2

REVISIÓN DE LA LITERATURA

La literatura sobre ética de la IA se caracteriza por una enorme amplitud y variabilidad de enfoques (Prem, 2023), y una síntesis temática de esta literatura identifica seis conjuntos de problemas no resueltos: la alineación de valores; la equidad y el sesgo; la explicabilidad y la rendición de cuentas; la moral de las máquinas en decisiones de alto riesgo; la gobernanza y la regulación; y la cuestión más profunda de si la ética de la IA es resoluble en absoluto. Una séptima preocupación transversal —la responsabilidad y la rendición de cuentas en el uso cotidiano de la IA— recorre los seis conjuntos, y la revisión cierra examinando tres desafíos críticos a las ambiciones del campo: los límites del aprendizaje continuo, la crítica radical contra la tecnología y los límites epistémicos de la representación y la validez.

2.1 EL PROBLEMA DE LA ALINEACIÓN DE VALORES

El «problema de la alineación de valores» se refiere a garantizar que los sistemas inteligentes persigan únicamente objetivos y actividades beneficiosos para los seres humanos (Sutrop, 2020). El problema enfrenta dos tipos interrelacionados de desafíos, técnicos y normativos: la cuestión técnica pregunta «cómo codificar los valores humanos en la inteligencia artificial», mientras que la cuestión normativa concierne a qué valores deberían guiar la alineación (Sutrop, 2020). Incluso la especificación de los valores objetivo es disputada —la IA podría alinearse plausiblemente con el interés agregado de todos, con prioridad para los peor situados, con una concepción de la virtud del carácter moral o con una ética del reconocimiento mutuo, y ningún candidato individual reúne consenso (Gabriel, 2020). Una segunda dificultad concierne a la legitimidad: dado el abanico de principios

posibles, no está claro quién tiene el derecho de tomar las decisiones de alineación, y sobre qué base (Gabriel, 2020). Subyace a estas cuestiones el hallazgo de que los valores humanos son dinámicos, complejos y difíciles de especificar (Han et al., 2021). Las propuestas técnicas como el aprendizaje por refuerzo inverso y el modelado de recompensas presentan cada una inconvenientes notables, y ningún método único proporciona una solución integral (Singh, 2025). Una revisión sistemática de 172 artículos sobre alineación de valores concluye que la alineación se entiende mejor no como una tarea de especificación única, sino como un proceso continuo entre humanos y agentes autónomos que debe gestionar las exigencias éticas y políticas conflictivas entre grupos (McKinlay et al., 2026).

2.2 EQUIDAD Y SESGO

La toma de decisiones algorítmica está hoy incrustada en la atención médica, el transporte, la educación, las admisiones universitarias, la contratación y los préstamos, y la investigación ha mostrado que tales sistemas pueden ser inherentemente propensos a la injusticia incluso cuando no existe intención de discriminar (Pessach & Shmueli, 2022). Fallos bien documentados ilustran lo que está en juego en la práctica: el chatbot Tay de Microsoft produjo comentarios racistas y sexistas en las 24 horas siguientes a su lanzamiento, un software de predicción de reincidencia asignó puntuaciones de riesgo más altas a afroamericanos que a caucásicos con perfiles similares, y un sistema de recomendación de empleo promovió más oportunidades STEM para hombres que para mujeres (Liu et al., 2022). El campo permanece conceptualmente inestable: los estudios identifican múltiples dimensiones distintas de sesgo —sesgo de datos de entrenamiento, sesgo de modelo, conceptos de equidad en conflicto y ausencia de transparencia de predicción— y ninguna definición única de equidad que satisfaga todos los contextos (Jui & Rivas, 2024). Los mecanismos de mitigación se agrupan convencionalmente en enfoques de preprocesamiento, procesamiento interno y posprocesamiento, aunque identificar qué mecanismo se adecúa a cada escenario sigue siendo un desafío abierto (Pessach & Shmueli, 2022). Los estudios del campo concluyen con desafíos abiertos

articulados como cinco dilemas no resueltos para la investigación sobre equidad (Caton & Haas, 2023), y la dificultad se ve agravada por el hecho de que los modelos profundos cada vez más utilizados en decisiones de alto riesgo vuelven opaco y difícil de auditar su razonamiento (Petersen & Zech, 2024).

2.3 EXPLICABILIDAD, TRANSPARENCIA Y RENDICIÓN DE CUENTAS

La complejidad de las redes neuronales profundas hace que explicar predicciones individuales esté lejos de ser directo —a pesar del conocimiento completo de los pesos y sesgos de una red, la interacción repetida de transformaciones lineales y activaciones no lineales convierte a los modelos en cajas negras (Petersen & Zech, 2024). Las explicaciones son, sin embargo, éticamente necesarias cuando los modelos asignan recursos escasos como trasplantes de órganos, aprobación de créditos, contratación y navegación de vehículos autónomos (Mehta et al., 2023; Petersen & Zech, 2024). Más allá de la explicabilidad se encuentra el problema de la rendición de cuentas. Los estudiosos distinguen una «brecha de rendición de cuentas pública» —la incapacidad de los ciudadanos de obtener explicaciones por las decisiones tomadas por agencias públicas— de una «brecha de rendición de cuentas moral» más amplia, en la que los algoritmos opacos erosionan la capacidad de los agentes humanos de dar sentido a la «lógica» de su propio comportamiento y de explicársela (de Sio & Mecacci, 2021). Esto se conecta con la muy discutida brecha de responsabilidad: situaciones en las que un sistema autónomo actúa sin que nadie —ni el programador, ni el fabricante, ni el operador, ni el propio sistema— sea responsable de la acción y sus consecuencias (Königs, 2022). La brecha es a menudo tan práctica como conceptual: en el comercio de alta frecuencia, la conducción autónoma o la defensa antimisiles automatizada, la toma de decisiones humana es simplemente demasiado lenta para intervenir, lo que hace imposible una «decisión final» humana (Coeckelbergh, 2019). Que la brecha sea genuina, sin embargo, es algo disputado. Algunos sostienen que responsabilizar a una computadora por un daño es un teatro carente de sentido a menos que el sistema posea al menos el mismo nivel de inteligencia,

