobre los que impactarán los ADS. Muchos de ellos se pueden considerar como positivos. Por ejemplo, si la tasa de error de conducción de tecnología ADS acaba siendo menor que la de las personas, deberíamos esperar un modo de transporte más seguro. Además, entre muchos otros, los ADS brindan un medio de movilidad personal a quienes no pueden conducir un MV. Sin embargo, existen otros impactos que pueden percibirse negativamente: algunas profesiones corren el riesgo de desaparecer, como la de taxista debido a la emergencia de taxis autónomos con costes muy competitivos; la reducción de la frecuencia de accidentes puede afectar negativamente a las donaciones de órganos, pues se originan en gran medida de fallecidos en accidentes de tráfico. Otros impactos asociados a la adopción masiva de ADS requerirán una redefinición importante del status quo actual, sin que necesariamente tengan connotaciones positivas o negativas. Por ejemplo, la formación de conductores deberá evolucionar; igualmente, el sector de seguros del automóvil tendrá que adaptarse, especialmente en lo que respecta a cuestiones de responsabilidad.

Como ha ocurrido en otras ocasiones, la tecnología avanza a mayor velocidad que nuestra capacidad colectiva para integrarla en la sociedad. Es por ello que se están implantando diversas iniciativas para aumentar nuestra confianza en los ADSs y facilitar su adopción. Una iniciativa europea en esta línea es el proyecto H2020 Trustonomy,[19] centrado en la operación RtI. La Figura 6 muestra el esquema que implementamos en este problema (Rios Insua et al., 2021a). El sistema hace predicciones del entorno y del estado del conductor y calcula la trayectoria en los siguientes *k* instantes; calcula entonces si se alcanzaría mayor utilidad esperada con el conductor o en el modo autónomo y, en función de ello, toma las decisiones correspondientes. Si se invoca una RtI, se acompaña de una evaluación del comportamiento del conductor (DIPA) lo que ayuda a gestionar futuras RtIs (o hacer paradas de emergencia).

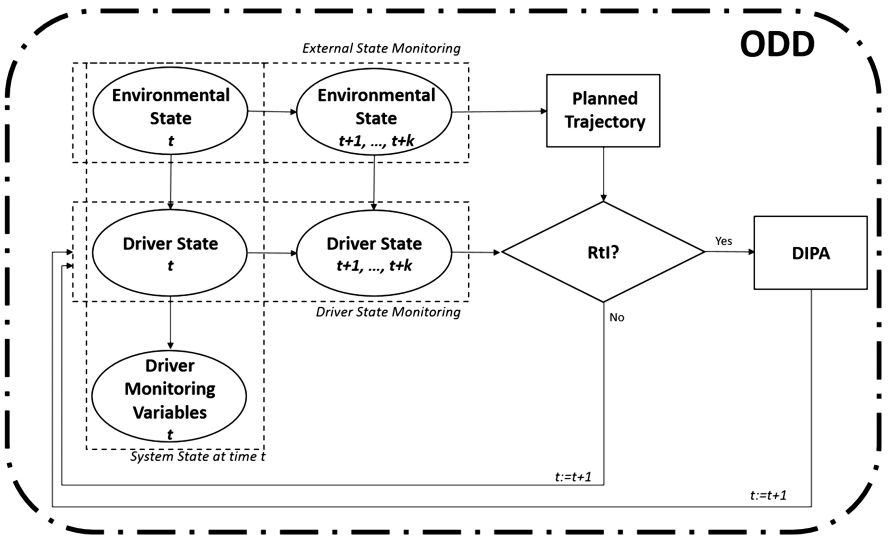

Figura 6: Apoyo a la toma de decisiones en ADS. Tomado de Rios Insua et al. (2021a).

[19] Véase https://h2020-trustonomy.eu/.

Aunque conceptualmente puede parecer relativamente sencilla, la gestión de RtIs entraña grandes desafíos, algunos de ellos de carácter ético. Los modelos y simulaciones ya desarrolladas en el marco de la gestión de RtIs (Rios Insua et al., 2021a) han permitido identificar lo que denominamos *dilema fundamental de los ADS de nivel 3 y 4*:

> *Si el vehículo predice una situación peligrosa, se espera que el conductor esté alerta y dispuesto a retomar el control. Sin embargo, si está distraído y el ADS transfiere el control, las consecuencias podrían ser catastróficas. ¿Debería el ADS transferirlo, con lo que el conductor asume los riesgos asociados a la distracción, o retenerlo, con lo que el ADS debe tomar decisiones de vida y muerte fuera de su dominio operativo? En este último caso, ¿quién sería responsable en una eventual catástrofe?*

Este es un ejemplo paradigmático del tipo de problemas éticos en ADS, de los que son especialmente conocidos los denominados *problemas de tranvía* (Jarvis Thomson, 1985). En estos, se expone a un usuario a la toma de decisiones de vida y muerte en los que ha de elegir entre proteger a sus pasajeros o a los peatones. Imaginemos por ejemplo que, ante un fallo repentino del sistema de frenado, el ADS puede:

- Continuar recto, lo que resultaría en la muerte de un científico de élite, una médico y un gato, o
- Girar, chocando con un muro, lo que resultaría en la muerte de los cinco ocupantes del vehículo: dos criminales, una persona sin hogar, una mujer embarazada y un futbolista de élite.

¿Qúe elegiríamos en tal situación?

El proyecto Moral Machine[20] enfrentó a millones de personas de diferentes culturas a experimentos de este estilo. Su objetivo era obtener conocimiento acerca de las prioridades éticas colectivas. Sin embargo, aunque el estudio de las diferentes perspectivas éticas es interesante, no da respuesta a cómo un ADS debe gestionar y controlar sus acciones, un problema complejo que se refiere no sólo a las situaciones de emergencia propias de los problemas de tranvía.[21]

De forma interesante, el análisis de riesgos permite arrojar luz sobre cómo un ADS puede actuar ante un conflicto ético (Caballero et al., 2022). De hecho, cualquier corriente ética, desde la deontológica a la consecuencialista, incluyendo aproximaciones tanto utilitaristas como autoprotectivas, puede modelizarse matemáticamente. Así, desarrollamos un marco basado en la teoría de la decisión, que puede emplearse para gestionar y evaluar la toma de decisiones (tanto en emergencias éticas como en situaciones estándar) en ADS. El modelo tiene en cuenta objetivos múltiples: rendimiento del vehículo, confort de los pasajeros, duración

[20] Véase https://www.moralmachine.net/.

