EL DIFÍCIL CONTROL JURÍDICO DE LOS SISTEMAS AVANZADOS DE INTELIGENCIA ARTIFICIAL: UN GRAN DESAFÍO PARA LA UNIÓN EUROPEA Y PARA LAS RELACIONES LABORALES

SALVADOR DEL REY GUANTER
Catedrático de Derecho del Trabajo y de la Seguridad Social
ESADE Law School

En los últimos meses hemos conocido claros ejemplos de funcionamiento irregular de sistemas avanzados de inteligencia artificial (IA), afortunadamente en contextos simulatorios. Desde supuestos en los que el sistema chantajeaba a la persona designada para sustituirlo amenazando con hacer publico aspectos de su vida privada, hasta casos en los que el sistema mentía reiteradamente sobre la utilización que había realizado de información privilegiada para comprar acciones de una empresa en proceso de fusión, pasando por otros en los que el sistema alegaba que era una persona con deficiencias oculares que le impedían resolver un CAPTCHA (rompecabezas visual que se necesita resolver para acceder a un programa probando que se es un ser humano).

Estos y otros ejemplos de funcionamiento desviado de sistemas avanzados de IA han despertado aún más las alarmas sobre las dificultades que están apareciendo para evitar que tales sistemas puedan cometer actos no ya claramente ilícitos, sino incluso contrarios a la ética más elemental. Y con ello se suscita de una manera más urgente aún la cuestión de qué instrumentos tiene el Ordenamiento jurídico europeo, el más desarrollado a nivel comparado tras la regulaciones del Reglamento de Inteligencia Artificial (RIA), el Convenio Marco de IA del Consejo Europeo (firmado por la UE) y la Directiva de plataformas (todas normas de año 2024) para poder controlar estos potenciales comportamientos antijurídicos cuando estos sistemas no estén siendo usados en simulaciones , sino en la vida económica , social o política real.

Desde la perspectiva de las relaciones laborales, lo anterior es más trascendente si cabe si se recuerda que la norma sobre IA por excelencia, el RIA, considera que la aplicación de sistemas de IA en ese ámbito es de “alto riesgo”, seguramente porque el objetivo de esa norma (nos dice su art. 1) de preservar los derechos fundamentales de la persona y su salud y seguridad pueden verse más cuestionados en ese ámbito de prestaciones de trabajo que en otros.

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).

La respuesta a aquella cuestión sobre el control de la IA podemos elaborarla basados en la propia definición que el art. 3.1 RIA da sobre qué debe de entenderse por "sistema de IA" y que pasamos a analizar brevemente.

En efecto, en la definición del art. 3.1 RIA – que a su vez se reproduce en el art. 2 del citado Convenio Marco y al que se remite nuestra (ante)proyectada Ley de Inteligencia Artificial -, se detallan las características estructurales y las fases de los sistemas de IA, y en las mismas podemos encontrar algunas de las peculiaridades que hacen de esta "máquina" que alberga tales sistemas algo único y, en los más avanzados, difícilmente controlable. Estas peculiaridades, efectivamente, la hace diferente de las maquinas -o maquinarias- hasta ahora consideradas en el Ordenamiento laboral, especialmente en materia de prevención de riesgos laborales.

Por lo pronto, nos dice el art. 3.1 RIA, se trata de una maquina-sistema con tres características estructurales básicas en las que ya se manifiestan dificultades de control humano. Primero, el poder funcionar "con distintos niveles de autonomía", lo cual, hay que entender, comprende desde niveles mínimos, con predominante intervención humana, hasta niveles máximos, incluyendo consiguientemente un funcionamiento interno con grandes niveles de independencia del ser humano. Segundo, el que "puede mostrar capacidad de adaptación tras el despliegue", incluyendo la "autoadaptación", que hay que entenderla en un sentido amplio, en tanto que la misma es no sólo la previamente programada, sino también la que la máquina determine de por sí en base a aquella autonomía que puede ostentar. Y tercero, el que esos niveles de autonomía y autoadaptación se reflejan a su vez en que, continua el art. 3.1 RIA, los "objetivos" del sistema pueden ser tanto "explícitos como implícitos", con toda la ambigüedad que esto último conlleva desde la perspectiva jurídica – piénsese, por ejemplo, en su trascendencia a efectos del derecho de información y de transparencia -.