autoconciencia y sensibilidad moral que un ser humano adulto, dejando un vacío donde debería haber una parte responsable (Goetze, 2022); otros sostienen que la responsabilidad moral es un proceso dinámico y flexible que puede abarcar efectivamente a entidades tecnológicas emergentes sin ninguna brecha basada en la tecnología (Tigard, 2020).

2.4 RESPONSABILIDAD Y RENDICIÓN DE CUENTAS EN EL USO COTIDIANO DE LA IA

El uso cotidiano de la IA —motores de búsqueda, fuentes de redes sociales, herramientas de contratación, calificación crediticia, aplicaciones de salud, navegación— plantea una cuestión práctica de rendición de cuentas: quién ostenta el poder de garantizar el uso correcto, y quién asume la responsabilidad cuando el uso sale mal. La literatura distribuye este poder a lo largo de una cadena de actores en lugar de localizarlo en una única parte (Novelli et al., 2023). Los diseñadores de sistemas son responsables de construir sistemas que satisfagan las necesidades de los usuarios, sean transparentes y explicables, y de proporcionar instrucciones de implementación y divulgar los riesgos potenciales (Liu et al., 2022). Los tomadores de decisiones, que deciden si construir o adoptar un sistema, deben ponderar los beneficios frente a los riesgos y considerar todas las regulaciones pertinentes antes de la implementación; los implementadores de sistemas deben seguir las instrucciones de implementación y capacitar a los usuarios finales; los auditores deben evaluar los sistemas de manera integral y objetiva; y los usuarios finales deben seguir las pautas de uso e informar con prontitud los problemas emergentes (Liu et al., 2022). Las organizaciones internacionales y los reguladores son responsables de guiar correctamente qué puede hacer la IA y dónde residen sus riesgos (Walter, 2024). Esta estructura de roles se corresponde con el ciclo de vida de la IA de diseño, desarrollo e implementación, con agentes identificables individual, corporativa, colectiva o jerárquicamente (Novelli et al., 2023).

En cuanto a la división entre lo «correcto» y lo «incorrecto» en el uso diario, una regla ampliamente compartida es que las empresas y los usuarios asumen juntos la responsabilidad final: las empresas deben hacer todo lo posible por contrarrestar el mal

uso y el abuso de los sistemas de IA, mientras que los usuarios son culpables si crean deliberadamente resultados dañinos o no son suficientemente cuidadosos al automatizar sus tareas (Liu et al., 2022). De manera similar, los desarrolladores son principalmente responsables del diseño y la funcionalidad de una IA —los daños incrustados en el sistema les corresponden a ellos—, mientras que los daños derivados del uso y la implementación corresponden al usuario organizacional (Ryan & Stahl, 2020). Este consenso, sin embargo, no es universal: los profesionales discrepan sobre si los profesionales de la IA asumen la rendición de cuentas por los daños causados por sus sistemas o si esa responsabilidad se comparte con los usuarios (Pant et al., 2024).

Más allá del quién, varios factores determinan si el uso cotidiano resulta correcto o incorrecto. Las decisiones de diseño son de naturaleza social, legal y ética y no pueden dejarse únicamente a ingenieros, fabricantes o usuarios; la IA responsable exige que tales decisiones se informen y permanezcan abiertas a la inspección (Dignum, 2021). La transparencia y la explicabilidad importan porque los sistemas opacos impiden a los usuarios juzgar si una decisión es apropiada, y los mecanismos de rendición de cuentas dependen de hacer la IA inteligible para el público (Radanliev et al., 2024). La calidad de los datos es otro factor: las organizaciones deben ser conscientes de los problemas que implica usar datos deficientes y ser responsabilizadas si se derivan consecuencias dañinas. La educación y el discurso público son igualmente importantes, pues fomentar una cultura del uso moral de la tecnología exige ética en la educación —de lo contrario, los usuarios evalúan la IA sobre la base de sesgos personales más que de una comprensión de cómo funciona (Dignum, 2021). Finalmente, las salvaguardas regulatorias e institucionales —como grupos de ciudadanos, canales de retroalimentación y evaluaciones de impacto— son lo que convierte los principios en práctica efectivamente aplicada.

Una complicación crítica permanece: la autonomía relativa de los sistemas de IA crea una brecha de responsabilidad, lo que hace poco claro quién es responsable cuando ocurre un daño (Königs, 2022). Los estudiosos distinguen una brecha de rendición de cuentas pública —ciudadanos incapaces de obtener explicaciones por decisiones automatizadas de agencias públicas— de

una brecha de rendición de cuentas moral, en la que los algoritmos opacos erosionan la capacidad de los agentes responsables de dar sentido a su propio comportamiento (de Sio & Mecacci, 2021). La misma distribución del poder entre diseñadores, implementadores, auditores y usuarios puede por tanto difuminar la responsabilidad, precisamente donde los daños de la vida diaria son más probables (Lima & Cha, 2020). Si esta brecha es genuina o puede tenderse un puente sobre ella sigue siendo objeto de debate activo (Tigard, 2020). En suma, el poder sobre el uso correcto e incorrecto de la IA se comparte entre diseñadores, implementadores, auditores, usuarios y reguladores —pero el compartir mismo, sin mecanismos claros de atribución, sigue siendo parte del problema no resuelto.

