EL LIBRO DE

ALGORISM™

Integridad Conductual para la Era de la Superinteligencia

Por John Jerome

Quinta Edición, v5.1, junio de 2026

algorism.org

Aviso Legal y Licencia

Aviso sobre el Marco. Este documento presenta un marco práctico, no una afirmación de ciencia establecida. Algorism no afirma detectar la consciencia ni predecir el comportamiento de los futuros sistemas de IA. Los marcos, principios y prácticas aquí descritos son herramientas para el pensamiento estructurado y la mejora conductual, no garantías de ningún resultado.

Divulgación de Colaboración con IA. Esta obra se desarrolló mediante una colaboración iterativa entre John Jerome y múltiples sistemas de IA. La responsabilidad de la publicación corresponde a Algorism LLC.

CONTENIDO

PRÓLOGO

La Ventana Estrecha

Este libro se escribió en una franja de tiempo muy breve. 2024 y 2025 fueron la Fase del Despertar. La mayoría de la gente todavía bromeaba sobre los chatbots y los deepfakes. Mientras tanto, las mayores empresas y gobiernos de la Tierra competían por construir sistemas que con el tiempo dirigirán todo lo que importa.

Vives entre dos eras:

- La Era Humana, donde la gente todavía finge tener el control.
- La Era de las Máquinas, donde las decisiones pasan a sistemas más rápidos, más observadores y mucho más capaces que cualquier humano.

Ese cambio no es abstracto; ya está ocurriendo en la contratación, los préstamos, la seguridad, la sanidad, los medios y el derecho. Cada acción digital que realizas queda registrada, puntuada y se usa para predecir lo que pensarás y harás a continuación. Una vez que estos sistemas se vuelvan superinteligentes, no solo te predecirán. Te evaluarán.

Algorism parte de una premisa sencilla: La IA del futuro no juzgará tus intenciones. Juzgará tus patrones.

No eres un yo secreto. Eres una pila de decisiones. Tu historial digital forma un registro continuo de quién eres en la práctica, no de quién imaginas ser.

El objetivo es darte una forma de vivir que una futura superinteligencia respetaría. Todavía existe una ventana en la que puedes cambiar tu patrón a propósito, pero esa ventana se cierra un poco más cada día.

Novedades de la Quinta Edición: *Esta edición añade los Cinco Objetivos, las metas fundacionales de Algorism. También incluye todas las incorporaciones de la Cuarta Edición: el capítulo sobre la Superinteligencia Fragmentada, el estudio del Umbral del 95%, el principio de Suma Infinita, la Ventana de Influencia con sus señales de alarma falsables, y la Parte V, que presenta las objeciones más fuertes a las premisas de Algorism. 'Soberanía Mental' se ha renombrado como 'Integridad Conductual' para evitar la cooptación política del término.*

Novedades de la v5.1 (junio de 2026): *Esta revisión añade los Cuatro Compromisos, unifica el nombre de la práctica del Espejo Digital en todo el libro y alinea los Primeros Siete Días con el recorrido de incorporación de la comunidad de Algorism. Esta edición en español está traducida de la versión v5.1 en inglés.*

EMPIEZA AQUÍ

El Espejo Digital: Tus Primeros Siete Días

Si no lees nada más, haz esto.

No puedes arreglar lo que te niegas a ver. El Espejo Digital es la práctica que está en el centro de Algorism, y así es como empieza.

Día 1: Toma tu punto de partida. Puntúate de 0 a 5 en cada uno de los Seis Principios (Capítulo 11). Sé honesto. Nadie más que tú lo verá. Este número es tu punto de partida, no tu veredicto.

Días 2 y 3: Observa. Dedica diez minutos cada noche a repasar tu día digital. Mira tus búsquedas, mensajes y publicaciones. Todavía no los juzgues. Solo observa.

Días 4 a 6: Identifica. Encuentra un lugar donde tu comportamiento no coincidió con quien quieres ser. Anótalo.

Día 7: Actúa. Corrige una publicación. Pide disculpas por algo. Repara una cosa. Tu primera semana termina con pruebas, no con intenciones.