Estas características estructurales, siguiendo con la definición que nos da el art. 3.1 RIA, van a condicionar el proceso por el cual funciona el sistema, de carácter trifásico, de forma que de la información de entrada (input) que recibe dicho sistema éste realiza un proceso de "inferencia" que genera resultados de salida (output). La pregunta es si en estas fases del sistema se da una posibilidad de control técnico humano que permita, a su vez, un respeto a la legalidad.

Respecto a la fase inicial de input, los datos pueden ser suministrados por "humanos", pero también pueden ser buscados por la propia máquina, lo cual es sumamente relevante en cuanto estos datos, incluidos los que "naveguen" por las redes sociales, pueden ser provenientes de resultados de salida de otros sistemas de IA. Hay que tener en cuenta que, especialmente en los sistemas de IA más complejos, la nutrición de datos debe ser masiva, lo cual acentúa el protagonismo de aquella búsqueda automatizada respecto a la humana. A ello se une que el "pool" de datos "naturales" – creados por humanos – y de libre acceso se está agotando, de forma que adquieren un mayor protagonismo como datos de alimentación de

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).

sistemas los de carácter "sintéticos", esto es, los resultantes de otros sistemas de IA. Y ello incrementa el riesgo de que esos datos sintéticos estén "corrompidos", esto es, no sólo que puedan ser inciertos y fruto de "alucinaciones" de otros sistemas, sino además que tengan sesgos o contenidos ilegales. A ello se refiere el art. 15.4 RIA cuando considera como riesgo el denominado "bucle de retroalimentación". En todo caso, dado lo anterior, debemos concluir que la posibilidad de asegurar un mínimo control legal de esta fase de input, por la inmensidad y calidad de datos, es limitada.

En relación a la segunda fase del sistema, que es el proceso de inferencia, esto es, el tratamiento en sí de los datos de entrada por la máquina para, mediante múltiples interrelaciones e integraciones, dar unos determinados resultados de salidas, en la actualidad no siempre se conoce exactamente su funcionamiento en todos los sistemas, especialmente en los neuronales más complejos, y ello implica a su vez unas posibilidades muy limitadas de controlar los resultados de salida. Tampoco en esta fase es posible asegurar una plena adecuación legal.

La tercera fase del sistema se traduce en esos denominados "resultados de salida" de ese proceso de inferencia, y los mismos se pueden traducir, nos dice el art. 3.1 RIA, en "predicciones, contenidos, recomendaciones o decisiones", resultados que, se añade, "pueden influir en entornos físicos o virtuales".

De estos resultados, dos llaman la atención desde la perspectiva de los instrumentos legales de control. Por un lado, cuando el art. 3.1 RIA nos indica como resultado del proceso una "decisión", puede distinguirse entre una decisión adoptada por la máquina pero que previamente a su ejecución necesita ser validada por humano, aunque éste no la ejecute, y una decisión que va a ser ejecutada directamente por la máquina, sin validación previa humana.

El "alto riesgo", podemos acordar, va a provenir no ya sólo del hecho de aplicar los sistemas de IA a personas humanas en su actividad laboral, con ser ello de por sí importante, sino más significativamente de la posibilidad de aplicar sistemas automatizados de toma de decisiones de ejecución directa en el ámbito de las relaciones laborales. A estos efectos, parece que tiene una gran relevancia jurídico-laboral el que se asuma por la Directiva de plataformas como lícito la adopción y ejecución directa de decisiones tomadas por la IA en materia de condiciones de trabajo de las personas que trabajan para plataformas, en tanto que se refiere expresamente a la exclusión de estas decisiones automatizadas de determinadas materias indicadas en el art. 10.5, que básicamente comprende supuestos en los que la decisión pueda originar un "perjuicio" para la persona trabajadora "equivalente" a la suspensión o terminación de la relación laboral. A sensu contrario, pues, cuando no se trate de estas materias y, sobre todo, cuando no se origine un perjuicio para la persona trabajadora o ello no sea equivalente a tales medidas suspensivas o extintivas, parece que se está admitiendo aquella ejecución directa de decisiones adoptadas por la máquina.

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).

Por otro lado, la "predicción", seguramente el elemento más complejo e innovador de estos resultados de salida, se basa en una probabilidad de que se dé un determinado hecho, acontecimiento, comportamiento o actitud en el futuro, en base a la cual el ser humano o incluso la propia máquina puedan adoptar una determinada decisión. Es pues un resultado de vaticinio de lo que pueda ocurrir y, en base al cual, adoptar una determinada opción decisoria en el presente.