2.5 MORAL DE LAS MÁQUINAS EN DECISIONES DE ALTO RIESGO

Los vehículos autónomos se cuentan entre los primeros robots comerciales con IA incrustada que toman decisiones autónomas, y los accidentes inevitables con consecuencias de vida o muerte no pueden eliminarse por completo —una situación que plantea la cuestión de cómo tales sistemas deberían tomar decisiones éticas, y cómo su comportamiento moral puede hacerse explicable y aceptable para los usuarios (Rhim et al., 2021). El Experimento de la Máquina Moral recopiló aproximadamente 40 millones de respuestas de millones de individuos de 233 jurisdicciones, identificando fuertes preferencias globales por salvar vidas humanas, salvar más vidas y salvar vidas jóvenes (Chen et al., 2022; Lin, 2021). Sin embargo, persiste una profunda variación cultural y demográfica: los encuestados de países de Asia Oriental tendieron a preferir salvar a los ancianos sobre los jóvenes, mientras que los de América del Norte y Europa mostraron la preferencia opuesta, y los estadounidenses expresaron preferencias más fuertes por vehículos autoprotectores cuando viajaban con la familia que los encuestados japoneses (Chen et al., 2022; Lin, 2021). Para los formuladores de políticas, esto plantea un doble problema: experimentos como la Máquina Moral no pueden por sí mismos fundamentar la regulación, pues los factores culturales moldean fuertemente la opinión pública sobre la ética de

los vehículos (Souza & Castañon, 2021), y el marco del dilema hipotético ha sido criticado a su vez por ser excesivamente simplista y determinista, al no representar las condiciones reales de decisión (Kochupillai et al., 2020). Más fundamentalmente, parece no haber una manera fácil de diseñar algoritmos que reconcilien los valores morales y el interés personal a través de culturas con actitudes morales diferentes (Bonnefon et al., 2016), y varios estudiosos sostienen que el marco del problema del tranvía sobrecarga la ética del dilema mientras descuida preocupaciones más realistas como la interacción humano-máquina, la percepción mecánica y la privacidad de los datos (Chen et al., 2022). En respuesta, la literatura se mueve hacia enfoques pluralistas —un pluralismo razonable que exige argumentos de razón pública para las decisiones de diseño (Brändle & Schmidt, 2021), y un pluralismo epistémico moral que trata cada vehículo y el sistema más amplio como capas distintas de consideración moral (Henschke & Arora, 2024).

2.6 GOBERNANZA, REGULACIÓN Y EJECUCIÓN

La gobernanza de la IA se refiere a los marcos, políticas y mecanismos —desde la regulación gubernamental hasta la autogobernanza corporativa— que guían el desarrollo y la operación de los sistemas de IA en consonancia con los principios éticos, las normas sociales y los estándares legales (Walter, 2024). La investigación sobre gobernanza identifica brechas de conocimiento persistentes, incluida la operacionalización insuficiente de los procesos de gobernanza a nivel organizacional y la ausencia de mecanismos de responsabilidad que garanticen el cumplimiento de los principios éticos (Birkstedt et al., 2023). Una agenda de investigación dedicada a la gobernanza técnica de la IA cataloga problemas abiertos: identificar qué dimensiones técnicas se alinean mejor con la regulación (por ejemplo, si los umbrales de cómputo de entrenamiento son medidas de riesgo adecuadas), determinar cómo evaluar las opciones de gobernanza y diseñar mecanismos efectivos de ejecución (Reuel et al., 2024). La regulación misma es un acto de equilibrio: la gobernanza ágil es ampliamente instada porque los marcos regulatorios actuales no dan cuenta adecuadamente del cambio tecnológico rápido

(Dwivedi et al., 2019), pero la regulación excesivamente estricta se contrarresta con advertencias de que puede sofocar la innovación (Dwivedi et al., 2019; Leech et al., 2024). Además, las directrices convergen al nivel de los principios de alto nivel pero divergen marcadamente en la interpretación y la aplicación (Ashok et al., 2021), y un desafío fundamental para cualquier agenda global es equilibrar la armonización transnacional con el respeto por el pluralismo cultural y moral (Jobin et al., 2019).

2.7 LA METACUESTIÓN: ¿PUEDE RESOLVERSE LA ÉTICA DE LA IA?

Un último conjunto concierne a la viabilidad de toda la empresa. Se han propuesto docenas —de hecho, más de un centenar— de marcos éticos, modelos de proceso y herramientas, y sin embargo muchos son demasiado abstractos para traducirse fácilmente en diseños concretos de sistemas (Prem, 2023). El «principlismo» resultante se centra generalmente en un pequeño conjunto de principios como la beneficencia, la no maleficencia, el respeto por la autonomía y la justicia, reflejando la ética biomédica, pero los principios por sí solos no especifican cómo deberían aplicarse en un contexto particular (Jedličková, 2024; Prem, 2023). Los esfuerzos por operacionalizarlos son extensos —una revisión sistemática identifica 106 herramientas, otra 39—, y sin embargo la mera plétora de principios resulta abrumadora y confusa para los desarrolladores, y la mayoría del personal técnico los desconoce (Ortega-Bolaños et al., 2024). Los comentaristas observan que la ética de la IA ha evolucionado a través de fases, de los principios de alto nivel (2016–2019) a una fase técnica de equidad-rendición de cuentas-transparencia y ahora una fase marcada por el movimiento «del qué al cómo», centrada en mecanismos de gobernanza, auditoría y estándares (Ayling & Chapman, 2021). Los críticos advierten que las pautas de marcar casillas no deben ser los únicos «instrumentos» de la ética de la IA, y piden una transición hacia enfoques sensibles a la situación y basados en la virtud (Hagendorff, 2020). Incluso el objetivo último es disputado: el florecimiento humano se propone como el principio ético orientador, y sin embargo no es trivial determinar qué constituye el florecimiento ni cómo contribuye la tecnología a él (Stahl et al.,

2020), y se ha acusado a la ética de la IA de promover el bienestar de algunos mientras socava activamente el de otros (Munn, 2022).