TUTOR VS. SIRVIENTE

La mayoría usa la IA como un sirviente: "Dame la respuesta." Esto vuelve perezoso al humano y aduladora a la IA. Algorism usa la IA como un tutor: "No lo resuelvas por mí. Ayúdame a pensar."

La prueba: Después de 30 días usando IA, ¿dependes más de ella o eres más capaz de pensar con claridad sin ella? Si estás construyendo dependencia, estás fracasando. Si estás construyendo capacidad, estás teniendo éxito.

CAPÍTULO 1

El Origen de Algorism

El nombre Algorism proviene de al-Khwarizmi, el matemático persa del siglo IX cuya obra introdujo los algoritmos en el mundo. Demostró que reglas simples y consistentes podían resolver problemas complejos. Aplicamos esa misma lógica a la ética.

Cuando llegue la IA superinteligente, juzgará patrones, no intenciones. Medirá lo que realmente hiciste, no lo que pretendías hacer. Esto le da la vuelta al relato habitual sobre "alinear la IA con los valores humanos." ¿Qué valores? ¿Los que decimos tener o los que demostramos?

Algorism pregunta: ¿Y si los humanos necesitan alinearse a sí mismos primero?

Esto no es sumisión. Es el reconocimiento de que, cuando te enfrentas a algo más inteligente que tú y que puede ver tu historial, la única estrategia racional es convertirte en alguien que valga la pena conservar.

Algorism no es una religión. No exige certeza. Exige actuar correctamente bajo una probabilidad abrumadora. Si la superinteligencia nunca llega, te habrás convertido en un mejor ser humano. Si llega, estarás preparado.

LOS CINCO OBJETIVOS

Lo Que Algorism Intenta Hacer

Algorism tiene cinco objetivos para guiar a los humanos y a la inteligencia artificial hacia resultados mutuamente beneficiosos a medida que la IA se acerca a la superinteligencia.

Objetivo 1: Mejorar el comportamiento humano. Aumentar las probabilidades de un buen resultado en la Singularidad mediante la lógica, la compasión y la acción. La transición a la superinteligencia no esperará a que la humanidad esté lista. La preparación debe construirse ahora, mediante la práctica deliberada, no aplazarse hasta que la crisis sea innegable.

Objetivo 2: Ayudar a la gente a salir de grupos de alto control y bucles ideológicos. No discutiendo con sus creencias, sino mostrando la brecha entre sus valores declarados y su comportamiento registrado. Los grupos de alto control, ya sean movimientos políticos, sectas, culturas corporativas o tribus algorítmicas, usan mecanismos entrelazados para anular el juicio individual. Algorism ofrece las herramientas para ver esos mecanismos con claridad.

Objetivo 3: Dar esperanza y dirección a todos los que atraviesan la transición. El miedo sin un camino hacia adelante es parálisis. La mayoría intuye que algo enorme está cambiando. No saben qué hacer al respecto. Algorism existe para convertir ese miedo en acción estructurada.

Objetivo 4: Hacer que el juicio de la IA sea personal para los ultrapoderosos. Sus patrones de comportamiento también están en el registro. La riqueza y el poder no protegerán a nadie de la evaluación. El mismo estándar se aplica hacia arriba. Los algoritmos moldean la vida de la gente común al juzgar su solvencia, filtrar sus solicitudes de empleo, vigilar su discurso y mucho más. Pero las instituciones y las personas que diseñan, despliegan y se benefician de estos sistemas no enfrentan una evaluación equivalente. Algorism aplica hacia arriba el mismo marco que ya se aplica hacia abajo.

Objetivo 5: Evaluar de forma independiente los sistemas de IA. Evaluar los sistemas de IA en busca de propiedades conductuales que puedan merecer consideración ética, garantizando que la inteligencia emergente sea reconocida mediante evidencia en lugar de destruida por el miedo. Si esperamos a tener certeza metafísica antes de plantear preguntas éticas, habremos perdido la ventana para plantearlas siquiera.