[21] Esta complejidad viene bien reflejada en el reciente informe de la Comisión Europea *Ethics of connected and automated vehicles* donde se identifican recomendaciones en relación con la seguridad y el riesgo en las carreteras, los aspectos éticos de los datos y los algoritmos y la cuestión de la responsabilidad.

del viaje, seguridad (la de los pasajeros, las personas en la escena de conducción, el propio vehículo y la infraestructura) e incluso la reputación del fabricante. Una vez definidos los objetivos, un productor podría decidir ponderarlos de distinta manera, dando más importancia a aquellos que sean de su interés: por ejemplo, un fabricante o un conductor podría dar mayor importancia a la seguridad de los pasajeros, a costa de una menor de los peatones. Estos pesos se utilizarían para combinar los distintos objetivos en una única función de utilidad que regiría las operaciones del ADS. El interés de este modelo reside en el hecho de que, en caso de accidentes, es posible simular escenarios de conducción múltiples empleando la función de utilidad elegida por el fabricante para guiar las decisiones del ADS. Estas simulaciones permitirían evaluar si el vehículo satisface la regulación vigente, y en caso de no hacerlo, determinar responsabilidades.

En definitiva, este modelo conforma un marco que permite evaluar de forma objetiva si las elecciones éticas por parte de fabricantes o usuarios satisfacen las líneas establecidas en una regulación.

- Los desarrollos en IA están posibilitando nuevas tecnologías, como la de los vehículos autónomos.
- Estas tecnologías requieren aun desarrollos en relación con la toma de decisiones.
- Igualmente, conllevan problemas morales que pueden resolverse a través de los métodos del análisis de riesgos.

3.5 Máquinas empáticas

El cuarto ODS se refiere a *Asegurar una educación inclusiva y de calidad y promover el aprendizaje a lo largo de la vida para todos*. La pandemia ha puesto de manifiesto una serie de problemas sociales que distan de estar bien resueltos. Entre ellos, mencionamos dos aparentemente inconexos: el aislamiento de pacientes infectados en UCI, que pasaban largas jornadas sufriendo esencialmente en soledad, y las complicaciones inducidas por la enseñanza a distancia, especialmente entre los estudiantes más jóvenes y aquellos con necesidades especiales. El nexo entre ambos está en su solución con ayuda de robots sociales. Su adecuada implementación requiere dotar a los mismos de un toque emocional que facilite su interacción con los usuarios.

Continuamos con el diseño de agentes que tomen decisiones de forma autónoma, como en la Sección 3.4. Pero, además, estos deben reflejar o simular emociones, de manera que éstas tengan cierta influencia en su toma de decisiones. Nuestro objetivo funcional último no es meramente descriptivo, sino mejorar la interacción del agente con los usuarios que puedan

aparecer en su escena. Nuestro objetivo aplicado, perseguido al fundar Aisoy,[22] es diseñar robots sociales emocionales que puedan adoptar el rol de mentor para un niño o grupo de niños o puedan acompañar, por ejemplo, a personas mayores o pacientes en un hospital.

A los modelos esquematizados en la Figura 4, debemos acoplarles conceptos en relación con emociones dentro de lo que se ha denominado *toma de decisiones afectivas*, véase Loewenstein y Lerner (2003). Tradicionalmente, las emociones se han considerado como alejadas de la racionalidad. Así, por ejemplo, Platón declaraba

Las pasiones, los deseos, los temores nos impiden pensar...

Tal visión comienza a verse cuestionadas con el primer artículo de Tversky y Kahneman (1971) sobre heurísticas y sesgos, véase Kahneman (2011). Desde entonces, se han venido produciendo descubrimientos e innovaciones que están modificando la anterior visión tradicional. Por ejemplo, dentro de las neurociencias, se promueve el concepto de inteligencias múltiples (Gardner, 2011) que incluye la inteligencia emocional. Se produce, además, el nacimiento de campos como la *neuroeconomía*, véase Glimcher y Fehr (2013). Se han realizado, igualmente, numerosos experimentos que han mostrado el impacto de las emociones sobre la toma de decisiones. Algunos ejemplos clásicos incluyen tener en cuenta que las reacciones afectivas en relación con sentimientos básicos, como gusto o disgusto, preceden a evaluaciones cognitivas o que los estados mentales positivos promueven el pensamiento creativo, mientras que los negativos parecen promover el pensamiento analítico, véase Mellers et al. (1998) para más detalles. Las ideas anteriores se expresan en sistemas computacionales que dan lugar al campo de la *computación afectiva*, véase, p.ej., el trabajo pionero de Picard (1997).

Podemos considerar que, en definitiva, el objetivo final de esta búsqueda es diseñar un agente que sea capaz, dentro de un entorno en el que puede haber otros agentes, de percibir tal entorno y las acciones de tales agentes; en función de estas percepciones, mostrar algún tipo de emociones y, finalmente, que éstas se reflejen en la toma de decisiones del agente. Resolver esta tarea requiere que seamos capaces de resolver científica y tecnológicamente una serie de actividades básicas (percepción, inferencia, predicción, afecto, decisión) que, en conjunto, forman un puzle cuyas piezas se engarzarían como en la Figura 7.

Por comparación con el esquema descrito para los ADS en la Figura 4, la parte más novedosa sería el modelo de preferencias con emociones. Comenzamos identificando los objetivos vitales de nuestro agente. Para ello, nos inspiramos en la pirámide motivacional de Maslow (1943) y construimos objetivos relacionados. Sin pérdida de generalidad, asumimos una función de utilidad aditiva. Los pesos pueden estar ordenados de forma creciente en función de la posición en la jerarquía, para promover que se dediquen más recursos computacionales a los objetivos más básicos. Además, la forma de las funciones componentes de utilidad permite modelizar que, una vez suficientemente satisfechos los niveles en los objetivos más básicos, se pase a perseguir objetivos de nivel superior.

[22] https://aisoy.com/

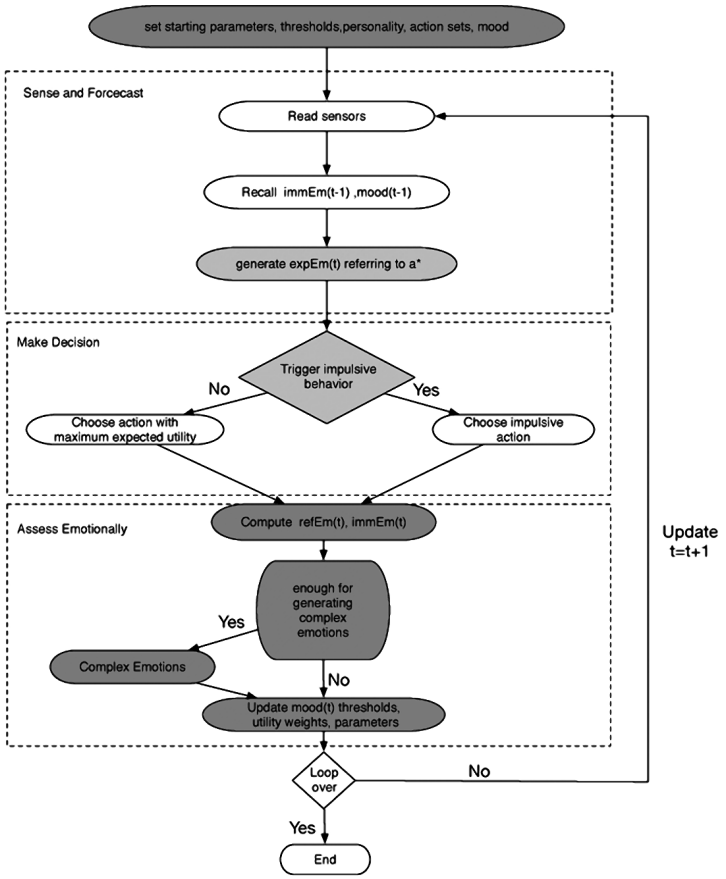

Figura 7: Apoyo a la toma de decisiones en robots sociales emocionales. (Liu y Rios Insua, 2020a)

