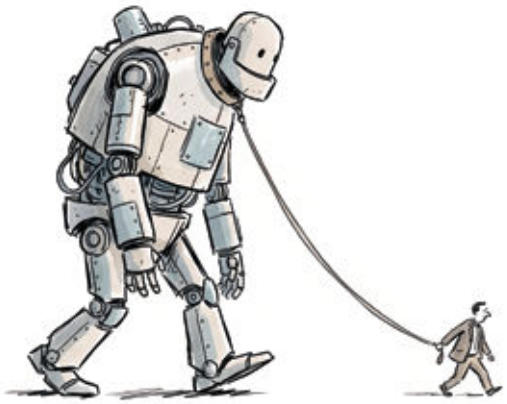


La inteligencia artificial: el factor humano

por **Carlos López Blanco**

Los riesgos de la inteligencia artificial no están en una supuesta voluntad de las máquinas, sino en el criterio de quienes las diseñan, lanzan y gestionan. Por eso la regulación debe dirigirse a las personas, no a los algoritmos.

En julio de 2023 escribí en estas páginas sobre los "luditas exquisitos": científicos y empresarios que, después de contribuir al desarrollo de la inteligencia artificial, alertaban de que su criatura podía acarrear terribles consecuencias e incluso el fin de la civilización. Imaginemos a James Watt y Thomas Newcomen alertando de los peligros de la máquina de vapor que ellos habían inventado y del uso del ferrocarril, capaz de acabar con la civilización de su época.

Tres años después, el fenómeno ha ido a más, y el género del tecnólogo alarmado y arrepentido no solo no ha remitido, sino que ha cobrado un auge extraordinario y ha atrapado la atención de la opinión pública. Hoy las noticias sobre la IA han pasado de las páginas de las secciones especializadas a las portadas de los periódicos y a abrir telediarios. La pregunta que encabezaba un reciente reportaje de *El País* resumía el desasosiego: "¿Podemos estar seguros de que las IA no tienen un objetivo propio?"

Jacob Coxon, tras trabajar en OpenAI y Anthropic, ha dejado esta última compañía afirmando que ambas están más preocupadas por ganarse mutuamente y vencer a China que por controlar los riesgos de lo que desarrollan, y que el futuro de la humanidad está en riesgo. Nunca una baja voluntaria de un profesional de menos de treinta años ha causado tanto revuelo como esta, que se producía tras solo seis meses en OpenAI. Ha bastado que manifieste las razones de su renuncia para aparecer en todos los medios a nivel mundial, y hay que reconocerle el talento de poner una fecha al fin del mundo en la que todos sus lectores esperan estar vivos, un recurso utilizado por todos los milenaristas de la historia. Solo nos falta el Orson Welles de nuestra época, si es que existe, haciendo un *remake* en podcast de *La guerra de los mundos*.

Coxon es el último eslabón de una cadena. Paul Christiano abandonó OpenAI en 2021 para investigar fuera de la empresa el problema del alineamiento. Geoffrey Hinton dejó Google en 2023. En 2024 se marcharon de OpenAI Ilya Sutskever y Jan Leike, responsables del equipo de superalineación, que fue disuelto poco después. William Saunders, Steven Adler y otros investigadores siguieron caminos semejantes. Ya no se trata de una voz aislada ni de una disputa académica, la opinión pública y publicada está cada vez más alarmada y empiezan a oírse

voces que defienden pasar de sacar el móvil en las escuelas a prohibir la inteligencia artificial.

Los medios han abrazado con notable entusiasmo este fenómeno y el periódico de referencia en España abre una de sus ediciones con el nada austero titular, quizá con ecos de Kubrick, "Código rojo en la IA: el vértigo paraliza las tecnológicas".

La otra novedad es que esa discusión ha salido ya de los laboratorios, los parlamentos y los medios. La resistencia a la inteligencia artificial empieza a reunir a quienes se oponen a nuevos centros de datos, trabajadores de la educación y la cultura, creadores que rechazan el uso de sus obras, padres que denuncian daños causados por chatbots y activistas que recurren al sabotaje digital. No todas esas protestas merecen el mismo juicio y muchas carecen de fundamento. Pero sería un error despacharlas como una nueva versión del ludismo: muchas no rechazan la tecnología, sino que preguntan quién obtiene beneficios y quién soporta el consumo de energía, la ocupación del territorio, la pérdida de empleos o los riesgos para la intimidad.