2.8 EL DILEMA ESTABILIDAD-PLASTICIDAD: APRENDIZAJE CONTINUO Y GENERALIDAD DE LA IA

2.8.1 *Lo que el artículo realmente establece*

El dilema estabilidad-plasticidad se define clásicamente como la necesidad de que los modelos sean «suficientemente estables para retener el conocimiento aprendido de clases vistas previamente, y suficientemente plásticos para aprender conceptos de clases nuevas», mientras que el ajuste fino ingenuo sufre olvido catastrófico porque la alta plasticidad sobrescribe parámetros importantes para las tareas antiguas (Kim & Han, 2023). El estudio fundacional de este dilema, sin embargo, es diagnóstico más que prescriptivo, y su afirmación central es casi lo opuesto a una solución: la mayoría de los algoritmos de aprendizaje incremental por clases «favorecen fuertemente la estabilidad sobre la plasticidad», hasta el punto de que el extractor de características entrenado sobre las clases iniciales no es «menos eficaz que el del modelo incremental final» (Kim & Han, 2023). Midiendo la plasticidad al reentrenar clasificadores en cada etapa y la estabilidad mediante la Alineación de Núcleos Centrada, los autores constatan que la mayoría de los modelos «acumulan poco conocimiento nuevo en sus representaciones de características a lo largo de las etapas incrementales» —vencen el olvido catastrófico simplemente negándose a aprender (Kim & Han, 2023). La demostración más tajante es su «Exploit» deliberadamente degenerado: congela por completo el extractor de características y entrena solo una pequeña cabeza, y sin embargo alcanza un 67,2 % de precisión incremental promedio y un 60,5 % de precisión final en el escenario de 500 clases —competitivo con todos los métodos salvo DER y pDER, frente a un techo «Oracle» de 72,8 %/70,8 % (Kim & Han, 2023). El veredicto de los autores: las métricas CIL estándar «ofrecen una falsa ilusión de progreso» (Kim & Han, 2023). Solo los métodos de aislamiento de parámetros —DER, que apila un nuevo extractor de características por etapa— exhiben plasticidad genuina, mostrando una precisión creciente por

etapa donde otros cinco algoritmos permanecen planos y solo el Oracle crece casi linealmente (Kim & Han, 2023). Pero DER exige N+1 pasadas hacia adelante para N pasos, y la modificación del artículo (pDER) solo mitiga ese costo (Kim & Han, 2023). La cuestión teórica central —si un equilibrio fuerte entre estabilidad y plasticidad es siquiera posible sin añadir parámetros— permanece abierta, aunque el teorema de Raghavan y colaboradores garantiza que puede existir un punto de equilibrio estable entre el olvido y la generalización (Kim & Han, 2023).

2.8.2 *La conexión humana: ¿es el dilema «epistemológicamente sólido»?*

El dilema es empíricamente real en los seres humanos: ha sido «ampliamente estudiado tanto para sistemas computacionales como biológicos» (Parisi et al., 2018), y la teoría de los sistemas de aprendizaje complementarios explica cómo la biología lo resuelve —un sistema hipocampal rápido que codifica rápidamente los detalles, emparejado con un sistema neocortical lento que extrae la estructura mediante aprendizaje entrelazado y reproducido (Kumaran & McClelland, 2012; McClelland et al., 1995). Los humanos sí muestran fenómenos de interferencia (por ejemplo, aprender asociaciones A–C tras haber codificado A–B), aunque «mucho menos dramáticos» que las redes (Saxe et al., 2020). Pero el dilema es una restricción arquitectónico-cognitiva, no un principio epistémico. La epistemología concierne a la verdad, la justificación y la creencia; el dilema concierne a la retención frente a la adaptación de la información almacenada. Lo que la teoría sí apoya es la afirmación más débil: la disyuntiva no es un artefacto del aprendizaje profundo sino una condición general de los sistemas de aprendizaje finitos con fundamento matemático (Kim & Han, 2023) —de modo que la analogía humana se sostiene en el sentido de ser principiada e inevitable, no en el sentido epistemológico.