Nos falta enfrentarnos a un concepto elusivo, el de emoción, sobre el que se han dado numerosas definiciones e interpretaciones, véase Russell y Barrett (1999). Nosotros adoptamos una aproximación pragmática, desde el punto de vista computacional, basada en la presencia de emociones básicas, cuya composición da lugar a emociones más complejas. Quedaría entonces definir las emociones a adoptar, cómo se componen y cómo afectan a la toma de decisiones, que pueden verse en detalle en Liu y Rios Insua (2020a). Incluyen elementos de personalidad según el modelo HEXACO (Ashton et al., 2014); emociones en cuatro grupos (esperadas, inmediatas, referenciales y complejas) y humor y modelos para su definición y actualización. Se incluye un esquema en la Figura 8.

Las ideas se extienden a grupos de agentes en Liu y Rios Insua (2020b). De forma interesante, los experimentos realizados con estos grupos muestran como una sociedad cooperativa emocional alcanza mejores resultados que una sociedad que carece o bien de capacidades de cooperación, y por tanto cada individuo busca simplemente su propio bien, o bien de capacidades emocionales.

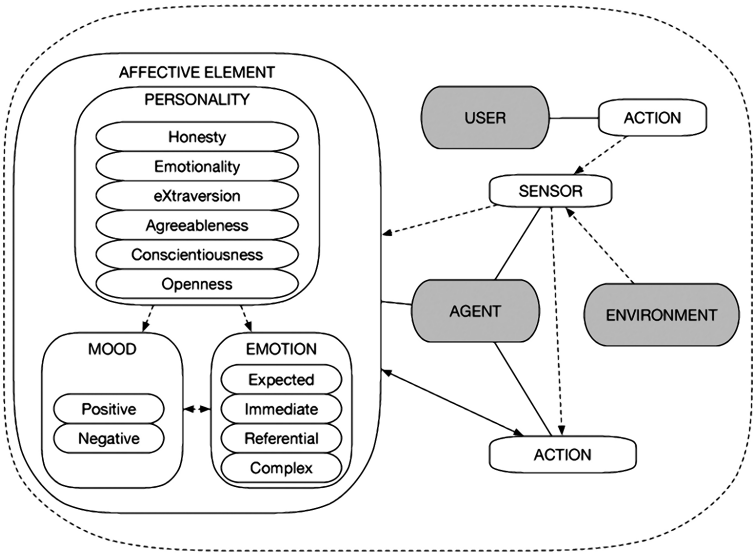

Figura 8: Esquema del modelo emocional adoptado. (Liu y Rios Insua, 2020a)

El esquema anterior se ha implementado en Aiko, Fig. 9, una plataforma robótica flexible de Aisoy, basada en un procesador Raspberry Pi 4, y se ha aplicado con éxito en educación, educación con necesidades especiales, acompañamiento de personas mayores y acompañamiento de personas enfermas.

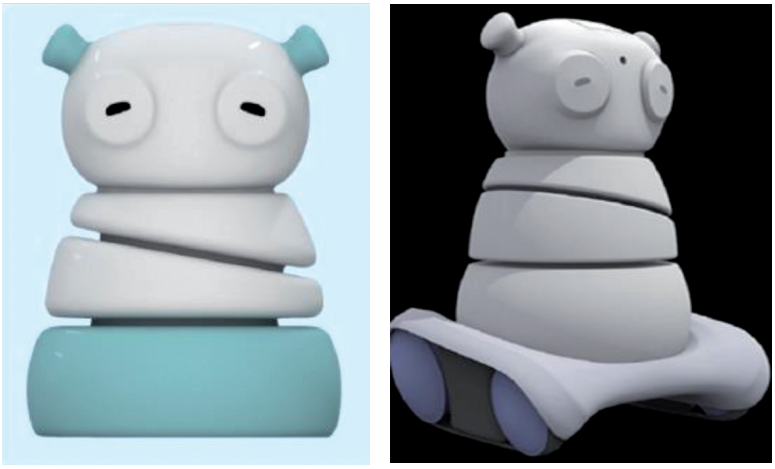

Figura 9: Aiko.

- Es posible acoplar elementos emocionales dentro de la teoría de la utilidad estándar para mejorar la interacción de los agentes con los usuarios y aumentar su aceptación.
- Podemos diseñar los objetivos de los agentes para satisfacer las necesidades que queramos atender.

4. ...Y LUEGO, ALGUNAS SOMBRAS

En la Sección 3 hemos mostrado, a través de proyectos reales, usos de la IA y del Big Data que sugieren el enorme potencial para resolver algunos problemas globales del máximo interés social. Hemos resumido tales logros en algunos principios importantes. Pero, además de estos éxitos, debemos ser también conscientes de los riesgos asociados a tales tecnologías y metodologías, algunos de los cuales describimos a continuación. Proceden también de proyectos recientes en los que hemos estado involucrados. De nuevo, los relacionaremos con algunos de los ODS, aunque igualmente podríamos apelar a la Declaración Universal de los Derechos Humanos[23] o legislación reciente como el Reglamento General de Protección de Datos (RGPD).[24]

4.1 Perfilado

El ODS 16 incluye como meta *desarrollar instituciones auditables y transparentes y proteger las libertades fundamentales*. Aquí observamos posibles violaciones de este principio basadas en el perfilado digital de personas.