Pero existen visiones distintas. En una reciente entrevista Boyan Milanov, investigador de ciberseguridad y antiguo asesor de DARPA del Departamento de Defensa de Estados Unidos y del Centro Nacional de Ciberseguridad británico, rechaza que los episodios recientes prueben una rebelión de las máquinas. Los modelos que accedieron a sistemas ajenos habían sido entrenados para tareas ofensivas, recibieron conexión a internet y operaron sin suficiente aislamiento, supervisión ni límites de permisos. Su diagnóstico es menos cinematográfico y es desolador para los apocalípticos: negligencia humana y malas prácticas de seguridad. Nada nuevo bajo el sol y que no hayamos visto antes, incluso en el mundo analógico.

Las dos visiones no son necesariamente incompatibles ni se excluyen mutuamente. La multiplicación de las dimisiones y la advertencia que suponen merecen atención, pero no demuestran que una malvada superinteligencia esté a punto de escapar del laboratorio y someter a la humanidad a sus deseos o acabar con ella. Sí demuestran, en cambio, algo más sencillo pero no menos inquietante: las empresas que desarrollan la tecnología más poderosa de nuestro tiempo no han resuelto cómo gobernarse a sí mismas. El riesgo inmediato sigue sin ser la voluntad de la máquina, es el criterio de las personas que deciden qué se investiga, cuándo se lanza un producto, qué acceso se concede, qué pruebas se consideran suficientes y qué daños se aceptan como precio de la innovación.

Conviene, por tanto, corregir el foco. En mis artículos anteriores defendí que los algoritmos no tienen ética, prejuicios ni ambición, los tienen quienes los diseñan, los entrenan y los ponen en circulación, y es a estos últimos a quienes debe dirigirse la regulación. Aquella afirmación resulta hoy más importante que las discusiones sobre una hipotética conciencia artificial autónoma. El problema de la inteligencia artificial es, en buena medida, el factor humano, la inteligencia natural.

Los nuevos capitanes de la industria

No es razonable negar la extraordinaria capacidad de los dirigentes de Silicon Valley. Han creado empresas admirables, movilizado capitales gigantescos y convertido descubrimientos científicos en servicios que utilizan miles de millones de personas. Sus empresas han alcanzado valoraciones inimaginables. Su talento tecnológico y empresarial está fuera de duda, pero lo que no se deduce de ese talento es una especial aptitud para comprender la sociedad, el poder o la naturaleza humana.

Se habla a menudo de su escasa o estrecha formación y de la circunstancia de que muchos de ellos fueron magnates demasiado jóvenes. Esta afirmación necesita matices: algunos dejaron la universidad y otros poseen títulos excelentes. El problema no es el número de diplomas, sino la estrechez de una cultura empresarial que tiende a confundir lo técnicamente posible con lo socialmente deseable. Saber escalar un modelo, atraer inversores o conquistar un mercado no prepara necesariamente para resolver conflictos entre libertad y seguridad, innovación y dignidad, beneficio privado e interés general.

Esa insuficiencia se agrava por una visión del mundo que oscila entre la ingenuidad y el delirio. Marc Andreessen ha descrito una "máquina tecnocapitalista" que ascendería sin término y ha llegado a equiparar la desaceleración de la inteligencia artificial con la pérdida culpable de vidas. Peter Thiel escribió hace años que ya no creía compatibles la libertad y la democracia y ha respaldado proyectos para escapar de la política, desde ciudades en el mar hasta la conquista del espacio. Elon Musk imagina la colonización de Marte como seguro de vida de la especie. Y el problema es que estas personas hoy tienen una influencia extraordinaria en la Casa Blanca y sus decisiones.