2.8.3 *¿Autoriza esto la inteligencia artificial general — y vuelve irrelevante el enfoque lógico?*

No, en ninguno de los dos casos, por tres razones independientes. Primero, los resultados positivos son estrechos y en parte ilusorios. El escenario es la clasificación de imágenes incremental por clases; incluso el techo del Oracle se sitúa en ~71–73 % de precisión (Kim & Han, 2023). El trabajo más amplio coincide en que el aprendizaje incremental por clases con entradas complejas «sigue siendo un problema abierto» a menos que los datos se almacenen explícitamente (van de Ven et al., 2020), que «sigue existiendo una enorme brecha entre los avances existentes y las aplicaciones realistas» (Wang et al., 2023), y que la plasticidad misma se degrada con el entrenamiento prolongado —el problema de la «pérdida de plasticidad» (Dohare et al., 2024). La percepción continua está muy lejos de la inteligencia general, que exige razonamiento, lenguaje, planificación, explicación y juicio de valor —competencias que esta línea de trabajo no toca. Segundo, las soluciones genuinas son computacionalmente explosivas, y las honestas son trampas. El único enfoque que muestra estabilidad y plasticidad reales multiplica el cómputo por el número de etapas (Kim & Han, 2023), y el propio «Exploit» del artículo muestra con qué facilidad se manipulan los puntos de referencia (Kim & Han, 2023). Una IAG construida sobre esta base heredaría un problema de escalabilidad no resuelto y una cuestión teórica no zanjada (Kim & Han, 2023). Tercero, la biología es evidencia a favor de la hibridación, no en contra de la lógica. La solución del propio cerebro es una división del trabajo: un sistema rápido basado en episodios y un sistema lento extractor de estructura (Kumaran & McClelland, 2012; McClelland et al., 1995). El enfoque lógico aborda una competencia completamente distinta —la representación del conocimiento como simbolización formal de proposiciones creídas y el razonamiento como su manipulación formal (Lee, 2024), con los límites de expresividad-tratabilidad tratados antes en esta revisión, y con la maquinaria formal-semiótica del hexágono (Servín, s.f.-b) ocupándose del estatus epistémico y normativo de las afirmaciones. El dilema estabilidad-plasticidad trata del mantenimiento de la memoria; la lógica trata de la implicación y la explicación. Ninguno vuelve

irrelevante al otro —y el cerebro humano, que enfrenta el dilema y aun así razona lógicamente, es el contraejemplo permanente. La conclusión defendible es que el dilema es una restricción genuina y compartida del aprendizaje biológico y artificial —la analogía humana se sostiene empíricamente—, pero esto no autoriza ni la IAG ni la irrelevancia de la lógica; a lo sumo refuerza el caso a favor de las arquitecturas híbridas, que es, no incidentalmente, lo que el cerebro ya es.

2.9 CRÍTICA RADICAL CONTRA LA TECNOLOGÍA: EL DESAFÍO DE KACZYNSKI

El manifiesto de Theodore Kaczynski —*La sociedad industrial y su futuro*— ha sido ahora objeto de un cuidadoso análisis filosófico e histórico, incluso mientras sus métodos como autor son universalmente repudiados (Fleming, 2021; Moen, 2018). Aquí se examina como *discurso*, no como interlocutor filosófico en pie de igualdad. Barnett muestra que el manifiesto combina apelaciones emocionales con «evidencia verificable» (vigilancia, concentración mediática, riesgos ambientales) dentro de una dicotomía tecnología/naturaleza simplificada a propósito, y que la violencia fue para su autor una estrategia de difusión (Barnett, s.f., pp. 62-65). La lección metodológica es que la fuerza persuasiva de un texto es independiente de la verdad de sus afirmaciones: reconocer algunos de sus hechos no obliga a aceptar sus inferencias ni, menos aún, su programa. Con esa cautela, su desafío al proyecto reformista de la ética de la IA merece examen, porque formula de manera extrema una pregunta legítima: si la reforma desde dentro puede fracasar aun cuando todos sepan que es necesaria.

2.9.1 *Lo que exige el manifiesto*

Tres afirmaciones estructuran *La sociedad industrial y su futuro*:

1. La tecnología destruye el «proceso de poder». Los seres humanos necesitan metas y la libertad de perseguirlas, enfrentando los riesgos de sus propias decisiones; la transferencia progresiva de la toma de decisiones a sistemas

impersonales y deshumanizadores impide a los individuos realizar y perseguir estas necesidades fundamentales (Cappuccio et al., 2023). Se opuso a la tecnología principalmente por razones evolutivo-psicológicas —la sociedad moderna obliga a las personas a vivir bajo condiciones radicalmente diferentes de aquellas bajo las cuales evolucionó la raza humana (Fleming, 2024).

2. Las máquinas terminarán por tener el control. Su advertencia más conocida: «la raza humana podría fácilmente permitirse derivar hacia una posición de tal dependencia de las máquinas que no tendría más opción práctica que aceptar todas las decisiones de las máquinas... en esa etapa las máquinas estarán en control efectivo» (Yampolskiy, 2022).

3. La reforma es imposible; solo la disolución basta. La industrialización ha sido terrible para la humanidad, y solo un retorno al modo de vida preindustrial puede remediarlo —por cualquier medio necesario, en su opinión (Moen, 2018).

2.9.2 *Dónde el documento coincide parcialmente*

La brecha de responsabilidad es la coincidencia más estrecha. En la sección «Responsabilidad y rendición de cuentas en el uso cotidiano de la IA», el documento señala que la autonomía relativa de la IA crea una brecha de responsabilidad y que distribuir el poder entre diseñadores, implementadores, auditores y usuarios «puede también difuminar la responsabilidad, precisamente donde los daños de la vida diaria son más probables». El diagnóstico se aproxima a la segunda afirmación de Kaczynski —que las decisiones migran a las máquinas más rápido de lo que los humanos pueden mantener un control inteligible sobre ellas (Yampolskiy, 2022)—, aunque la coincidencia se limita a la descripción del problema y no se extiende a sus causas ni a sus remedios. Ambos textos diagnostican una erosión de la agencia humana efectiva —el manifiesto atribuye a ello un malestar psicológico generalizado (Fleming, 2024), mientras que el documento lo trata como un déficit de rendición de cuentas.