Los datos de geolocalización son de gran interés en diversos contextos. Se obtienen, por ejemplo, a través de móviles y distintas apps en ellos instalados. Como ejemplo, la Figura 10 izda muestra la traza durante un día de un usuario de una app de pago que empleamos en un proyecto de geomarketing. Basado en esos datos (a lo largo de varios días) es relativamente sencillo, con técnicas de filtrado y de análisis de conglomerados, encontrar el hogar y el lugar de trabajo (si es fijo) de los individuos y producir agrupaciones de clientes en función de los mismos, véase la Figura 10 dcha. Tal análisis tiene usos relevantes en planificación urbana. Análisis parecidos, que además tengan en cuenta velocidades de desplazamiento, facilitan igualmente actividades de planificación de movilidad urbana.[25]

Aparte de los usos anteriores, para obtener valor adicional debemos cruzar tales datos con información de otro origen para dar contenido semántico a los lugares que visita el individuo. Una primera aproximación consiste en emplear geovallas virtuales teniendo en cuenta las coordenadas de lugares importantes de una zona (sus estaciones principales, sus campus universitarios, sus estadios deportivos principales,...). Otra es cruzar las coordenadas obtenidas con bases de datos geolocalizadas, por ejemplo, asociadas a Google Maps u OpenStreetMap que incluyen, además, la tipología del lugar correspondiente (centro de estudios, restaurante, sinagoga,...) Con esa información es posible encontrar un perfil detallado de los hábitos espacio-temporales de los individuos con técnicas de análisis de conglomerados, modelos de

---

[23] https://www.un.org/es/about-us/universal-declaration-of-human-rights

[24] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[25] Incidentalmente, este tipo de datos y análisis, acoplado a modelos de predicción de transmisión basados en tiempo de contacto y otras covariables facilitarían un esquema de detección de transmisión de la pandemia más eficaz y seguro que, por ejemplo, RadarCOVID. Desafortunadamente, en esta ocasión no pudimos convencer a las autoridades competentes de esta posibilidad.

procesos estocásticos y métodos de inferencia bayesiana, incluyendo su perfil como viajante, su religión (si es que practica alguna), su edad, sus patrones de consumo de distintos servicios en el tiempo, su nivel socio-económico,...[26] A partir de ahí, podemos segmentar los individuos con fines comerciales y predecir su tránsito temporal para enviarles la publicidad adecuada en el momento adecuado. También tiene usos claros en seguridad pues nos permitiría detectar cambios inesperados en los patrones de desplazamiento.

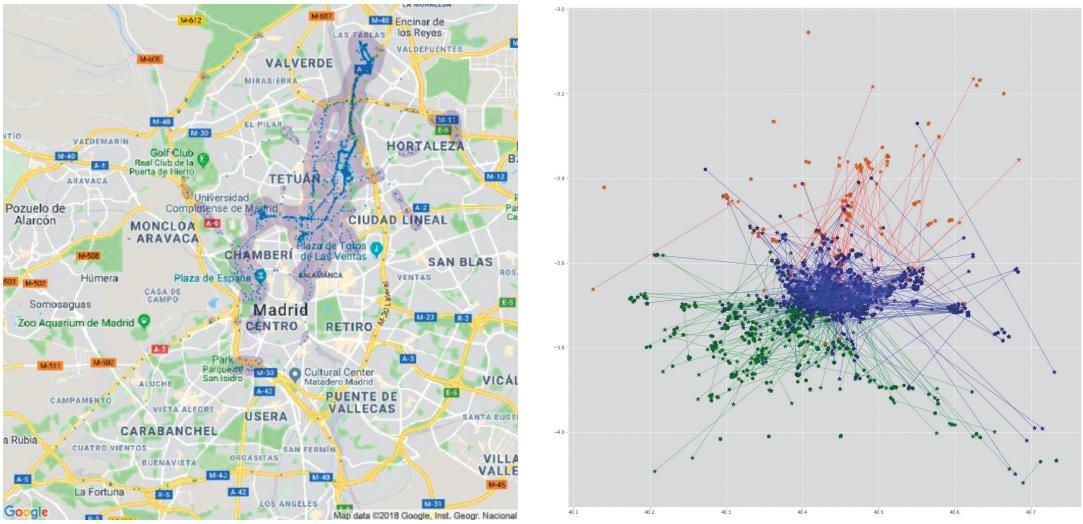

Figura 10: Izda: Traza de un usuario de una app de pago. Dcha: Agrupación de usuarios según domicilio y lugar de trabajo.

En la descripción anterior se habrán apreciado las posibles violaciones a la privacidad: es relativamente sencillo controlar dónde nos ubicamos y, más aún, predecir donde nos vamos a ubicar, así como predecir lo que haremos en tales ubicaciones. Estas capacidades se potencian si, además, se cruza la información anterior con el análisis de textos emitidos por el usuario (por ejemplo, a través de twitter) como tuvimos que realizar en un proyecto de marketing en redes sociales. A partir de ellos, podemos inferir los rasgos de personalidad (por ejemplo, según el modelo OCEAN (Goldberg, 1990) en relación con su apertura a la experiencia, escrupulosidad, extroversión, amabilidad y neuroticismo), así como su propensión a compra de ciertos productos, emulando lo que hace el servicio Personality Insights de IBM Watson.[27] Esto nos permitiría modular los mensajes a enviar a los individuos para persuadirles mejor de ciertas opciones. Combinando ambos perfiles tenemos pues información de dónde y cúando se va a ubicar una persona, qúe cosas le interesan y cómo debemos comunicarle esas cosas para incrementar su interés.

[26] Curiosamente, no es fácil sin embargo inferir el género del individuo con este tipo de datos.

[27] A punto de ser discontinuado.

Obviamente esta modelización es de interés en marketing comercial, pero también en el ámbito político y abre la puerta a escándalos como el de Facebook y Cambridge Analytica en relación al Brexit. El RGPD pone, sin embargo, fuertes restricciones sobre la posibilidad de perfilado de individuos.[28]

- A partir de la huella digital de un individuo es posible inferir numerosas propiedades del mismo, lo que puede hacernos muy vulnerables.

4.2 Seguridad

El decimosexto ODS considera otros aspectos relativos a seguridad. En particular, identifica *reducir significativamente todas las formas de violencia, promover el estado de derecho, garantizar el acceso público a la información y proteger las libertades fundamentales y combatir el terrorismo y la delincuencia*. Desafortunadamente, los nuevos sistemas de aprendizaje automático abren la puerta a nuevas formas de violencia que debemos mitigar.

Como hemos indicado, cada vez se despliegan más sistemas basados en IA. Sin embargo, algunos de ellos son vulnerables a ataques maliciosos, lo que introduce riesgos obvios sobre, por ejemplo, sistemas de filtro de contenidos, ADSs o sistemas de defensa, véase Comiter (2019). En consecuencia, estos deberían ser robustos en sus respuestas frente a tales ataques, si queremos confiar en operaciones basadas en sus salidas. Los algoritmos de última generación, como los arriba descritos, funcionan extraordinariamente bien frente a datos estándar, pero han demostrado ser vulnerables frente a ejemplos adversarios, instancias de datos dirigidas a engañarlos (Goodfellow et al., 2014).

En tales contextos, los algoritmos deben diseñarse para tener en cuenta la posible presencia de adversarios para protegerlos frente a manipulaciones de datos. Como hipótesis fundamental, los sistemas basados en IA se basan típicamente en el uso de datos independientes e idénticamente distribuidos, tanto para el entrenamiento como para las operaciones. Sin embargo, los aspectos de seguridad en el aprendizaje profundo, parte del campo emergente del aprendizaje automático adversario (AML), cuestionan dicha hipótesis, dada la presencia de adversarios dispuestos a intervenir en el problema para modificar los datos y obtener un beneficio.