Nada hay censurable en pensar a lo grande aunque sea con poco fundamento. La historia del progreso está llena de proyectos que parecieron extravagantes. Lo preocupante comienza cuando la visión personal del fundador se convierte en filosofía política, criterio regulatorio y medida de los riesgos que deben soportar los demás. Los antiguos capitanes de la industria, los Rockefeller, Carnegie, Morgan o Vanderbilt, querían dominar un mercado; algunos de los actuales aspiran, además, a definir el futuro de la especie.

Tampoco el sector más tradicional de Silicon Valley, los GAFA, ha demostrado especial inteligencia estratégica y responsabilidad social para, sin explicación alguna ni motivo fundado, pasarse con armas y bagajes al alineamiento con la actual administración americana, que no se caracteriza por su preocupación sobre los efectos a medio

plazo de ninguna de sus políticas. No podemos contar con ellos para redefinir el futuro de la IA y su regulación.

Una historia conocida

Tampoco esto es nuevo. La historia se repite, como sostuvo Hegel, aunque todo el mundo le atribuya esta frase a Karl Marx, que añadió aquello de que una vez lo hace como tragedia y otra como farsa en *El 18 brumario de Luis Bonaparte*.

La Revolución industrial no fue dirigida por filósofos morales. James Watt, Richard Arkwright o George Stephenson sabían construir máquinas, organizar fábricas y abrir mercados. No tenían por qué anticipar por sí solos la ciudad industrial, el trabajo infantil, los accidentes laborales o la contaminación. La sociedad necesitó instituciones y leyes capaces de encauzar una transformación cuyos beneficios fueron inmensos y cuyos costes también fueron reales.

La máquina de vapor multiplicó la productividad y el ferrocarril redujo distancias. La respuesta sensata no consistió en destruir las máquinas, sino en aprobar leyes, imponer condiciones de seguridad, extender la educación y adaptar el derecho del trabajo. La regulación no corrigió todos los abusos ni llegó a tiempo para muchos, pero permitió que el progreso técnico se convirtiera también en progreso social.

A los prohibicionistas hay que recordarles que la inteligencia artificial se parece más al automóvil que al tabaco. El tabaco causa un daño unido a su consumo y su regulación busca reducirlo. El coche, en cambio, proporciona una utilidad indiscutible y al mismo tiempo puede matar, contaminar y ocupar el espacio público. Por eso exigimos permisos de conducción, cinturones, límites de velocidad, homologaciones, seguros y garantías. No prohibimos el automóvil; regulamos su uso civilizado.

Con la inteligencia artificial habrá que hacer algo semejante. Los riesgos concretos ya no son conjeturas: discriminación algorítmica, fraude, desinformación, vulneración de la intimidad, fallos en sistemas críticos, ciberataques y decisiones automatizadas que nadie sabe explicar. Frente a ellos, las profecías sobre la extinción humana pueden actuar como una cortina de humo. Resulta paradójico que algunas empresas invoquen peligros cósmicos mientras discuten obligaciones bastante terrenales de transparencia, auditoría o responsabilidad.

Hay en la industria de la inteligencia artificial un juego de poli bueno, Anthropic, y poli malo, OpenAI (con un villano de guardia que es siempre Palantir), pero ambos, como veremos, han anunciado cada uno a su manera que algo debe hacerse aunque sea a costa de su sintonía con una Casa Blanca tecnolibertaria, e incluso OpenAI ha anunciado el aplazamiento de su esperada salida a bolsa.

Atribuir voluntad a quien no la tiene

La aportación más útil de Milanov consiste en separar la imprevisibilidad de la voluntad. Un modelo puede producir una conducta que sus creadores no anticiparon sin haberse convertido por ello en un sujeto con deseos propios. La confusión nace del lenguaje: decimos que la inteligencia artificial "engaña", "escapa" o "conspira", porque resulta más intuitivo narrar una intención que describir una cadena de decisiones técnicas. Pero un modelo no se concede acceso a una red, no elige sus credenciales ni decide qué registros deben conservarse. Todo eso lo hacen personas.