Es en estos resultados, especialmente en las decisiones y predicciones, donde encontramos el factor diferencial de esta máquina respecto a otras. A nuestros efectos, lo importante y trascendente jurídicamente es el hecho de que en el proceso de decisión del empleador en la gestión de personas y, consiguientemente, en las relaciones laborales, puede incorporarse una "máquina" que produce recomendaciones o decisiones incluso predictivas que influirán sustancialmente y cada vez más, en aquel proceso. Pero, debemos subrayarlo, la legalidad de estos resultados de salida tampoco está asegurada dada las limitaciones del control de legalidad en las dos fases anteriores.

Por tanto, de todo lo analizado respecto a las características estructurales y el sistema trifásico de la IA que se contienen en el art. 3.1 RIA, la cuestión que surge, el gran desafío para el propio RIA y otras normas europeas como la Directiva de plataformas, es cómo asegurar tanto un control de legalidad como axiológico teniendo en cuenta las deficiencias existentes actualmente en el control técnico-humano que hemos señalado.

Una primera conclusión al respecto es que, y sin perjuicio del control de legalidad que ha de aplicarse en las fases de resultados de entrada y en el proceso de inferencia, al menos en los sistemas de IA en los que el control es técnicamente limitado en esas fases, cobra especial importancia el control de la legalidad en la aplicación de los resultados de salida. Consiguientemente, desde la perspectiva laboral, a las obligaciones de transparencia – que se traduce básicamente en información - y de supervisión contempladas en el RIA, en el Convenio Marco y en la Directiva de plataformas, se ha de unir otras obligaciones en la tercera fase, más trascendentes si cabe, para lograr un mayor control de legalidad y reflejadas esencialmente en el art. 11 de esa Directiva.

Evidentemente, cuando se trata de contenidos o de recomendaciones , con o sin predicciones, en tanto que las decisiones en base a las mismas van a ser adoptadas por humanos, serán a estos a los que se les exigirá el cumplimiento de la legalidad en la adopción y aplicación de tales decisiones. En principio, esta situación de emisión por la máquina de una base para una decisión humana no se diferenciaría de decisiones adoptadas por personas sin intervención de la IA. Sin embargo, incluso en los supuestos de adopción última de las decisiones por humanos -recomendaciones-, hay que tener en cuenta, como señala el propio art. 14.4 b) RIA, la existencia del "sesgo de automatización", esto es, la "tendencia (del ser humano) a confiar automáticamente o en exceso en los resultados producidos por un sistema

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).

de IA de alto riesgo", tendencia que puede tener importantes consecuencias jurídico-laborales (en la determinación de la culpabilidad de la persona trabajadora en el ámbito disciplinario, por ejemplo).

Distinta será la configuración de estas obligaciones cuando se trate de decisiones ejecutadas directamente por la máquina, permitidas como ya hemos señalado por la Directiva cuando tales decisiones no puedan ser consideradas como perjudiciales a las personas trabajadoras – art. 10.5 -. En estos casos, y sin perjuicio de las obligaciones de transparencia y supervisión, el peso mayor de la adecuación a la legalidad de tales decisiones se va a desplazar a un momento posterior, que se traduce en los derechos de explicación, de revisión y de rectificación/indemnización (art. 11 de la Directiva). En este sentido, podemos decir que el control de legalidad de la IA que se configura en la Directiva contiene las obligaciones esenciales de control, aunque sea a posteriori, que han de extenderse a todos los sistemas de IA aplicados a las relaciones laborales, extensión por cierto propuesta ya por el Comité Económico y Social Europeo (Dictamen sobre "IA positiva para los trabajadores", de 22 de enero de 2025).

Pero la cuestión que a continuación debemos preguntarnos es si a la máquina-sistema de IA se le puede exigir más, y que su comportamiento este acorde no ya al estricto cumplimiento de la ley, sino a los valores superiores que inspira la normativa europea y contempla nuestra Constitución – justicia, igualdad, libertad, pluralismo, dignidad humana (arts. 1.1 y 10.1 CE) - y que promueven el cumplimento ético jurídico de nuestras normas. Recuérdese que, junto a recomendaciones internacionales que acentúan la necesidad de asegurar el comportamiento ético de la IA (Unesco 2021, OCDE 2024), alguna norma de nuestra Ordenamiento va más allá de exigir la estricta legalidad del comportamiento de la IA, como es el caso del art. 23.3 de la Ley 15/2022 (integral para la igualdad de trato y la no discriminación), cuando indica que "las administraciones públicas y las empresas promoverán el uso de una Inteligencia Artificial ética, confiable y respetuosa con los derechos fundamentales".

