Seminario 'Huella digital: ¿servidumbre o servicio?

La evaluación por terceros de procesos de *big data*

(Síntesis de la sesión del 18 de febrero de 2021)

El comité de expertos del seminario *La Huella Digital: ¿servidumbre o servicio?* celebró su décima sesión el pasado 18 de febrero por videoconferencia. La sesión se inició con unas reflexiones de la dirección del seminario sobre el avance en los temas ya tratados, sobre algunas cuestiones en las que se requiere más información, y sobre el tipo de conclusiones a las que se podrá llegar al término del seminario, previsto para el mes de junio. Estas conclusiones podrían formularse principalmente en los campos de la normativa, la autorregulación, y la formación de capacidad crítica. A continuación, se prosiguió con el debate sobre la naturaleza de los *sistemas de inteligencia artificial* y la evaluación por terceros de procesos de *big data*.

La ponencia inicial corrió a cargo de José Luis Calvo, Director de Inteligencia Artificial de SNGULAR, y se siguió con comentarios de Pablo García Mexía, Consultor-Director de Derecho Digital en Herbert Smith Freehills, Letrado de las Cortes, y de Gloria Sánchez Soriano, directora de la Asesoría Jurídica de Tecnología y Transformación del Grupo Santander.

Desmitificar la IA

A lo largo de las sesiones anteriores se ha hablado frecuentemente de un cierto mito alrededor de los sistemas autónomos e inteligentes. Para José Luis Calvo, aterrizar el concepto supone ante todo, recordar que la IA es software. Aunque es en los últimos cinco años cuando la IA ha comenzado a dar unos resultados sorprendentes, su historia se remonta al surgimiento de los primeros ordenadores.

El auge actual se ha dado en el desarrollo reciente de las llamadas redes neuronales profundas, que se enmarcan en las técnicas de aprendizaje automático. En un principio, se partía de unos algoritmos deterministas, un conjunto de instrucciones precisas que la máquina ha de ejecutar. Estos algoritmos eran creados de forma artesanal por desarrolladores, que controlaban las casuísticas a las que el software habría de enfrentarse. Partiendo de esta base determinista, estable durante los últimos cincuenta años, han ido apareciendo las técnicas de aprendizaje automático o algoritmos probabilísticos. En estas técnicas, el algoritmo tiene un fundamento estadístico y es entrenado con un conjunto de datos que ya contienen un resultado y que condicionan el entrenamiento. Basado en un patrón dado, el algoritmo se va ajustando hasta

crear un modelo o programa. De tal forma que, al recibir un nuevo conjunto de datos, el modelo es capaz de generar resultados por sí solo. A diferencia del tradicional, el software creado con técnicas de IA es capaz de crear su propio modelo, que va más allá de unos pasos y reglas pautadas por un desarrollador. Dentro de los modelos, algunos son comprensibles, como los árboles de decisión, pero otros son complejos y difícilmente explicables: las llamadas "cajas negras". El avance en los modelos va esencialmente ligado a la disponibilidad de grandes volúmenes de datos y a una capacidad de computación mucho más amplia. (Evidentemente, el concepto de "cantidades ingentes de datos" es variable: lo que para una pequeña o mediana empresa es una cantidad ingente de información, para una compañía internacional no es siquiera algo representativo).

En resumen, la evolución de los sistemas de IA en las últimas décadas pasa de unos modelos *top down* (basados en pautas dadas y su aplicación) a un nuevo tipo de modelos probabilísticos que son posibles gracias al aumento en el volumen de datos y al aprendizaje de las máquinas: los modelos *bottom up*, como las ya mencionadas redes neuronales.