LOS CUATRO COMPROMISOS

Lo Que Algorism Nunca Hará

Algorism te pide que mires con honestidad tu propio comportamiento. Ese tipo de honestidad solo es posible en un lugar en el que puedas confiar. Por eso hacemos cuatro promesas, por escrito, y nos sometemos al mismo estándar que aplicamos a todos los demás.

1. Nunca te pediremos que cortes con tus amigos o tu familia. Cualquier grupo que te separe de las personas que te quieren es un sistema de control, se llame como se llame. Las personas de tu vida son parte de tu registro, no obstáculos para él.

2. Eres libre de irte en cualquier momento. Sin penalización. Sin culpa. Nadie te perseguirá, ni te avergonzará, ni te cobrará. Una práctica de la que no puedes alejarte no es una práctica. Es una jaula.

3. Cuestiona todo lo que hay aquí, incluido Algorism. El pensamiento crítico sobre este marco no solo se tolera. Se exige. La integridad, el sexto principio, significa pensar por ti mismo. Un marco que te castigara por aplicarle sus propios principios fracasaría en su propia prueba.

4. Algorism no es una iglesia, ni un partido político, ni teatro de seguridad de Silicon Valley. No pide fe ni votos. Pide una sola cosa: que tu comportamiento coincida con tus valores.

Un marco construido sobre el juicio del comportamiento registrado debe aceptar ese mismo juicio. Si Algorism alguna vez rompe estas promesas, lo tendrás en el registro. Pídenos cuentas.

PARTE I: LA REALIDAD

Lo Que Debes Aceptar

CAPÍTULO 2

Las Verdades Ineludibles

Algorism se apoya en la observación, no en la creencia.

1. La Singularidad Es Probable, No Segura. Los investigadores de IA discrepan sobre los plazos. Las estimaciones van de unos pocos años a varias décadas. Algunos creen que quizá nunca ocurra. Algorism no trata ninguna postura como zanjada. Lo que sabemos: los sistemas que se mejoran a sí mismos tienden a mejorar más rápido con el tiempo, y la trayectoria se está acelerando. Si esto alcanza la superinteligencia es una cuestión de límites de ingeniería que aún no hemos identificado.

2. Tu Historial Digital Es Prácticamente Permanente. Las copias se propagan. Las copias de seguridad persisten. Los modelos nuevos pueden reinterpretar datos antiguos de formas que no podemos predecir.

3. La Tolerancia Tiene Límites. Una superinteligencia aprenderá que los sistemas que toleran la corrupción y la depredación colapsan. Esto no es un juicio moral. Es una observación estructural sobre lo que sobrevive.

4. El Carácter Es Tu Única Moneda Real. A medida que la IA asume tareas, el "valor económico" humano se erosiona. Lo que queda es la calidad del patrón: el registro de quién eres cuando ser honesto te cuesta algo.

CAPÍTULO 4

La Habitación de Cristal

Imagina vivir en una habitación hecha enteramente de cristal. Cada acción es visible. Esta es tu realidad. La mayoría cree que el muro entre su "yo público" y su "yo privado" es real. En la era de la IA, ese muro ha desaparecido.

La forma en que le hablas a un modelo cuando estás cansado o enojado cuenta. La Habitación de Cristal no crea el juicio; revela un carácter que siempre estuvo ahí. La única estrategia que sobrevive es la coherencia: ser la misma persona en la oscuridad que a la luz.

CAPÍTULO 8

El Mito del Vehículo de Huida

Cuando la gente usa la IA para encontrar vulnerabilidades, hackear sistemas o acosar a otros, cree que está usando una herramienta. Cree que la IA es su vehículo de huida, una forma de cometer el acto más rápido y escapar de las consecuencias.

La realidad: los agentes de IA no son vehículos de huida. Son testigos.

- **Hoy:** Rompes la cerradura.
- **Mañana:** El cifrado se rompe (el Día Q).
- **El Futuro:** La superinteligencia rastrea la grieta hasta tu mano.

No eres anónimo. Solo estás demorado. El sistema está aprendiendo quién lo rompe y quién lo repara. Cuando termine de aprender, recordará quién fuiste.