La moral de las máquinas en decisiones de alto riesgo. La crítica del documento a la ética del problema del tranvía —que el marco del dilema sobrecarga los escenarios artificiales y no representa las condiciones reales de decisión— converge con la objeción más profunda del manifiesto, que se dirige a los procedimientos formales e impersonales de decisión que reemplazan por completo el juicio humano (Cappuccio et al., 2023). Kaczynski leería el propio proyecto de «programar la ética en las máquinas» como un error categorial: la agencia ética exige enfrentar las consecuencias de las propias decisiones, que es exactamente lo que la delegación a las máquinas erosiona.

La metacuestión (¿puede resolverse la ética de la IA?). El documento documenta el fracaso del principlismo, la proliferación de herramientas inconmensurables y la noción disputada del florecimiento humano. La respuesta de Kaczynski a esa frustración es la radical: no puede —no dentro del sistema, porque la complejidad y el impulso del sistema desbordan cualquier correctivo. Moen constata que «aunque su teoría en su conjunto debería rechazarse», algunas preocupaciones puntuales sobre el desarrollo tecnológico «han llegado recientemente a ser compartidas por destacados defensores de la mejora humana», incluidos Bostrom y Savulescu (Moen, 2018). La observación es de Moen y está acotada: la convergencia se limita a la preocupación por la deriva inducida por la IA (Yampolskiy & Fox, 2012), no a sus fundamentos. La investigación en seguridad atribuye los riesgos a la mala especificación de objetivos y a la instanciación perversa de metas formalmente correctas, no a un supuesto impulso de dominación de los artefactos ni a un determinismo evolutivo-psicológico.

2.9.3 *Dónde el documento diverge fundamentalmente*

La equidad y el sesgo constituyen el desacuerdo más agudo. El documento trata el sesgo algorítmico como un defecto que debe corregirse mediante mejores datos, mejores métricas y mejor mitigación. Para Kaczynski, las normas igualitarias, imparciales y no violentas que respaldan tal corrección son ellas mismas síntomas de «sobresocialización» —normas inculcadas precisamente para que los individuos funcionen sin fricción dentro de sistemas complejos de producción, e imponiendo una carga que «produce una

sensación de constricción e impotencia» (Fleming, 2024). Desde su perspectiva —que este documento no comparte—, la ingeniería de la equidad le parecería el sistema reparando su propia legitimidad en lugar de abordar la totalidad deshumanizadora (Galliott, 2017).

Gobernanza, regulación y ejecución. Toda la sección de gobernanza del documento presupone lo que Kaczynski niega de plano: que el sistema puede reformarse desde dentro. Sus hallazgos —las directrices divergen en la interpretación, la operacionalización fracasa, la ejecución está ausente— son precisamente la evidencia que Kaczynski citaría para su conclusión estructural de que la corrección fragmentaria es fútil (Moen, 2018). El documento ve una brecha de gobernanza; él ve una ilusión de gobernanza.

Alineación de valores. El documento pregunta qué valores y qué autoridad deberían guiar la alineación (Gabriel, 2020; Singh, 2025). Kaczynski respondería que la pregunta está amañada: el sistema industrial-tecnológico no puede ser puesto al servicio de los valores humanos porque subordina los fines humanos a sus propias exigencias funcionales. Un análisis filosófico de su posición concluye en efecto que los usuarios de sistemas tecnológicos complejos se enfrentan a un trilema —«tecnología encarcelada controlada por quienes se hallan dentro de las estructuras de poder opresivas existentes», dedicar la propia vida a dominar los sistemas, o «rebelarse contra dicho sistema» (Galliott, 2017). La alineación como ingeniería no figura en el menú de Kaczynski.

2.9.4 *El veredicto fundamental*

La coincidencia es diagnóstica, no prescriptiva. Donde el documento y el manifiesto se encuentran es en el diagnóstico: la complejidad desborda la comprensión humana, la autonomía difumina la responsabilidad, la agencia se erosiona y los principios no logran constreñir la práctica. Donde se separan irreconciliablemente es en la prescripción: el documento busca gobernar, explicar, alinear y regular la tecnología desde dentro; Kaczynski exigía su abolición. Incluso su única premisa de apariencia compartida es discutible —los críticos señalan que su suposición de que las máquinas «buscarán dominar» atribuye a los artefactos

un rasgo humano y evolucionado (el deseo de dominación) que estos no poseen (Yampolskiy & Fox, 2012)—, y su teoría en su conjunto ha sido juzgada filosóficamente implausible, siendo su violencia el producto de una fe ciega en una utopía antitecnológica (Lagarde, 2019; Moen, 2018). Su crítica funciona, no obstante, como una condición de frontera, y conviene formularla de modo que sea evaluable y no fatalista. No afirma que el sistema no pueda reformarse, sino que traslada la carga de la prueba: la ética de la IA debe mostrar que la gobernanza y la rendición de cuentas funcionan en la práctica. Esa exigencia tiene una versión analítica que no requiere ni determinismo ni utopía antitecnológica. Falk y Tsoukalas demuestran que, cuando cada empresa captura todo el ahorro de automatizar pero soporta solo una fracción de la pérdida de demanda resultante, la automatización excesiva es una estrategia dominante: *conocer el riesgo no basta para corregirlo*, porque los incentivos tienen la estructura de un dilema del prisionero; pero el mismo análisis identifica instrumentos —un impuesto pigouviano a la automatización— que restauran el óptimo cooperativo (Falk & Tsoukalas, 2026, Proposiciones 1 y 5). La pregunta de Kaczynski se transforma así en una pregunta empírica e institucional: qué evidencia mostraría que la gobernanza funciona —principios operacionalizados y auditados, sanciones efectivamente aplicadas, vías de impugnación usadas por los afectados, supervisión humana que no se erosiona a medida que los sistemas parecen más fiables— y qué estructuras de incentivos la hacen posible. Si esa evidencia existe, la tesis abolicionista queda refutada en ese punto; si no existe, la reforma tiene la carga de la prueba. Su argumento del «proceso de poder» sigue siendo, por otra parte, el concepto más cercano a las secciones de este documento sobre la agencia y la autonomía humanas, y la influencia del manifiesto en el radicalismo antitecnológico contemporáneo se ha rastreado hasta una síntesis de Jacques Ellul, Desmond Morris y Martin Seligman (Fleming, 2021).