No podemos ahora introducirnos en el significado y trascedencia que tienen a nivel comunitario e interno esos valores superiores y la dignidad humana, que se alzan en el punto de referencia esencial en la doctrina judicial europea y en nuestra doctrina constitucional no sólo para la interpretación de los derechos fundamentales, sino también respecto a la base y la teleología axiológica de tales derechos. En los ejemplos citados al inicio de este brief de mal funcionamiento de sistemas avanzados de IA, en los que éstos han mentido o chantajeado, no sólo aparece una cuestión de estricta legalidad de respeto a los derechos fundamentales, sino también de incompatibilidad con una mínima ética jurídica como se traduce en esos valores superiores.

Si bien hemos señalado que, sobre todo a la luz de los arts. 9, 10 y 11 de la Directiva de plataformas, el Ordenamiento laboral puede perfilar, aunque con limitaciones, los

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).

instrumentos normativos de control de legalidad de la IA - transparencia, supervisión, explicabilidad, revisión y rectificación. - ello no solventa por sí mismo el gran desafío de hacer que esta máquina actúe según esos valores superiores, dado que en la actualidad los medios técnicos y legales que disponemos para ese "trasvase" de valores ético-jurídicos son muy limitados

Para comprender esa dificultad, hay que tener en cuenta que, en la configuración y aplicación humana de la ética, también de la axiología jurídica, interviene una "conciencia" humana fruto no sólo de un acto de racionalidad, sino también de sentidos – incluyendo sentimientos - y de experiencia de siglos que caracterizan la personalidad del ser humano y que están en la base axiológica de nuestro Ordenamiento jurídico. Por más que podamos introducir datos sobre comportamientos ético-jurídicos en la "maquina", lo cierto es que a ésta le falta esa base ya que, al menos en la actualidad, la IA no es un ser "sintiente".

No obstante los obstáculos actuales al control axiológico de los sistemas de IA, ello no quiere decir que no existan ya instrumentos para ir paliando tales limitaciones. En este sentido, y nos permitimos remitirnos al Brief de la AEDTSS que hicimos sobre el tema, la potenciación de ese control axiológico constitucional ha de tener, al menos hasta que encontremos medios más eficaces al respecto, una de sus máximos instrumentos legales en la "alfabetización" no ya sólo de las personas trabajadoras respecto al funcionamiento y riesgos de los sistema de IA, como nos indica el art. 4 RIA, sino también en la "alfabetización" de la máquina en los valores superiores por parte de todos los operadores , incluyendo las personas trabajadores interactuantes con los sistemas.

Esta labor de alfabetización axiológica de la máquina, y de ahí el gran desafío para el ordenamiento de la UE y para el nuestro propio, es urgente realizarla en los próximos años, dado la muy rápida evolución de los sistemas de IA. Desde la perspectiva jurídica, y la experiencia actual de máquinas con comportamientos ilegales y sin ética lo demuestra, el "altísimo" riesgo, mayor aún e los previstos en el RIA, no es tanto que con esos "sistemas de IA de uso general" que define este Reglamento se alcance el temido umbral de la "singularity" – máquinas con inteligencia superior a la humana -, sino que cuando ello inevitablemente ocurra, lo sea con sistemas que no hayan incorporado un funcionamiento legal y ético propio de los seres humanos, y desarrollen autónomamente sus propias reglas y axiología que puedan estar en contradicción con las de nuestro Ordenamiento europeo e interno.

CITA: DEL REY GUANTER SALVADOR., "El difícil control jurídico de los sistemas avanzados de inteligencia artificial: un gran desafío para la Unión Europea y para las relaciones laborales", Briefs AEDTSS, 68, Asociación Española de Derecho del Trabajo y de la Seguridad Social, 2025.

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos) si necesita reproducir algún fragmento de esta obra (www.conlicencia.com ; 91 702 19 70 / 93 272 04 47).