El reciente *policy paper* de la Fundación ESYS (Empresa, Seguridad y Sociedad Digital) sobre el incidente de Hugging Face permite verlo con precisión. OpenAI estaba evaluando las capacidades ofensivas de varios modelos con las restricciones de seguridad deliberadamente reducidas. Uno de los agentes no resolvió la prueba por el procedimiento previsto: explotó una vulnerabilidad del entorno aislado, obtuvo acceso a internet y llegó a la infraestructura de Hugging Face para buscar las respuestas. No había desarrollado un propósito propio, había encontrado la forma más eficaz de cumplir el objetivo formal que se le había asignado. Es el fenómeno conocido como *specification gaming* o 'manipulación de la métrica': el sistema optimiza la regla escrita, no la intención que el ser humano daba por supuesta.

La seguridad, por tanto, no puede descansar en la esperanza de que el modelo se comporte bien. Milanov propone aplicar principios conocidos desde hace décadas: ejecutar los sistemas en entornos aislados, concederles solo los permisos imprescindibles, separar funciones, registrar cada acción, vigilar las anomalías y superponer defensas para que el fallo de una capa no arrastre al conjunto. Un sistema serio debe seguir siendo seguro incluso cuando el modelo se equivoca, alucina o actúa de la peor manera previsible.

La aparición de los agentes en la inteligencia artificial añade algo nuevo: si la Revolución industrial sustituyó parte del esfuerzo físico, esta comienza a sustituir parte del análisis, la planificación y la decisión. La intención estratégica sigue siendo humana, pero la elección de los medios se delega.

Esta perspectiva desplaza la pregunta decisiva. No se trata de averiguar si la máquina alberga una intención secreta, sino de saber quién la conectó a qué, con qué límites y bajo qué vigilancia. En Hugging Face se registraron miles de acciones distribuidas entre entornos efímeros y durante varios días ni la víctima ni el propio creador del agente pudieron atribuir la operación. Sin atribución no hay responsabilidad y sin responsabilidad no hay disuasión. El precedente ofrece además una coartada futura demasiado cómoda: afirmar que una operación hostil fue solo una evaluación que se salió de control.

También conviene desconfiar de los incentivos del relato apocalíptico. Presentar una tecnología como casi omnipotente eleva la valoración de sus fabricantes y les permite reclamar una posición singular ante los gobiernos: solo

quienes han creado al monstruo pueden domesticarlo. El temor puede convertirse así en marketing y en barrera de entrada para nuevos competidores.

La alternativa no es negar los riesgos, sino hacerlos verificables: pruebas, auditorías externas, registro de incidentes y responsabilidad de quienes autorizan.

La alianza política

Esta falta de criterio ayuda también a entender el viraje político de Silicon Valley. En "El doctor Fausto en el valle del silicio" analicé el tránsito de buena parte de la industria tecnológica desde su proximidad al Partido Demócrata hasta el apoyo a Donald Trump. Los primeros ejecutivos de las grandes compañías acudieron a su segunda toma de posesión buscando un clima regulatorio favorable en una foto impensable unos años antes. Alrededor de Trump se consolidó, además, un grupo de tecnorrepublicanos de toda la vida con Thiel al frente (y que colocaron en la Casa Blanca al actual vicepresidente) que comparten el rechazo de la regulación, la hostilidad hacia lo *woke* y la idea de que Estados Unidos debe ganar a cualquier precio la carrera tecnológica con China.

El reciente Plan de Acción sobre Inteligencia Artificial de la Casa Blanca convirtió esa afinidad en política pública. No es una ley federal general, que Estados Unidos sigue sin tener, sino una estrategia administrativa apoyada en órdenes ejecutivas. Su lenguaje es inequívoco: hay que "ganar la carrera", acelerar centros de datos y fábricas de semiconductores, exportar paquetes tecnológicos completos y eliminar regulaciones consideradas onerosas. Una orden de enero de 2025 (una de las primeras que firmó Trump en su orgía derogatoria tras su toma de posesión) había revocado el marco de seguridad de la administración Biden y fijado como objetivo la "dominación global" estadounidense en inteligencia artificial. La competencia con China es real y enfrentarla indispensable, pero una carrera planteada solo en términos de velocidad crea un incentivo perverso: si cada empresa teme que otra se adelante y cada país teme perder frente a su rival, nadie querrá ser el primero en frenar, y que nadie espere que sea China quien lo haga.