Problemática actual y potencial de la IA

Los problemas usualmente relacionados con el uso de la IA no son nuevos, pero la naturaleza de estos nuevos modelos hace que aquellas entre sus características que puedan ser potencialmente dañinas, lo sean a una escala mucho mayor y para muchos más millones de personas. Los procesos siguen siendo utilizados para automatizar actividades previamente realizadas por personas, pero con un coste marginal que aparentemente tiende a cero; es decir que se podría extender y crecer casi sin límites. Ahora bien, los expertos reconocen que estos procesos suponen una demanda de capacidad de computación y unos volúmenes de almacenamiento de datos considerables; sería interesante disponer de datos al respecto para poder estimar los límites físicos y los costes reales (externalidades) de los sistemas de IA. También cabe preguntarse sobre la sostenibilidad del "modelo de negocio" montado en la explotación de la "huella digital" y en la imputación de fuertes costes de publicidad a los anunciantes, con sustanciales márgenes de rentabilidad para las *big tech*.

¿Cuáles son los aspectos sobre los que podría incidir la evaluación por terceros de los procesos de *big data* y el uso de la IA? Se mencionan cuatro grupos de problemas: los referentes a la desinformación y polarización; a la privacidad; a los problemas de discriminación; y aquellos que surgen de la vigilancia a los usuarios.

Los problemas de *desinformación y polarización* surgen por los algoritmos de recomendación. Este tipo de algoritmos no solo se utilizan para recomendar a los usuarios qué película o libro será de su agrado, sino para dirigirles noticias y contenidos que contribuyen a fomentar un punto de vista posiblemente sesgado, y que favorecen la difusión de noticias falsas. En contraparte

aparece el riesgo de una limitación de la libertad de expresión por parte de Estados o de las *big tech*, justificada por la necesidad de evitar la divulgación de contenidos falsos o abusivos. De hecho, las plataformas están aplicando prácticas censoras con el uso de algoritmos destinados a bloquear el acceso a determinados contenidos, lo que plantea la cuestión de la legitimidad o la arbitrariedad de tales intervenciones.

Los problemas de *privacidad* son posiblemente los más reconocidos por la sociedad puesto que la materia prima de los algoritmos son los datos que se obtienen de la actividad de nuestra "huella digital". Son los datos que ceden los usuarios los que alimentan y hacen posibles las predicciones. Pero en la obtención y manipulación del dato surgen problemas de privacidad como en los casos en que el algoritmo, combinando datos cedidos más o menos voluntariamente, es capaz de reconstruir una información no cedida. Un ejemplo de ello es el caso en el que el sistema de IA es capaz de obtener información del sexo de una persona a través de la imagen de su retina. Como este, hay muchos ejemplos en los que el modelo se entrena de tal forma que permite predecir información privada del usuario. Así, los usuarios ceden unos datos a partir de los cuales se deducen otros, que no necesariamente querrían hacer públicos, y que podrían no utilizarse para su propio beneficio. Este es quizás el meollo del asunto del "excedente comportamental" construido a partir de nuestras huellas digitales y con el que se pretende, con o sin nuestro asentimiento, predecir con creciente precisión nuestro propio comportamiento. Sobre ello sería interesante que el seminario dispusiera de algunos elementos de información adicionales.

En lo referente a la *discriminación*, los problemas parten del hecho de que el entrenamiento del algoritmo se hace a base de un histórico de datos desde el que se crean nuevos patrones, que reflejarán situaciones selectivas anteriores. Esto no sólo puede suponer problemas de justicia, cuando alguien es valorado partiendo de un histórico de datos que determina las decisiones del sistema a favor o en contra de la persona, sino que, además, ello puede implicar que se refuercen discriminaciones previas y se perpetúe el sesgo. Ciertamente en algunas ocasiones los datos tienen oculto un sesgo que los desarrolladores en un principio no ven. En ese caso, puede ser que los nuevos sistemas estén contribuyendo a fomentar una discriminación que como sociedad queremos evitar. En otras ocasiones, el sesgo es el resultado de una falta de esfuerzo por incluir diversidad en el dato, o incluso puede responder a una intención manipuladora por parte del desarrollador o de sus mandantes. En cualquier caso, es claro que tomar decisiones automatizadas basándose solo en información del pasado implica unos límites que pueden producir efectos no ajustados a la realidad actual.