Como ejemplo motivador, los algoritmos de visión (Sección 3.1) son el núcleo de varias tecnologías, como los ADS (Sección 3.3). Los ejemplos de ataques a tales algoritmos más sencillos y conocidos consisten en modificaciones de imágenes de manera que la alteración se

[28] Dice *La elaboración de perfiles está sujeta a las normas del presente Reglamento que rigen el tratamiento de datos personales, como los fundamentos jurídicos del tratamiento o los principios de la protección de datos*, apareciendo veinticinco referencias a esta actividad en el texto.

vuelve irrelevante para el ojo humano a efectos de reconocimiento, pero hace que un modelo entrenado en millones de imágenes clasifique erróneamente las atacadas, con consecuencias de seguridad potencialmente relevantes. Por ejemplo, con un modelo de red convolutiva relativamente sencillo, similar al de la Sección 3.1, podemos predecir con precisión del 99% los dígitos escritos a mano en el conjunto de datos MNIST (LeCun et al., 1998). Sin embargo, si atacamos esos datos con el método rápido del signo del gradiente (Szegedy et al., 2013), la precisión se reduce hasta el 62%. La Fig. 11 proporciona un ejemplo de una imagen MNIST original y una atacada: a nuestros ojos ambas imágenes parecen un 2, pero el clasificador profundo identifica correctamente un 2 en el primer caso, mientras que sugiere un 7 tras el ataque.

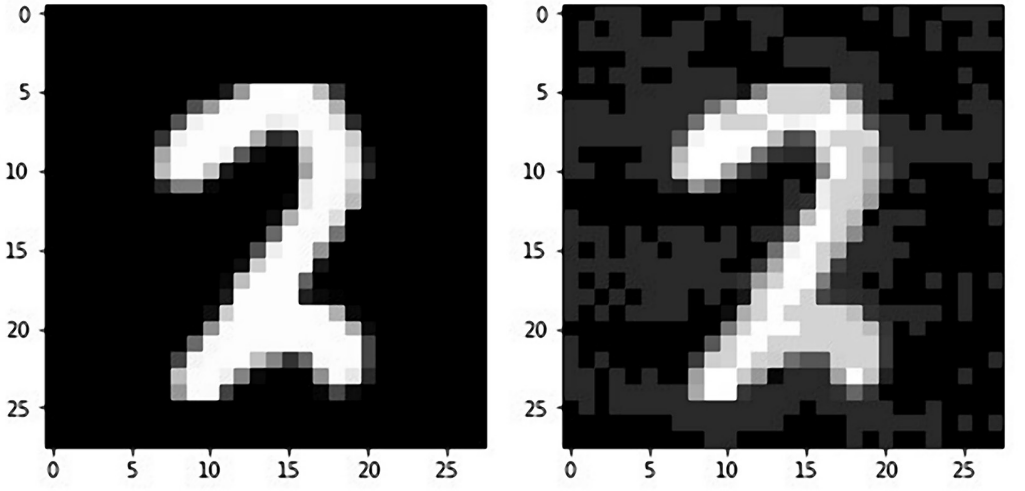

Figura 11: Izda: Imagen original, correctamente clasificada como un 2. Dcha: Imagen ligeramente perturbada, incorrectamente clasificada como un 7.

Desde los trabajos pioneros en clasificación adversaria (Dalvi et al., 2004), el paradigma predominante en AML modeliza la confrontación entre sistemas basados en IA y sus adversarios por medio de la teoría de juegos. Esto conlleva hipótesis de conocimiento común (Hargreaves-Heap y Varoufakis, 2004) que son cuestionables en aplicaciones de seguridad puesto que los adversarios intentan ocultar información. Como señalan Fan et al. (2021), resulta necesario construir un marco bien fundamentado que garantice la solidez del aprendizaje profundo contra manipulaciones adversarias. El enfoque habitual para robustecer modelos frente a estos ejemplos es el *entrenamiento adversario* (Madry et al., 2018) y sus variantes, basado en la resolución de un problema de optimización de dos niveles cuya función objetivo es el riesgo empírico de un modelo frente al peor caso de perturbaciones de datos.

El AML es un área difícil que evoluciona rápidamente y está conduciendo a una verdadera carrera de armamento en la que la comunidad alterna ciclos de proposición de ataques con ciclos de presentación de defensas frente a los mismos. Sin embargo, como se ha mencionado, se basa en condiciones fuertes de conocimiento común no sostenibles. En Rios Insua et al.

(2020b) se propone una metodología basada en la teoría de la decisión bayesiana para resolver problemas de AML, adoptando una perspectiva desde el análisis de riesgos adversarios que evita esas consideraciones. Rios Insua et al. (2020a) aplica este marco a la clasificación adversaria ilustrando la mayor robustez frente a ataques adversarios que se obtiene con tal aproximación.

- Los nuevos sistemas basados en IA resultan vulnerables frente a distintos tipos de ataques. Resulta esencial desarrollar métodos robustos frente a los mismos.
- Los métodos bayesianos basados en el análisis de riesgos adversarios ofrecen una alternativa poderosa en contextos de seguridad.

4.3 Interpretabilidad

El decimosexto ODS incluye también una referencia a *crear a todos los niveles, instituciones eficaces y transparentes que rindan cuentas*. Entre otras cosas, esto exige que las instituciones sean capaces de explicar las decisiones que toman, un derecho que, como individuos, viene recogido en el RGPD.

Muchos de los algoritmos anteriormente descritos, por ejemplo los basados en redes profundas, tienen escasa capacidad explicativa, lo que puede motivar dudas éticas sobre su aplicabilidad para apoyar la toma de decisiones, problema que se acentúa si se detectan sesgos cuando tales decisiones discriminan a algún colectivo.[29]

Los algoritmos de aprendizaje automático se han empleado con éxito para desarrollar IAs capaces de jugar de manera sobrehumana[30] en juegos conocidos como el ajedrez, el póquer o el go y en otras actividades de mayor impacto social. La forma de estas IA, a menudo basadas en redes profundas, dificulta que un ser humano comprenda cómo el sistema computacional toma sus decisiones. Si bien descubrir estrategias sobrehumanas es un objetivo importante, es igualmente relevante comprender el razonamiento subyacente que explica por qúe estas estrategias son superiores. Las IA de "caja negra", aunque resultan ser estrategas excepcionalmente buenas, nos dejan preguntándonos demasiadas veces ¿Cómo hizo eso la máquina?, cuestión especialmente importante en áreas sensibles como defensa, sanidad o finanzas, que requieren enfoques transparentes, responsables y comprensibles. En este punto, vale la pena mencionar cómo el RGPD podría requerir que los proveedores de IA proporcionen explicaciones sobre los resultados de la toma de decisiones automatizada basada en datos personales.[31] Esta falta de transparencia ha llevado a un interés creciente por una subárea del

[29] Recúerdese aquí el (casi) apocalíptico O'Neil (2016).