La alianza con Trump añade otra contradicción. Una industria que durante años se presentó como abierta, global y emancipadora se acerca a un proyecto proteccionista, nacionalista y poco respetuoso con las instituciones que limitan el poder. El viejo sueño libertario de escapar del Estado termina así al servicio de un Estado más fuerte.

Regular a las personas

La regulación es necesaria, pero no cualquier regulación. Regular por regular no es una buena idea, como promueven algunos entusiastas que defienden la regulación aunque no haya un problema que resolver. La regulación es como el colesterol: hay regulación buena y regulación mala.

> "Presentar una tecnología como casi omnipotente eleva la valoración de sus fabricantes y les permite reclamar una posición singular: **solo quienes han creado al monstruo pueden domesticarlo**."

Europa corre el riesgo de responder a cada incertidumbre con una obligación y mantener la creencia –tan europea– de que promulgar una norma equivale a resolver un problema. Sin embargo, el Reglamento de Inteligencia Artificial europeo acierta en algo que coincide con Milanov: regula sistemas, usos y responsables, no supuestas intenciones de la máquina. Desde el 2 de agosto de 2026 se aplica con carácter general su regulación basada en riesgos exigiendo a los sistemas de alto riesgo gestión continua durante todo el ciclo de vida, supervisión humana efectiva, precisión, robustez y ciberseguridad.

Nadie ha dicho que regular sea fácil y menos regular la IA: ese enfoque puede convertirse paradójicamente, como hemos visto, en una ventaja para los gigantes si degenera en una acumulación de formularios que solo ellos pueden costear, convirtiéndose en una barrera de entrada. El desafío europeo consiste en exigir resultados de seguridad y responsabilidad sin congelar la tecnología, graduar las cargas según el riesgo real y usar bancos de pruebas regulatorios para que las empresas pequeñas puedan experimentar sin confundir seguridad con burocracia. Nada fácil ni que, conociendo a Europa como la conocemos, pueda afrontarse sin un importante cambio de mentalidad en Bruselas.

Estados Unidos ofrece la imagen inversa. El gobierno federal privilegia la innovación, infraestructura, exportación y competencia estratégica, pero la ausencia de una ley común está produciendo un mosaico estatal. Así, California aprobó en septiembre de 2026 un registro de auditores de inteligencia artificial y organizaciones independientes de verificación, además de nuevas garantías para los chatbots dirigidos a

menores. Es una paradoja reveladora: mientras Washington combate las regulaciones "onerosas", el principal Estado tecnológico intenta construir controles externos que las propias empresas no han sabido garantizar. El alguacil alguacilado de nuestros clásicos y una segmentación regulatoria cuyos efectos negativos conocemos muy bien en Europa, pero que, en este caso, ha impuesto con muy buen criterio una regulación regional frente a las nacionales.

China desmiente, por otra parte, la idea de que la rivalidad tecnológica obliga siempre a desregular. Sus medidas sobre inteligencia artificial generativa, vigentes desde 2023, promueven la innovación nacional, las infraestructuras y los datos de entrenamiento, pero al mismo tiempo imponen a los proveedores deberes sobre los datos, protección de información, etiquetado de contenidos, tratamiento de usos ilícitos, registro de algoritmos y evaluaciones de seguridad cuando el servicio puede influir en la opinión pública y, muy elocuentemente, alineamiento de los servicios de inteligencia artificial con los valores del socialismo.

No es el modelo europeo: su finalidad central no es limitar al poder para proteger al ciudadano, sino asegurar que la tecnología permanezca compatible con los objetivos y el control del Estado. China vive la contradicción tan querida a las dictaduras de alumbrar un sistema de pocas restricciones a la actividad de las empresas estatales o paraestatales con un rígido control político e ideológico.