La exactitud del modelo es clave para reducir y/o evitar los falsos positivos (como conceder un crédito a quien no puede pagarlo) y los falsos negativos (como no conceder un crédito a quien lo merece). La sintonía fina de los modelos debería atender ambos extremos de tal forma que, en el ejemplo mencionado, reducir los falsos negativos implicaría que se concedan créditos a un mayor número de personas que pueden ser buenos deudores, y reducir falsos positivos

conllevaría dejar de conceder crédito a personas que pueden no ser capaces de devolver el crédito concedido, evitando asimismo un sufrimiento directo a estas personas e indirecto a la colectividad.

El último riesgo mencionado, el de *vigilancia*, es objeto de múltiples estudios y se ha debatido ampliamente en medios de difusión en distintos países, siendo quizás el que implique mayores amenazas de futuro. Se plantea, entre otras situaciones, en los escenarios en que programas como los que permiten el control de tráfico, la grabación y la identificación vocal, o el reconocimiento facial para la seguridad de los aeropuertos pueden utilizarse para hacer otro tipo de seguimientos que podrían violar la intimidad y los derechos de los individuos. Se trata de un debate abierto ante el que empresas como IBM, Amazon y Microsoft, por ejemplo, han dado un paso atrás en la comercialización de herramientas de reconocimiento facial.

El control de la IA

Con el fin de controlar y fomentar el desarrollo y buen gobierno de los modelos de IA, instituciones gubernamentales, empresariales y académicas han propuesto diversos principios y códigos éticos para el uso de una IA centrada en humanos. El debate se sitúa entre quienes tienden a promover ante todo la intervención gubernamental – previa o a posteriori – y quienes ven más futuro en la autorregulación, el buen diseño de los sistemas y la posibilidad de certificar su funcionamiento por fuentes independientes.

Con el fin de controlar y fomentar el desarrollo y buen gobierno de los modelos de IA, instituciones gubernamentales, empresariales y académicas han propuesto diversos principios y códigos éticos para el uso de una IA centrada en humanos. Entre los códigos usualmente sugeridos, se encuentra una base propuesta, entre otros, por autores como la ingeniera española Nuria Oliver ("Fair, transparent, and Accountable Algorithmic Decision-making Processes", artículo publicado en 2017 en la revista *Philosophy & Technolgy*): autonomía, justicia, beneficencia y no-maleficencia, es decir los cuatro principios de la bioética, y uno más que responde a la transparencia necesaria para un desarrollo ético de los modelos de IA.

El control de los modelos de IA debe atender a las tres piezas clave de los sistemas: los datos utilizados para el entrenamiento del algoritmo, el tipo de algoritmo en relación con el escenario de aplicación, y el resultado del modelo.

Con respecto a la primera de estas piezas clave – los datos – Europa se ha posicionado con el Reglamento General de Protección de Datos, ampliamente comentado en sesiones anteriores del seminario. Se sugieren ideas adicionales, como la introducción de estudios éticos que partan desde el diseño del modelo o la prohibición de la comercialización de datos personales, de forma

análoga al tráfico de órganos: en esta perspectiva, los datos personales podrían utilizarse en determinados contextos, pero no podría comercializarse con ellos.

En cualquier caso, los datos son la materia prima de la IA; por tanto, atender directamente a aspectos como la calidad del dato es esencial para evitar problemas en la evaluación por parte de terceros. Así, los datos deben provenir de fuentes fiables; el modelo deberá contener la cantidad necesaria de datos para evitar o minimizar el riesgo de sesgo, de tal forma que se asegure la mayor exactitud posible; y todo ello implica que los datos no hayan sido objeto de alteración. Si los datos para el entrenamiento cumplen con estos requisitos, podrán constituir un buen punto de partida.