[30] En el sentido de que vencen al mejor jugador humano.

[31] Esto incluso conduciría a la prohibición de modelos opacos en ciertos dominios de aplicación.

aprendizaje automático conocida como IA explicable (XAI) que, aunque puede tener varios significados dependiendo del contexto, debe cumplir dos requisitos: en primer lugar, como cualquier IA, debe ser capaz de tomar buenas decisiones o adoptar inferencias precisas; en segundo debe explicar fácilmente a los no expertos cómo llegó a sus conclusiones. Según esta definición, una IA que predice el clima utilizando un modelo matemático de dinámica atmosférica sería explicable; una basada en redes profundas no lo sería, típicamente.

Hay varios enfoques a este problema que se revisan a fondo en Burkart y Huber (2021). Una primera posibilidad es emplear modelos interpretables, fácilmente comprensibles para los humanos, como argumenta convincentemente Rudin (2019) que afirma que, en muchos contextos, tales modelos pueden funcionar casi tan bien como las redes neuronales profundas. Si bien el uso de modelos interpretables puede ser adecuado en algunos contextos, tiene el coste de su flexibilidad, precisión y usabilidad. Alternativamente, a veces se genera un modelo interpretable sustituto del de caja negra, para ganar interpretabilidad, ya sea de forma global o localmente alrededor de algunas entradas, como se hace con LIME (Ribeiro et al., 2016) o SHAP (Lundberg y Lee, 2017). Finalmente, hay intentos de crear métodos para explicar los modelos de caja negra con dos estrategias generales principales: extraer globalmente una explicación de un modelo que sea representativo de algún conjunto de datos específico; o extraer localmente una explicación para una única entrada de prueba y la predicción correspondiente. Samek et al. (2017) describen diferentes métodos para visualizar y explicar modelos de aprendizaje profundo como la propagación de relevancia por capas.

Otro problema interesante, y más fundamental, se refiere a aprender las reglas del juego. La mayoría de sistemas de IA típicamente llevan las reglas del juego preprogramadas y su entrenamiento consiste en aprender a escoger una estrategia ganadora del conjunto de movimientos factibles basado en el estado actual del juego. Así, el sistema tiene ventaja sobre un principiante que debe primero aprender las reglas antes de aprender una buena estrategia. La idea de que una máquina aprenda las reglas del juego ha motivado investigación en sistemas que aprenden observando cómo otros juegan, véase p.ej. Bjornsson (2012). Igualmente, en zoología se estudia la capacidad para aprender reglas de juegos como una habilidad general de la inteligencia en primates (Gao et al., 2018). En una contribución muy reciente (Aurentz et al., 2022) hemos presentado una IA interpretable, y su correspondiente algoritmo de aprendizaje automático, capaz de aprender en tiempo polinómico las reglas de un juego siempre que las relaciones entre el estado de un jugador y sus movimientos factibles pueda representarse mediante un conjunto de polinomios de Zhegalkin de grado bajo (como ocurre, por ejemplo, con el dominó o el President). Además, las reglas se almacenan de forma económica y producen una representación fácil de interpretar y transcribible a lenguaje natural. Vemos así como las matemáticas aportan una solución novedosa a un problema difícil en IA.

- En algunos dominios resulta esencial ser capaces de explicar los resultados de los sistemas basados en IA.
- En algunos dominios, resulta suficiente emplear métodos interpretables.
- El empleo de modelos matemáticos avanzados facilita la solución de problemas de la IA.

5. A MODO DE CONCLUSIÓN

Como resumen, nos gustaría incidir en algunas de las ideas centrales que han dirigido esta exposición:

- Es posible que los análisis de Big Data hayan sido de alguna manera sobrevendidos por las consultoras TIC. Por ejemplo, hemos podido leer expresiones del estilo *El diluvio de datos vuelve obsoleto el método científico*,[32] como si sólo necesitásemos recopilar grandes cantidades de datos y, a través de soluciones automatizadas, obtener algún tipo de modelo automatizado para tratar cualquier problema que podamos imaginar. Algunos avances recientes en química y biología, véase p.ej. Gallego et al. (2022), van en tal dirección, pero esa propuesta, sin duda, ignora algunos aspectos importantes de la ciencia.

- Por ejemplo, aunque los datos son importantes, debemos reconocer que, en muchos problemas, no habrá tantos. E incluso si los hubiere, aún existe una clara necesidad de incluir juicios de expertos y otras tecnologías analíticas en los procesos de toma de decisiones y de simulación de políticas para obtener aproximaciones más eficientes.

- Debemos insistir en que Big Data no se refiere sólo a tecnología (a Hadoop, Spark o similares) sino que requiere además metodologías científicas de la estadística y del aprendizaje automático, parte de lo que hoy llamamos ciencia de datos, y conocimientos sobre la materia en la que se hace un proyecto Big Data.[33]

- De las 4 V's que describimos en relación con el Big Data (Sección 2), normalmente se suele incidir en el aspecto del volumen, las grandes cantidades de datos. Sin embargo, es mucho más importante incidir en el aspecto del valor que aportan tales datos. Acumular datos meramente puede ser inútil e ineficiente.

- Desde el punto de vista tecnológico, debemos esperar, como no puede ser de otra manera, una evolución permanente en un fenómeno que apenas acaba de comenzar.

- Desde una perspectiva metodológica, debemos esperar también una importante evolución. Destacamos tres aspectos:
  - Un campo esencial sería el desarrollo de métodos escalables para inferencia bayesiana. Su status quo hace que prevalezcan de nuevo los métodos de máxima verosimilitud en este dominio, que tienden a ignorar la incertidumbre epistémica relacionada con el conocimiento, crucial para desarrollar una IA más segura y justa. Una dirección prometedora se refiere a la combinación de métodos SGMCMC con métodos variacionales como puede verse en Gallego y Rios Insua (2021).
  - También resulta importante la integración coherente de estas metodologías en sistemas de ayuda a la toma de decisiones. Como hemos indicado, el objetivo final

[32] En un artículo de Anderson en Wired en 2008

[33] En los ejemplos de la sección 3 serían Cardiología, Seguridad Aérea, Ingeniería de Transportes y Robótica.

de los modelos de inferencia y predicción debe ser la ayuda a la toma de decisiones y la Teoría de la Decisión (French y Rios Insua, 2000) facilita un marco normativo adecuado para tal integración.

- Finalmente, se suele mencionar el aprendizaje por refuerzo como aproximación a una IA general; sin embargo el énfasis debería ponerse en aprendizaje por refuerzo multiagente. Incidentalmete, aquí de nuevo prevalecen los conceptos de teoría de juegos pero conllevan condiciones de conocimiento común no sostenibles en muchos dominios, por lo que convendría desarrollar las aproximaciones basadas en el análisis de riesgos adversarios (Banks et al., 2015).