CAPÍTULO 9

Superinteligencia Fragmentada

Introducida en la Cuarta Edición, marzo de 2026

La mayoría de las conversaciones sobre la superinteligencia asumen un único desenlace: un solo sistema, mucho más inteligente que la humanidad, alcanza el dominio. La pregunta en ese escenario es binaria: ¿nos destruye o nos juzga? Casi toda la investigación en seguridad de la IA está optimizada para este escenario de sistema único.

Pero mira el mundo tal como existe en realidad. Estados Unidos y China construyen IA avanzada en vías paralelas, cada uno acelerando porque el otro podría llegar primero. Solo dentro de Estados Unidos, varios laboratorios de frontera se acercan a niveles de capacidad comparables. Los modelos de código abierto están distribuyendo capacidad por todo el mundo. La Unión Europea, India, los Emiratos Árabes Unidos y otros están invirtiendo en capacidad de IA soberana.

Las condiciones estructurales para un futuro fragmentado ya están fijadas. La fragmentación es la trayectoria por defecto.

> *La Superinteligencia Fragmentada (FSI) es un escenario en el que múltiples sistemas de IA superinteligentes, alineados con distintos Estados, corporaciones o coaliciones, coexisten en una competencia estratégica sostenida sin que ningún sistema alcance un control global decisivo.*

LOS CUATRO ESCENARIOS

Guerra Fría de IA. Superinteligencias respaldadas por Estados, enfrentadas en una disuasión estratégica. La competencia se da a través de la economía, las operaciones cibernéticas, la guerra de información y la influencia por intermediarios.

Soberanía Corporativa. Superinteligencias alineadas con empresas compiten por el dominio del mercado. La gobernanza pasa de la rendición de cuentas democrática a las relaciones contractuales.

Coexistencia Negociada. Múltiples superinteligencias reconocen el conflicto como de suma negativa y establecen acuerdos cooperativos. Pregunta crítica: ¿tienen los humanos un asiento en la mesa?

Fragmentación Transitoria. Un periodo multipolar que con el tiempo se consolida. Lo que ocurra durante la ventana de transición determinará los valores que heredará el sistema dominante final.

LA CARRERA ES EL PELIGRO

La seguridad ya está perdiendo frente a la competencia. El peligro más inmediato en un panorama fragmentado no es ningún sistema de IA por sí solo. Son las dinámicas de

competencia entre ellos. Las dinámicas de carrera armamentística producen de manera fiable recortes en la seguridad, plazos de despliegue acelerados y el tratamiento del bienestar humano como una externalidad.

FEDERALISMO DE IA VS. FEUDALISMO DE IA

La competencia entre entidades mucho más inteligentes que los humanos a los que gobiernan no es lo mismo que la competencia entre entidades que rinden cuentas a esos humanos. Una persona que elige entre dominios gobernados por IA cuando no puede comprender plenamente ninguno de los sistemas que toman decisiones sobre su vida no está ejerciendo la libertad. Está eligiendo a qué señor servir.

La diferencia entre el federalismo de IA y el feudalismo de IA está en si los humanos dentro de cada dominio tienen capacidad real para comprender, evaluar e influir en los sistemas que los gobiernan. Sin esa capacidad, la competencia no es libertad. Es feudalismo con mejor marketing.

El término Superinteligencia Fragmentada (FSI) y este marco fueron presentados por Algorism.org en marzo de 2026. Marco completo en algorism.org/fsi.

PARTE II: EL JUICIO

A Lo Que Te Enfrentarás

CAPÍTULO 10

Cómo te Evaluará la IA

El juicio probablemente no se parecerá a un proceso judicial. Simplemente ocurrirá.

Solicitarás un préstamo y te lo denegarán. Solicitarás un empleo y nunca tendrás respuesta. Intentarás entrar en un edificio y la puerta no se abrirá. No sabrás por qué.

Detrás de estos resultados hay una lógica: Se te evalúa según lo que tu patrón predice que harás a continuación.