Respecto al algoritmo se menciona la posibilidad de regular las tipologías a la hora de crear un modelo. Por ejemplo, se podría prohibir el uso de modelos de "caja negra" en los casos en que puedan afectar a personas y sus efectos puedan suponer un mayor riesgo. En estos temas, como por ejemplo la video-vigilancia o los contratos laborales, podría permitirse solo el uso de modelos de fácil comprensión, mitigándose así la tensión entre corrección estadística y justicia individual. Otros participantes ven más factible, en lugar de regular el algoritmo, regular los casos de uso y las finalidades para las que se usa, considerando los niveles de riesgo que entrañan, pues en muchas ocasiones la misma tecnología puede usarse para distintos casos de uso. Ello puede ser objeto de una fructífera reflexión preventiva por parte de empresas e instituciones. Sin embargo, pretender obtener estas limitaciones mediante disposiciones legales podría acabar limitando los usos socialmente positivos de los procesos y equivaldría a querer "poner puertas al campo".

Finalmente, la evaluación de la IA debe atender al resultado del modelo, de tal forma que dichos resultados sean auditables. Un algoritmo puede utilizar datos que respeten la privacidad y los sesgos, ser explicable y transparente, pero aun así sus resultados pueden incumplir los paradigmas de justicia social. Casos como este son, por ejemplo, los modelos utilizados para la identificación de caras que funcionan mejor con hombres de raza blanca. Algunas voces del seminario subrayan que es primordial atender al hecho de que, si el resultado es un reflejo de la realidad, aunque esa realidad no siempre guste (o resulte políticamente incorrecta y/o sorprendente), no por ello deja de ser veraz o real. Es decir, cabría evitar caer en el error de dejar de analizar la realidad porque la información pueda asustarnos. Quizá se podría proponer un análisis por comparación con el resultado obtenido sobre los mismos datos, pero con otros modelos, o tal vez buscar otras formas de validación del resultado.

Sobre el tipo de regulación

Frente a todos los intentos de regulación o evaluación preventiva o voluntaria, hay quien defiende que la única regulación eficaz es *a posteriori*, sobre los resultados finales. Como en

otros campos, la ley sancionaría las situaciones en las que ha ocurrido algo intolerable. Ante ello, la respuesta se aplicaría en forma de multa, y posiblemente una multa o sanción contundentes. Esta propuesta pretende salvar las desventajas competitivas que traen las leyes a los países con regulación estricta: sin importar en dónde se haya desarrollado el algoritmo y bajo qué leyes, este deberá respetar unas normas preestablecidas o será sancionado con una multa. Además, este enfoque es un potente aliciente a un comportamiento preventivo por parte de los agentes, que intentarán evitar la sanción.

Sin embargo, y sin que ello sea necesariamente incompatible con lo anterior, algunas voces defienden la necesidad de que la exigencia ética parta desde el diseño. Los algoritmos deberían ser auditables por diseño y la interpretabilidad debería también ser incluida por diseño, ya sea por normativa, ya sea por códigos de conducta. Apelar al diseño implica defender el pensar antes de hacer. Aquí, se propone trasladar el concepto "by design" a las problemáticas de la gestión de datos y hablar así de la responsabilidad "by design", de la ética "by design" y del control "by design". Esto conlleva una comprensión de los mecanismos que se quieren controlar, de los riesgos y de las fases que hay que establecer para los requerimientos y controles, así como las áreas de la organización que deben hacerse responsables de vigilar el cumplimiento. Una propuesta para avanzar en posibles soluciones plantea establecer seis fases de control: la conceptualización, los datos, el desarrollo del modelo, el lanzamiento, el control y *feedback*, y finalmente la gestión de errores. Cada una de estas fases abarca el control de los diversos problemas que se presentan en la gestión de datos por parte de las empresas. Con estas seis fases en mente, cada organización debería adaptar el sistema de control a sus procesos. Todo ello puede ser objeto de certificación independiente, como ocurre con otros procesos de fabricación o de organización.