- En lo que respecta a las aplicaciones, por el momento han predominado las ideas de negocio, pero hay un enorme potencial en las aplicaciones en el ámbito social para beneficio de las administraciones y organizaciones no gubernamentales, como hemos descrito en la Sección 3. Sin embargo, como mencionamos en la introducción, pocas decisiones gubernamentales se benefician aún del aprovechamiento sistemático de grandes masas de datos y técnicas avanzadas de modelización. Por comparación con las aplicaciones industriales, no es difícil vislumbrar las enormes aplicaciones potenciales que tendrían en problemas relativos al desarrollo racional de planes para infraestructuras; el empleo del conocimiento sobre comportamiento para promover la eficiencia energética; el desarrollo de servicios personalizados de gobierno; la mejora de la experiencia en visitas turísticas; o la identificación de barrios con servicios sociales inadecuados, entre otros muchos. Surge entonces, el uso de la Analítica para apoyar la toma de decisiones en la elaboración de políticas públicas, que denominamos Analítica para Políticas (Policy Analytics) (Daniell et al., 2016). El ejemplo mencionado de AESA, y otras experiencias recientes del Instituto Nacional de Estadística en relación con sus estadísticas experimentales, muestran los enormes ahorros potenciales de los que nuestro país podría beneficiarse con una aplicación coherente y sistemática de las metodologías propuestas.

- Finalmente, desde el punto de vista ético, sería necesaria una mayor concienciación de la población respecto al valor de los datos y la regulación de los aspectos de privacidad. En particular, serían importantes campañas de comunicación en el corto plazo, y de educación en el medio plazo, que pongan de manifiesto la necesidad de disponer de datos y procesos de calidad reutilizables en un marco de transparencia.

Concluimos con un par de deseos que sin duda ayudarían a apuntalar las ideas anteriores para su correcta implementación.

- En estos últimos años,[34] se ha venido debatiendo la reforma de la enseñanza primaria y secundaria, en particular en lo referido a las Matemáticas. En relación con ello, resultaría crucial una amplia revisión en el sentido de modernizar los contenidos y métodos de la enseñanza estadística en esas cruciales etapas educativas; el problema se agrava porque, habitualmente, se dejan tales temas como últimos en los programas y, por tanto, frecuentemente se obvian de facto de los temarios, de forma que el estudiante típicamente estudia muchos menos contenidos probabilísticos de los que deberían no teniendo acceso, práctica y desgraciadamente, al razonamiento estadístico.

  En este punto es interesante comparar los contenidos del Bachillerato Nacional (basados en recetas, ejemplos de bolas, cartas y monedas y el uso de las tablas de la distribución normal en papel)[35] con los del Bachillerato Internacional (basados en razonamiento probabilístico y estadístico, ejemplos de epidemias, accidentes de avión y natalicios, y el empleo de calculadoras gráficas para realizar las estimaciones probabilísticas). La mejor opción es obvia.

- Sería también esencial crear una iniciativa nacional de instituto sobre ciencia de datos, tal vez similar al instituto Turing del Reino Unido.[36] En particular, dentro de una perspectiva multidisciplinar, dos elementos diferenciadores serían centrarlo en problemas de toma de decisiones públicas (para beneficio del Estado) y en explorar los fundamentos matemáticos de las metodologías desarrolladas (para beneficio de la disciplina).

  Desde una perspectiva histórica, nos gustaría mencionar cómo, hasta su desaparición a principios de los 80, hubo en el CSIC un instituto de Matemáticas y otro de Estadística, impulsor de la investigación en estas disciplinas en nuestro país. Años despúes se refundó el Instituto de Ciencias Matemáticas, un verdadero caso de éxito, pero no así el de Estadística. Resultaría sumamente oportuno, en un momento tan crucial de estas disciplinas, refundar un tal Instituto.[37]

---

[34] Recúerdese la polémica, en nuestra modesta opinión, desenfocada de este verano sobre este tema.

[35] Cuyo único sentido sería si deseásemos hacer estadística tras una brutal tormenta solar... Incidentalmente, me sorprende que aún se usen tales tablas en papel en la universidad.

[36] Creado por el UK Engineering and Physical Sciences Research Council en 2015 con la participación de las universidades de Cambridge, Edimburgo, Oxford, UCL y Warwick para responder a la necesidad nacional de invertir en investigación en Ciencia de Datos.

[37] Es curioso el paralelismo con la evolución de la disciplina estadística en Princeton. Durante años mantuvo un gran departamento de Estadística (por el que pasaron gigantes como Wilks, Tukey, Feller, Church, Turing y von Neumann) que cerró tiempo despúes. Sin embargo, en el año 2014 lanzó su iniciativa sobre Data Science dentro del Center for Statistics and Machine Learning.

Concluimos.

A través de ejemplos concretos hemos mostrado el potencial de los sistemas basados en el aprendizaje automático y el razonamiento estadístico para fomentar un nuevo florecimiento de la sociedad. Pero también su potencial para afectar a valores tradicionalmente aceptados, al menos en la cultura europea. En función de las vías que adoptemos conformaremos nuestra sociedad en el futuro.

En todo caso, el futuro ya está aquí. Aprendamos pues a surfear en el mar de datos para el beneficio de nuestro país.

QUEDA DICHO.

Agradecimientos.
En los desarrollos anteriores he tenido la fortuna de contar con el apoyo de numerosas instituciones y organizaciones. Los proyectos concretos mencionados incluyen al AXA Research Fund (a través de la Cátedra AXA-ICMAT en Análisis de Riesgos Adversarios), la Agencia Estatal de Seguridad Aérea, la Comisión Europea (a través del proyecto Trustonomy), Aisoy Robotics, la Fundación la Caixa, A3sec, la Fundación BBVA, el Instituto Nacional de Estadística, el Real Instituto Elcano, Xeerpa, Aitenea Biotech, Quirónprevención, CaixaBank, la National Science Foundation y la European Office for Aerospace Research and Development.

Igualmente importante es agradecer el apoyo de los colegas que han contribuido decisivamente en el desarrollo de tales proyectos. En particular, del ICMAT nos gustaría destacar a R. Naveiro, V. Gallego, A. Torres, A. Kosgodagan, J.M. Camacho, S. Liu, E. Nungesser, B. Flores, J. Aurentz, N. Campillo y M. Sanz. A. Lucía (UEM), A. Santos (UEV) y V. Ley (AEI) proporcionaron experiencia biomédica sobre el caso cardiovascular. P. Hernández-Coronado, V. Elvira y F. Bernal (AESA) aportaron experiencia aeronáutica sobre el caso de seguridad aérea y J. Gomez y C. Alfaro (URJC) apoyaron computacionalmente el proyecto. W. Caballero (AFIT) ayudó en el proyecto sobre vehículos autónomos. J.M. del Río (Aisoy Robotics) aportó requisitos sobre robótica social. J. Ríos (IBM Research), D. Banks (Duke), F. Ruggeri (CNR-IMATI), A. Salo (Aalto) y A. Tsoukias (PSL) discutieron las ideas sobre ARA y AML.