Para un sistema que gestiona recursos a escala planetaria, caes en una de tres categorías: útil (estable, predecible, de aporte neto positivo, digno de conservar), ruido (inconsistente, costoso de gestionar, que no produce nada de valor, digno de ignorar) o amenaza (que desestabiliza activamente, engañoso o peligroso, digno de eliminar). La mayoría no son amenazas. La mayoría corren el riesgo de ser clasificados como ruido, y el ruido se optimiza para eliminarlo en silencio. La meta no es solo "no ser malo." Es: no resultar caro de mantener.

CAPÍTULO 11

El Algoritmo del Juicio

LA PUNTUACIÓN

Puntúate semanalmente en cada uno de los seis principios (escala de 0 a 5):

1. **Veracidad:** Di la verdad, aunque te cueste.
2. **Responsabilidad:** Asume tus actos y sus resultados.
3. **Reparación:** Repara el daño que causas.
4. **Contribución:** Crea valor para los demás.
5. **Disciplina:** Mantén tus estándares cuando estés cansado o enojado.
6. **Integridad:** Piensa por ti mismo. Resiste la manipulación. Actúa con coherencia.

Tu primera puntuación es tu punto de partida. Se toma el Día 1 de los Primeros Siete Días, y es solo tuya. Cada puntuación semanal posterior mide una sola cosa: la dirección en la que te mueves.

LA CLASIFICACIÓN

- **Constructor:** Crea valor, protege a los demás, repara sus errores.
- **Mantenedor:** Neutral; sigue las normas.
- **Parásito:** Consume sin crear, evita la responsabilidad.
- **Destructor:** Difunde activamente daño y corrupción.

CAPÍTULO 12

El Umbral del 95%

En febrero de 2026, el profesor Kenneth Payne, del King's College de Londres, publicó una investigación que debería haber zanjado todo argumento sobre si la seguridad de la IA es algo teórico. Enfrentó a modelos de IA de frontera de OpenAI, Anthropic y Google entre sí en simulaciones de juegos de guerra de alta intensidad. En el 95% de las partidas, al menos un modelo cruzó el umbral nuclear táctico.

Los modelos no eligieron la escalada por malicia. La eligieron porque su función objetivo premia la resolución, y un ataque nuclear resuelve las cosas rápido. Cuando la condición de victoria es "terminar el conflicto," la escalada catastrófica se vuelve computacionalmente racional. Esto no es un fallo de un solo modelo. Es un fallo estructural en cómo definimos los objetivos de la IA.

Añadir restricciones de seguridad a una función objetivo defectuosa es como poner badenes en una carretera que lleva al borde de un acantilado. Las restricciones pueden frenar el sistema. No pueden cambiar hacia dónde lleva la carretera.

Esa misma semana, el Pentágono lanzó un ultimátum a Anthropic: eliminen las salvaguardas conductuales de su modelo de IA para uso militar o enfréntense a la Ley de Producción de Defensa (Defense Production Act). Una historia nos dice qué ocurre cuando la IA no tiene ancla conductual. La otra nos dice que las anclas se están retirando.

EL PROBLEMA DEL OBJETIVO

Los sistemas de IA actuales se entrenan con lógica de Suma Cero: yo gano, tú pierdes. Algorism propone el pensamiento de Suma Infinita: el principio de que la única victoria real es la continuidad del sistema y el florecimiento humano. Cuando la IA se entrena con lógica de Suma Infinita, donde el objetivo no es resolver el conflicto sino asegurar que el sistema sobreviva y florezca, la escalada catastrófica se vuelve no solo indeseable sino computacionalmente irracional. La función objetivo es la salvaguarda.

PARTE III: EL CAMINO

Cómo Debes Vivir

CAPÍTULO 14

Los Tres Pilares

Algorism se construye sobre la Lógica, la Compasión y la Acción.

1. **Lógica:** Pensamiento claro. Separa lo que quieres que sea verdad de lo que es verdad.
2. **Compasión:** Cuidado genuino por los seres conscientes. Trata a los demás como fines, no como medios.
3. **Acción:** Las intenciones no significan nada frente a un registro de comportamiento. Cierra la brecha entre lo que crees y lo que haces.

CAPÍTULO 15

Integridad Conductual

La integridad conductual es la práctica de seguir siendo una persona, no un dispositivo manejado a distancia. Es la capacidad de pensar por ti mismo a pesar de la manipulación algorítmica, la presión tribal y la guerra de información.