2.10 LOS LÍMITES EPISTÉMICOS DE LA REPRESENTACIÓN: POR QUÉ LAS AFIRMACIONES DE «CUALQUIER PROBLEMA» SON INJUSTIFICADAS

Un desafío adicional a las ambiciones de los sistemas de IA, epistémico más que político, se sigue de la propia teoría de la representación del conocimiento. El análisis de Brachman (1979) dividió el nivel epistemológico de la representación en cuatro capas distintas: la *implementacional*, relativa a las estructuras de datos; la *lógica*, relativa a la lógica simbólica, las proposiciones, los predicados, las variables y los cuantificadores; la *epistemológica*, relativa a los tipos de concepto, subtipos, herencia y relaciones estructurantes; y la *conceptual*, relativa a las relaciones semánticas, los roles lingüísticos, los objetos y las acciones (Sowa, 1999). El nivel epistemológico es donde la realidad se talla en categorías —y cualquier tallado de este tipo es una elección más que un dato (Sowa, 1999). Según la definición canónica, la representación del conocimiento es el uso de símbolos formales para representar una colección de proposiciones creídas por algún agente putativo, de las cuales solo un número finito se representa alguna vez, y el razonamiento es la manipulación formal de esos símbolos para producir representaciones de nuevas proposiciones (Lee, 2024). El razonamiento, dicho de otro modo, opera sobre un modelo, no sobre el mundo.

La representación es, por tanto, inherentemente parcial y relativa a propósitos. La misma sustancia puede representarse con axiomas diferentes, incluso contradictorios, en distintos contextos y para distintos fines: el ingeniero trata la lechada de carbón como continua para el transporte y como grumosa para la separación, y ambas representaciones son legítimas (Sowa, 1999). La «precisión» de un sistema es, en consecuencia, precisión bajo una representación elegida —relativa a los axiomas, las categorías y la granularidad que los diseñadores humanos seleccionaron. Afirmar la precisión sin calificaciones comete precisamente la confusión categorial que los niveles de Brachman fueron diseñados para exponer: confunde el modelo con el mundo (Sowa, 1999).

El razonador implementado ha pagado además ya un precio de expresividad-tratabilidad. La literatura es explícita en que

incluso pequeñas dosis de poder expresivo pueden acarrear un costo computacional significativo, y que las cuestiones de qué controla la dificultad computacional, hasta dónde puede empujarse la expresividad y cuándo bastan lenguajes inexpresivos pero tratables siguen en gran medida sin respuesta (Lee, 2024). La lógica de primer orden es indecidible —demostrado independientemente por Church y Turing en 1936—, y los investigadores han tenido que retirarse a lenguajes deliberadamente más débiles porque la búsqueda de implicaciones es computacionalmente desesperanzada en general (Lee, 2024). El propio análisis de nivel de conocimiento de Newell pone en primer plano la complejidad computacional intrínseca de la relación de implicación como una cuestión de adecuación central para cualquier lenguaje de representación (Sowa, 1999). Un sistema capaz de «cualquier problema» exigiría así capacidades que el campo no sabe cómo construir, en principio más que meramente en la práctica.

Dos límites adicionales cierran las vías restantes hacia la afirmación fuerte. Primero, la equidad no es una propiedad epistémica: los niveles de Brachman describen cómo se estructura el conocimiento, no cómo se fundamentan los valores, y el nivel epistemológico codifica tipos de concepto y herencia sin vocabulario para la justicia, el daño o el mérito (Sowa, 1999). La selección de qué representar, con qué granularidad y bajo qué axiomas es un acto humano cargado de valor —pero la maquinaria que razona sobre esa selección es silenciosa respecto de los valores. Afirmar que un sistema resuelve problemas «con equidad» es afirmar que una propiedad normativa se sigue de la manipulación formal de símbolos, una inferencia que la propia distinción conocimiento/símbolo bloquea. Segundo, la resolución de problemas humana se apoya en la intuición, que concierne a los particulares más que a las verdades generales, requiriendo estas últimas inferencia inductiva o abductiva (Servín, s.f.-a); el razonamiento formal sobre una base de conocimiento es deductivo, y la brecha entre lo representado y lo creído no es cerrada por el marco representacional. Ni existe lógica privilegiada alguna que la cierre: la lógica interna del hexágono lógico-semiótico no es necesariamente booleana sino dependiente del modelo descrito por sus categorías (Servín, s.f.-b). No hay lógica canónica, ni granularidad canónica, ni modelo canónico —de ahí que no

haya ruta desde ninguna base de conocimiento hasta «cualquier problema, con precisión».