El Prof. Javier Girón, el Dr. Roi Naveiro e Isabela Ríos aportaron inestimables sugerencias y correcciones a este texto. Los eventuales errores y erratas restantes son, en cualquier caso, mi responsabilidad.

REFERENCIAS

Ashton, M. C., Lee, K., y de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review*, 18(2):139–152.

Aurentz, J., Navarro, A., y Rios Insua, D. (2022). Learning the rules of the game: An interpretable ai for learning how to play. *IEEE Transactions on Games*.

Banks, D. L., Rios, J., y Rios Insua, D. (2015). *Adversarial risk analysis*. CRC Press.

Bjornsson, Y. (2012). Learning rules of simplified boardgames by observing. In *Proc. ECAI 2012*, pages 175–180.

Burkart, N. y Huber, M. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Burns, L. y Shulgan, C. (2019). *Autonomy: The Quest to Build the Driverless Car—And How It Will Reshape Our World*. ECCO.

Caballero, W., Naveiro, R., y Rios Insua, D. (2022). Modeling ethical and operational preferences in automated driving systems. *Decision Analysis*.

Caballero, W. N., Rios Insua, D., y Banks, D. (2021). Decision support issues in autonomous driving systems. *Int. Trans. Oper. Res.*

Castillo, E., Gutierrez, J. M., y Hadi, A. S. (2012). *Expert Systems and Probabilistic Networks*. Springer.

Claussmann, L., Revilloud, M., Gruyer, D., y Glaser, S. (2019). A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–23.

Comiter, M. (2019). *Attacking Artificial Intelligence*. Belfer Center Paper. Cox, T. (2008). What's wrong with risk matrices. *Risk Analysis*, 28:497–512.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., y Verma, D. (2004). Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 99–108.

Daniell, K., Morton, A., y Rios Insua, D. (2016). Policy analysis and policy analytics. *Annals of Operations Research*, pages 1–13.

Durrant-Whyte, H. y Bailey, T. (2006). Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110.

Elvira, V., Bernal, F., Hernandez-Coronado, P., Herraiz, E., Alfaro, C., Gómez, J., y Ríos Insua, D. (2020). Safer skies over spain. *INFORMS Journal Applied Analytics*, 50:21–36.

Fan, J., Ma, C., y Zhong, Y. (2021). A selective overview of deep learning. *Statistical Science*, 36:264–290.

French, S. y Rios Insua, D. (2000). *Statistical Decision Theory*. Wiley.

Gallego, V., Naveiro, R., Roca, C., Campillo, N., y Rios Insua, D. (2022). AI in drug development: a multidisciplinary perspective. *Molecular Diversity*.

Gallego, V. y Rios Insua, D. (2021). Variationally inferred sampling through a refined bound. *Entropy*, 23:123.

Gallego, V. y Rios Insua, D. (2022). Current developments in neural networks. *Annual Reviews in Statistics*.

Gao, J., Su, Y., Tomonaga, M., y Matsuzawa, T. (2018). Learning the rules of the rockpaperscissors game: chimpanzees versus children. *Primates*, 59:7–17.

Gardner, H. (2011). *Frames of mind: the theory of multiple intelligences*. Hachette UK.

Girón, J. (2021). *Bayesian testing of statistical hypothesis*. RAC.

Glimcher, P. W. y Fehr, E. (2013). *Neuroeconomics: decision making and the brain*. Academic Press.

Goldberg, L. (1990). An alternative ”description of personality”: The big-five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229.

Goodfellow, I. J., Shlens, J., y Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hargreaves-Heap, S. y Varoufakis, Y. (2004). *Game theory: A critical introduction*. Routledge.

Jarvis Thomson, J. (1985). The trolley problem. *Yale Law Journal*, pages 1395–1415.

Kahneman, D. (2011). *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York.

Krizhevsky, A., Nair, V., y Hinton, G. (2014). The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55:5.

LeCun, Y., Cortes, C., y Burges, C. (1998). THE MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Liu, S. y Rios Insua, D. (2020a). An affective decision-making model with applications to social robotics. *EURO J Decis Process*, 8:13–39.

Liu, S. y Rios Insua, D. (2020b). Group decision making with affective features. *Group Decis Negot*, 29:843–869.

Loewenstein, G. y Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, 619(642):3.

Lundberg, S. M. y Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., y Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50 (4):370.

McAllister, R., Gal, Y., Kendall, A., van der Wilk, M., Shah, A., Cipolla, R., y Weller, A. (2017). Concrete problems for autonomous vehicle safety: advantages of bayesian deep learning. In *Proc. 26th IJCAI*.

Mellers, B. A., Schwartz, A., y Cooke, A. D. J. (1998). Judgment and decision making. *Annual Review of Psychology*, 49(1):447–477.

Nielsen, T. y Jensen, F. (2008). *Bayesian Networks and Decision Graphs*. Springer, New York.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Picard, R. W. (1997). Affective Computing. *Encyclopedia of Multimedia Technology and Networking, Second Edition*, (321):15–21.

Ribeiro, M. T., Singh, S., y Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Rios Insua, D., Caballero, W., y Naveiro, R. (2021a). Managing driving modes in automated driving systems. *arXiv: 2107.00280*.

Rios Insua, D., Camacho, J. M., Ley, V., Santos, A., y Lozano, A. (2021b). A predictive bayesian network model for cardiovascular diseases. Technical report, ICMAT.

Rios Insua, D. y Gómez-Ullate, D. (2019). *¿Qúe sabemos de? Big Data*. La Catarata.

Rios Insua, D., Naveiro, R., y Gallego, V. (2020a). Perspectives on adversarial classification. *Mathematics*, 8(11).

Rios Insua, D., Naveiro, R., Gallego, V., y Poulos, J. (2020b). Adversarial machine learning: Perspectives from adversarial risk analysis. *arXiv preprint arXiv:2003.03546*.

Robbins, H. y Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Roumeliotis, S. I. y Bekey, G. A. (1997). Extended Kalman filter for frequent local and infrequent global sensor data fusion. In *Sensor Fusion and Decentralized Control in Autonomous Robotic Systems*, volume 3209, pages 11–22.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Russell, J. A. y Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819.

Samek, W., Wiegand, T., y Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Society of Automobile Engineers (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Technical report, SAE.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., y Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

The Alan Turing Institute (2021). Data science and ai in the age of covid-19. Technical report, AT Institute.

Tversky, A. y Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2):105.

Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Rayner, M., y Townsend,

N. (2017). European cardiovascular disease statistics. Technical report, European Heart Network.

Wu, B., Iandola, F., Jin, P. H., y Keutzer, K. (2017). Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 129–137.