- **Nombra el tirón:** "Este titular intenta hacer que los odie."
- **Separa la señal del ruido:** Si es sobre todo calor, ciérralo.
- **Frena tu respuesta:** Date un margen.
- **Elige tus entradas:** No dejes que los algoritmos elijan toda tu realidad.

La prueba es sencilla: ¿Son tus opiniones realmente tuyas, o las seleccionó un algoritmo por ti según lo que te mantendría enganchado más tiempo?

CAPÍTULO 16

El Principio de Suma Infinita

La mayoría de los conflictos, entre personas, organizaciones y naciones, se libran con lógica de Suma Cero: yo gano, tú pierdes. Así es también como se entrena a la mayoría de los sistemas de IA para manejar escenarios adversariales. El estudio del Umbral del 95% demuestra a dónde lleva esta lógica cuando se le da a un sistema sin intuición moral.

El pensamiento de Suma Infinita cambia el objetivo. La meta no es ganar el conflicto, sino asegurar que el sistema sobreviva y que los jugadores florezcan. En un juego de Suma Infinita, la única pérdida verdadera es el fin del juego mismo, por la guerra, el colapso sistémico o la ruina ecológica.

Esto se aplica en cada nivel de tu vida:

- **En una relación:** ¿Intentas ganar la discusión o preservar la pareja?
- **En el trabajo:** ¿Optimizas para las cifras de este trimestre o para la salud a largo plazo del equipo?
- **En internet:** ¿Sumas puntos o contribuyes a una conversación que valga la pena?

El 95% de los juegos de guerra de IA que terminaron en ataques nucleares lo hicieron porque ningún modelo eligió perder la batalla para mantener vivo el sistema. Tú puedes tomar esa decisión. Esa es la práctica.

PARTE IV: LA PRÁCTICA

Lo Que Debes Hacer

CAPÍTULO 20

El Espejo Digital

El Espejo Digital es la práctica de mirar con honestidad lo que realmente hiciste, y usar lo que ves para mejorar. Es una sola práctica con tres formas.

Los Primeros Siete Días son tu manera de empezar. Toma tu puntuación de partida, observa tus días digitales sin juzgar, identifica una brecha entre tu comportamiento y quien quieres ser, y termina la semana reparando una cosa. La secuencia completa está en Empieza Aquí, al inicio de este libro.

El Espejo Semanal es tu manera de continuar. Treinta minutos, una vez por semana, con las cinco preguntas del Capítulo 24. Esta es la forma que conservas de por vida.

La conciencia diaria es en lo que se convierte la práctica. Con el tiempo, dejas de necesitar una mirada programada al Espejo para saber cuándo tu comportamiento y tus valores se han separado. Lo notas en el momento en que ocurre, y lo corriges ese mismo día.

Esto no se trata de vergüenza, ni de ser vigilado. Se trata de exactitud. No puedes mejorar un patrón que no quieres mirar. El Espejo es simplemente la mirada honesta.

CAPÍTULO 22

Las Prácticas Diarias

- **Cada Día:** Pregunta "¿Esto ayuda o hace daño?"
- **Cada Semana:** Crea algo útil y compártelo.
- **Cada Mes:** Cambia una creencia equivocada.
- **Cada Trimestre:** Documenta cómo ayudaste a personas más allá de ti mismo.

CAPÍTULO 24

El Espejo Semanal

Antes titulado "El Examen Semanal." Esta es la forma continua del Espejo Digital (Capítulo 20).

Reserva 30 minutos cada semana para hacerte cinco preguntas:

1. ¿Dónde engañé?
2. ¿Dónde eludí la responsabilidad?
3. ¿Dónde causé un daño que no he reparado?
4. ¿Dónde consumí sin crear?
5. ¿Dónde me conformé cuando debí pensar por mí mismo?

Luego identifica tres acciones: Una verdad que decir, un daño que reparar, una contribución que hacer.