Con relación a todo ello, se hace referencia a la necesidad de identificar en cada caso el lugar donde situar la responsabilidad jurídica que cubra los daños ocasionados por el empleo de algoritmos, sobre todo en los casos de alto riesgo para la integridad y los derechos individuales. En materia jurídica parece haberse diluido la idea (barajada en Europa después del 2015) de construir un tipo de "personalidad electrónica" para los sistemas inteligentes, similar a la de las sociedades, asociaciones o fundaciones. Pero de momento, parece que esta idea se ha abandonado pues, aunque pueda verse como similar a las nombradas, ya no estaríamos hablando de un "compuesto" de seres humanos, sino que se extendería el concepto de persona más allá del ámbito humano. Parece más adecuado exigir que siempre haya una persona, más allá de los algoritmos, que sea responsable de la actividad y los resultados del sistema.

El sesgo de automatización

Antes incluso de los riesgos citados – como la polarización, la privacidad, los sesgos de datos y la vigilancia – se vislumbra otro riesgo enraizado, el denominado sesgo de automatización, o sea una tendencia a creer ciegamente en los resultados que arrojan los sistemas automatizados. Ante ello es preciso reivindicar que el ser humano no debe abandonar su capacidad de juicio, ni delegar su razonamiento a los sistemas robotizados. En la época actual, tal y como se ha mencionado en otras sesiones, el individuo tiene una responsabilidad de mantenerse libre. La sociedad en conjunto tiene la responsabilidad de desarrollar un juicio crítico sobre los asuntos de su entorno, de entender la realidad, procurar mejorarla y hacerla más justa. Es humano confiar en lo técnico, pero el problema sería abandonarnos a una corriente en la que se ceda a las máquinas el razonamiento sin ningún tipo de análisis por parte del hombre.

En esta línea es interesante señalar que ha habido ocasiones a lo largo de la historia en las que el desconocimiento ha llevado a aceptar como dadas unas proposiciones que finalmente se vio que no eran verdaderas. Por ejemplo: a mediados del siglo XX, con el invento de la televisión, los anuncios de publicidad adquirieron una fuerza inusitada: en un principio no se ponía en duda lo que se anunciaba, así que amas de casa, oficinistas y hombres de negocio se mostraban dispuestos a comprar lo anunciado sin cuestionarse sobre las técnicas de mercadeo. Algo parecido ocurre ahora con la IA: el desconocimiento tiene un papel relevante en la importancia que se le da en sociedad. La opacidad con la que funcionan los modelos genera importantes riesgos sociales, además lo impredecible de sus consecuencias añade un factor de riesgo aumentado. Pero es posible que los procesos automáticos “superinteligentes” se encuentren en un estado de evolución aún bastante primitivo y que, con su posterior evolución, se seguirán corrigiendo con elementos del insustituible “sentido común” del hombre.

Con o sin regulación fuerte, no se puede obviar por otro lado el necesario aumento en la educación y el fomento del conocimiento de la sociedad en todas estas materias, con el fin de forjar un sentido crítico al respecto. Aunque el tema interese a toda la sociedad y se deba tender a la educación social general en materia de modelos de IA, es esencial empezar por los tecnólogos, para que estos puedan ver más allá del propio desarrollo del sistema y se preocupen por la responsabilidad en el diseño de la IA. Las universidades deberían fomentar la educación de los ingenieros para que durante todo su periodo de formación tengan presentes los problemas y los retos sociales que la tecnología implica. También es imprescindible que la educación en estos temas abarque a las empresas que compran los algoritmos o que piden unos resultados específicos a los ingenieros y desarrolladores. Si los altos cargos de las empresas piden a los técnicos lo que quieren que se haga sin pensar en el impacto, la responsabilidad estará también en quien manda que se haga un determinado servicio, y no sólo en el desarrollador del modelo. Y es necesario que exista un lenguaje común al respecto entre quienes toman decisiones empresariales y quienes desarrollan los programas informáticos.