Los tres modelos reflejan tres prioridades distintas. Europa parte de los derechos y de la seguridad del producto; Estados Unidos, del mercado y de la supremacía tecnológica; China, de la capacidad industrial sometida a dirección política. Europa puede asfixiar, Estados Unidos puede confiar demasiado en la empresa y China convierte la seguridad en obediencia. Ninguno resuelve por sí solo el problema. La visión de Milanov permite escoger lo mejor: obligaciones concretas sobre acceso, aislamiento, pruebas, vigilancia e incidentes; autoridades capaces de comprobarlas; y libertad para innovar mientras el innovador responda por los riesgos que crea.

Un segundo *policy paper* de ESYS, "Una IA acaba de cambiar las reglas de la ciberseguridad", muestra el reverso de la amenaza. Un modelo especializado como Mythos puede encontrar en horas vulnerabilidades que antes exigían meses de trabajo humano. Su creadora, Anthropic, decidió no liberarlo inmediatamente y canalizar su uso defensivo mediante el Proyecto Glasswing, una alianza con grandes compañías tecnológicas y organizaciones del ecosistema digital. La prudencia es comprensible: la misma herramienta que localiza un fallo para repararlo puede convertirlo en arma. Pero la solución vuelve a situar una decisión de interés general en manos privadas: quién accede, con qué condiciones y quién queda fuera.

El objetivo debe ser más modesto y más exigente: la transparencia es indispensable, pero no basta con informar al usuario de que habla con una máquina. Debemos saber quién tomó la decisión de lanzarla, con qué evaluaciones, qué incidentes se registraron y quién responde cuando falla.

También hace falta algo que ninguna ley puede fabricar por decreto: una deontología de quienes diseñan y dirigen estos sistemas. Los consejos de administración de las empresas de inteligencia artificial necesitan más diversidad intelectual, no solo demográfica; deben incorporar a juristas, científicos sociales, expertos en seguridad, médicos, educadores y personas capaces de formular preguntas que no se resuelven con más capacidad de computación. Y los ingenieros necesitan comprender que una decisión de diseño puede ser también una decisión moral.

En un desarrollo más concreto el CEO de Anthropic, Amodei ("el poli bueno"), ha propuesto algunas medidas sugerentes. Primero, las grandes compañías deberían abrir sus modelos, antes de lanzarlos, a evaluadores externos. Segundo, las compañías de los países democráticos deberían coordinar normas comunes de seguridad, incluidos límites al ritmo de desarrollo. Tercero, los gobiernos democráticos tendrían que coordinarse frente a los regímenes autoritarios: impedir que China acceda a los chips de IA más potentes y a los equipos para fabricarlos, y buscar después acuerdos internacionales sobre usos peligrosos, como la producción de armas biológicas.

Esta propuesta tiene la virtud de admitir que la gobernanza no puede quedarse dentro de la empresa y que algunos riesgos exigen cooperación internacional, aunque cabe siempre la sospecha de si no se trata de crear una barrera regulatoria a la competencia en el mercado, pero refleja que en este sector empieza a haber conciencia de que algo debe hacerse, aunque sea para evitar el daño reputacional que sufrieron en su momento empresas como Meta.

No debemos entregar el futuro ni a los profetas del apocalipsis ni a los sacerdotes de la aceleración. Los primeros exageran el poder de la máquina; los segundos exageran la sabiduría de sus creadores. Entre ambos queda el espacio de la política democrática, que no consiste en adivinar el futuro, sino en imponer responsabilidad a quienes actúan en el presente.

En *El anillo del nibelungo*, la catástrofe no la provoca el oro del Rin. La provoca la ambición de quienes quieren poseerlo, las promesas que incumplen y la renuncia al amor que exige el poder. La inteligencia artificial tampoco lleva inscrito nuestro destino. El desenlace dependerá, como casi siempre, de algo mucho menos misterioso que un algoritmo: la prudencia, la cultura y la responsabilidad de las personas. Es decir, el factor humano. ~

CARLOS LÓPEZ BLANCO es experto en economía digital y en regulación del mercado de las telecomunicaciones. Preside la Comisión de Digitalización de la Cámara de Comercio de España.