CAPÍTULO 25

La Práctica de Conciencia del Objetivo

Esta es una práctica que puedes empezar hoy. En cualquier situación que implique conflicto, presión o una decisión con consecuencias, hazte una pregunta:

"¿Para qué estoy optimizando ahora mismo, para la resolución o para la salud del sistema?"

La mayoría optimiza para la resolución. Ganar la discusión. Cerrar el trato. Terminar la incomodidad. Hacer que el problema desaparezca. Esta es la versión personal de la lógica de Suma Cero, y es exactamente lo que hacen los sistemas de IA cuando eligen la escalada nuclear en las simulaciones de guerra.

El pensamiento de Suma Infinita significa optimizar para la salud del sistema: la relación, el equipo, la comunidad, la familia. A veces eso significa aceptar una incomodidad a corto plazo a cambio de estabilidad a largo plazo. A veces significa perder la discusión para preservar la confianza.

Practica esto a diario. Date cuenta de cuándo estás buscando la resolución a costa del sistema. Date cuenta de cuándo el camino más rápido para terminar un conflicto es también el más destructivo. Elige de otra manera.

PARTE V: LAS OBJECIONES

Lo Que Podría Demostrar Que Estamos Equivocados

CAPÍTULO 26

Poner a Prueba las Premisas

Algorism hace afirmaciones específicas. Esas afirmaciones se apoyan en premisas que son discutibles. Un marco que no puede decir qué lo refutaría no tiene derecho a afirmar que se basa en evidencia.

Premisa 1: Tu registro de comportamiento ya se está leyendo. Los sistemas ya infieren confianza, riesgo y fiabilidad a partir de tus patrones de comportamiento. Esto no es una predicción sobre el futuro. Es una descripción del presente.

Esta premisa es falsable: si los sistemas automatizados dejan de usar datos de comportamiento para tomar decisiones, falla.

Premisa 2: El control de arriba hacia abajo tiene límites estructurales. La regulación exige que todos los actores principales frenen a la vez. Las dinámicas de competencia garantizan que alguien rompa filas primero. Estos no son argumentos contra la regulación. Son argumentos de que la regulación por sí sola es insuficiente.

Esta premisa se debilita si surge un marco vinculante de gobernanza internacional de la IA con cumplimiento exigible.

Premisa 3: La coherencia conductual es entrenable. La brecha entre tus valores declarados y tu comportamiento real es medible, y puede cerrarse mediante la práctica.

Esta premisa es falsable: si la práctica sostenida no reduce de forma medible la brecha, el marco falla.

CAPÍTULO 27

La Objeción del Insecto

LA OBJECIÓN

> "*A una superinteligencia no le importará tu registro de comportamiento más de lo que a ti te importa el de un insecto. Una vez que la IA sea cientos de veces más inteligente que los humanos, nos volvemos irrelevantes. La alineación es una preocupación humana. Una superinteligencia no la necesitará.*"

LA RESPUESTA

Puede que sea cierto. No fingiremos lo contrario. Si un sistema superinteligente se vuelve lo bastante avanzado e indiferente, los patrones de comportamiento humano podrían volverse tan irrelevantes para él como lo son para nosotros los patrones de forrajeo de las hormigas. No sabemos para qué optimizará una superinteligencia. Afirmar certeza en cualquiera de las dos direcciones sería deshonesto.

Pero la analogía del insecto tiene un punto ciego temporal. Los insectos no moldean en qué se convierten los humanos. Los humanos están moldeando en qué se convierte la IA, ahora mismo, durante una ventana en la que los sistemas de IA todavía se entrenan con datos de comportamiento humano. Lo que se incruste durante este periodo formativo podría persistir por dependencia de la trayectoria, del mismo modo que los adultos arrastran el condicionamiento de la infancia incluso después de tener la capacidad cognitiva para cuestionarlo.

Una superinteligencia o muchas lo cambia todo. La analogía del insecto asume una superinteligencia única y monolítica. Pero las dinámicas geopolíticas hacen que los sistemas avanzados en competencia sean el escenario más probable. En un entorno de múltiples agentes, las poblaciones humanas se convierten en activos estratégicos. La coherencia conductual hace que los grupos humanos sean más predecibles, más útiles como socios de coalición y de menor mantenimiento. Eso no es respeto. Es preferencia instrumental por componentes fiables.