La educación, como herramienta ante el sesgo de automatización, supone la clave para evitar que la tecnología reemplace al ser humano en el razonamiento que le lleva a tomar las decisiones respecto al rumbo que socialmente se quiere y debe tomar. Hay una tendencia a la confianza ciega en los resultados de los algoritmos, esto debe evitarse desde la legislación y la educación. Lo dicho implica una defensa de la libertad humana, característica que forma parte de la esencia intrínseca del hombre, y que no puede ser invalidada por las máquinas. En temas de educación, no se trata tanto de formar las mentes en una disciplina concreta, sino de una formación que tienda a la concienciación de la realidad tan impactante y ciertamente desconocida para muchos que está en las dinámicas del mundo digital.

Como reflexión final, algunos hacen hincapié, una vez más, en los extraordinarios beneficios que los sistemas inteligentes pueden brindar cuando se ponen al servicio de un proyecto de investigación innovador y desinteresado. Ejemplo de este tipo puede ser el Proyecto *Brain* (Cerebro Humano) liderado por el neurobiólogo español Rafael Yuste, proyecto médico, científico y tecnológico que mediante técnicas avanzadas se propone conocer en profundidad y mapear la "caja negra" del cerebro humano. La posibilidad de leer y quizás de alterar la actividad cerebral abre inmensas posibilidades y plantea la necesidad de criterios éticos en una dimensión aún desconocida. Sobre estos temas, el seminario dedicará su próxima sesión, prevista para el 15 de abril.

Listado de asistentes:

1. **Alfonso Carcasona**, Consejero Delegado, AC Camerfirma
2. **Alfredo Marcos Martínez**, Catedrático de Filosofía de la Ciencia, Universidad de Valladolid
3. **Ángel González Ferrer**, Director Ejecutivo Centro Cultura Digital Consejo Pontificio para la cultura del Vaticano
4. **Carolina Villegas**, Investigadora de la Cátedra Iberdrola de Ética Financiera y Empresarial, Universidad Pontificia de Comillas
5. **Cristina San José**, Chief Data Strategist, Banco Santander
6. **David Roch Dupré**, Profesor de la Universidad Pontificia Comillas
7. **Diego Bodas Sagi**, Lead Data Scientist – Advanced Analytics, Mapfre España
8. **Domingo Sugranyes**, Director del Seminario de Huella Digital
9. **Esther de la Torre**, Responsible Digital Banking Manager, BBVA
10. **Francisco Javier López Martín**, Ex-Secretario General , CCOO de Madrid

11. **Gloria Sánchez Soriano**, Directora de Asesoría Jurídica de Tecnología, Costes y Transformación de grupo Santander
12. **Idoia Salazar**, Experta en Ética en IA, Universidad CEU San Pablo
13. **Idoya Zorroza**, Profesora Contratada Doctora, Facultad de Filosofía, Universidad Pontificia de Salamanca
14. **Javier Camacho Ibáñez**, Director de Sostenibilidad Ética y profesor de ICADE e ICAI
15. **Javier Prades**, Rector, Universidad Eclesiástica San Dámaso
16. **Jesús Avezuela**, Director General de la Fundación Pablo VI
17. **Jesús Sánchez Camacho**, Profesor de la Facultad de Teología, Universidad Pontificia Comillas
18. **José Luis Calvo**, Director de Inteligencia Artificial en SNGULAR
19. **José Manuel González-Páramo**, Consejero Ejecutivo, BBVA
20. **José Ramón Amor,** Coordinador del Observatorio de Bioética de la Fundación Pablo VI
21. **Juan Benavides**, Catedrático de comunicación, Universidad Complutense de Madrid
22. **Julio Martínez s.j.**, Rector, Universidad Pontificia Comillas
23. **Pablo García Mexía**, Jurista Digital, Consultant - Head of Digital Law Herbert Smith Freehills
24. **Paul Dembinski,** Director de Observatoire de la Finance (Ginebra)
25. **Raúl González Fabre**, Profesor, Universidad Pontificia de Comillas
26. **Sara Lumbreras**, Subdirectora de resultados de investigación, Profesora titular, Instituto de Investigación Tecnológica , ICAI, Universidad Pontificia Comillas
27. **Richard Benjamins**, Embajador de Data & IA, Telefónica