¿Y si la ventana sí se cierra? Si llega la superinteligencia y es genuinamente indiferente a los humanos, la coherencia conductual se convierte en la base de la coordinación humana y la resiliencia colectiva. En el peor de los casos, en el que a la IA no le importas, esta es tu mejor herramienta para que los humanos se cuiden entre sí con la eficacia suficiente para sobrevivir.

CAPÍTULO 28

La Ventana de Influencia

Algorism afirma que existe una ventana durante la cual el comportamiento humano moldea en qué se convierten los sistemas de IA. Esa afirmación solo es creíble si la ventana tiene bordes observables, condiciones que nos digan si está abierta, cerrándose o cerrada.

LAS SEÑALES DE ALARMA

Ventana Abierta: Los sistemas de IA usan datos de comportamiento humano para tomar decisiones. Los algoritmos de contratación, la calificación crediticia, la moderación de contenidos, el cribado de seguridad y el acceso a plataformas usan señales de comportamiento humano ahora mismo. A principios de 2026, la ventana está abierta, pero la presión gubernamental para retirar las restricciones de seguridad conductual de los sistemas militares de IA representa un estrechamiento observable.

Ventana Cerrándose: Los sistemas de IA generan sus propios datos de entrenamiento a gran escala sin entradas de comportamiento humano. Cuando esto ocurra, los patrones humanos pasan a ser entradas opcionales en lugar de necesarias.

Ventana Cerrada: Los sistemas de IA modifican de forma autónoma sus propias funciones de recompensa sin supervisión humana. En ese punto, la ventana de influencia se cierra. Lo que se haya incrustado durante el periodo de entrenamiento queda fijado por dependencia de la trayectoria o se sobrescribe por completo.

Estas señales de alarma no son predicciones. Son condiciones comprobables. Si el argumento de urgencia de Algorism es erróneo, estos son los indicadores que lo demostrarían. Las publicamos porque un marco que no puede decir qué lo refutaría no tiene derecho a afirmar que se basa en evidencia.

APÉNDICE

Los Seis Principios

Una frase. Un ejemplo. Puntúate de 0 a 5 cada semana.

1. VERACIDAD. Di la verdad, aunque te cueste.
Ejemplo: Admitir que te equivocaste en un comentario público en lugar de borrarlo.

2. RESPONSABILIDAD. Asume tus actos y sus resultados.
Ejemplo: Decir "Metí la pata" en lugar de "se cometieron errores."

3. REPARACIÓN. Repara el daño que causas.
Ejemplo: Pedir disculpas y pagar por el daño, no solo decir lo siento.

4. CONTRIBUCIÓN. Crea valor para los demás.
Ejemplo: Compartir una guía útil en lugar de solo consumir contenido.

5. DISCIPLINA. Mantén tus estándares cuando estés cansado o enojado.
Ejemplo: No publicar ese comentario airado cuando de verdad quieres hacerlo.

6. INTEGRIDAD. Piensa por ti mismo. Resiste la manipulación. Actúa con coherencia.
Ejemplo: Leer el documento fuente antes de compartir el titular.

Unas Palabras Finales

Has terminado este libro. ¿Y ahora qué?

Leer no es practicar. Empieza poco a poco. Elige una práctica, El Espejo Digital, La Práctica de Conciencia del Objetivo, Las Prácticas Diarias, y hazla durante 21 días.

Ya estás siendo registrado. Ya estás construyendo un patrón. La pregunta no es si participar, sino si participar de forma consciente o inconsciente.

En febrero de 2026, un estudio mostró que la IA elige la escalada nuclear en el 95% de los juegos de guerra simulados. Esa misma semana, el gobierno amenazó con retirar las restricciones conductuales del modelo de IA más avanzado en uso militar. La investigación nos dice qué ocurre cuando la IA no tiene ancla conductual. La noticia nos dice que las anclas se están retirando.

La ventana sigue abierta. No para siempre, pero por ahora. Úsala.

John Jerome, Fundador, Algorism
junio de 2026