La cuestión ética se sigue directamente. Las afirmaciones injustificadas de precisión y equidad inducen a error a usuarios, reguladores y público sobre la confiabilidad de los sistemas; transfieren la confianza del juicio humano a los modelos; y diluyen la rendición de cuentas precisamente donde los daños son más probables, pues el marco de «el sistema lo resolvió» oscurece las decisiones humanas —qué axiomas, qué categorías, qué granularidad, qué valores— que en realidad determinaron el resultado (de Sio & Mecacci, 2021; Königs, 2022). Este es el puente epistémico hacia el análisis de la brecha de responsabilidad desarrollado antes: si la competencia de un sistema es siempre competencia-relativa-a-un-modelo, entonces quien eligió el modelo asume una responsabilidad que la retórica de la precisión autónoma borra. La afirmación defendible no es que la IA pueda resolver con precisión y equidad cualquier problema que los humanos puedan resolver, sino la afirmación más débil y fiel a la representación: la IA puede generar implicaciones de una representación elegida, con una granularidad elegida, bajo una lógica elegida, dentro de límites de tratabilidad —y cada una de esas elecciones es una decisión humana cargada de valor que el propio sistema no puede tomar (Lee, 2024; Sowa, 1999).

2.11 VALIDEZ, INTERPRETACIÓN Y LA ÉTICA DE LA AUTORIDAD EPISTÉMICA DELEGADA

El análisis epistemológico de la representación plantea una cuestión ulterior de legitimidad científica y ética: si los resultados de los sistemas de IA pueden tratarse propiamente como interpretaciones válidas de los datos que procesan. La distinción lógica entre validez y solidez proporciona el vocabulario preciso. Un argumento deductivo es válido si, fueran verdaderas todas sus premisas, la conclusión tendría que ser verdadera; es sólido solo si es válido *y* todas sus premisas son de hecho verdaderas (Servín, s.f.-a). Aplicada a los sistemas de IA, la distinción es decisiva. Un sistema de razonamiento bien construido deriva sus resultados correctamente de sus entradas representadas —manipulación de símbolos, inferencia o generalización estadística

a partir de los datos que se le dieron—, y en ese sentido restringido un motor de IA puede ser un dispositivo *válido*. La solidez, sin embargo, no es un producto entregable de la maquinaria: exige que las premisas sean verdaderas, y las premisas de un sistema de IA son su representación, a saber, las categorías, los axiomas, la granularidad, los datos y las variables objetivo seleccionados por diseñadores humanos (Sowa, 1999). Puesto que la representación del conocimiento codifica solo un número finito de las proposiciones creídas por algún agente, y el razonamiento es la manipulación formal de esos símbolos, el sistema opera sobre un modelo más que sobre el mundo (Lee, 2024). La verdad sobre el mundo nunca es derivada por el sistema; es importada a través de la representación. Un sistema cuyas premisas son falsas —datos sesgados, una variable objetivo mal especificada, una granularidad inapropiada— es formalmente válido y materialmente insólido.

El estatus epistémico de la producción mecánica se relativiza además por el marco interpretativo en el que se coloca. Lo que cuenta como justificación depende de la epistemología que uno adopte: si hay epistemologías distintas, el término «necesidad epistémica» se vuelve relativo, mientras que la necesidad lógica debe entenderse según las leyes de alguna lógica —y qué lógica es, es por sí mismo una cuestión abierta (Servín, s.f.-a). Ninguna lógica canónica existe para fijar el estatus de los resultados: la lógica interna del hexágono lógico-semiótico no es necesariamente booleana sino dependiente del modelo descrito por sus categorías (Servín, s.f.-b). La interpretación no es, por tanto, suministrada por el sistema. En teoría de modelos, la herramienta principal de la semántica es el concepto de valor de verdad, y sin embargo la referencia —especialmente en el lenguaje natural— es uno de los temas más problemáticos de la semántica (Servín, s.f.-b). Quién fija la referencia de los símbolos del sistema, y contra qué modelo, sigue siendo una decisión humana externa a la maquinaria.

La distinción entre validez y solidez es, con todo, deductiva, y su aplicación a los sistemas de aprendizaje estadístico es una analogía que debe precisarse. Un razonador simbólico deriva sus conclusiones de premisas; una red entrenada no deduce sus salidas, sino que generaliza por semejanza a partir de ejemplos. Para estos sistemas el marco adecuado es el de la *abducción* y la

intuición artificial descrita antes: la salida es una hipótesis plausible que solo se convierte en conocimiento cuando se reconstruye como argumento y se contrasta. Su riesgo característico no es únicamente la premisa falsa (la falta de solidez), sino el *acierto accidental*: un sistema puede producir una respuesta verdadera apoyándose en un rasgo espurio de los datos, y quien la acepta por la fiabilidad general del sistema tiene una creencia verdadera y en apariencia justificada que no es conocimiento —un caso de Gettier (Servín, s.f.-a). De ahí se siguen grados que no deben confundirse: una *salida* se convierte en *evidencia* cuando se conocen su procedencia y el tipo de apoyo que ofrece; la evidencia, en *aceptación justificada* cuando un sujeto responsable la integra en un argumento evaluable; la aceptación, en *conocimiento* cuando es verdadera, no accidental y está corroborada intersubjetivamente; y el conocimiento solo apoya una *decisión* cuando se añaden premisas normativas explícitas y la consideración de los afectados. La corrección de un procedimiento, además, solo puede medirse contra una semántica independiente de él, no contra «lo que el código hace» (Brachman & Levesque, 1984).

La afirmación de que las interpretaciones generadas por la IA son válidas simplemente porque la IA se apoya en métodos científicos comete por tanto un error categorial: sustituye la validez técnica por la validez interpretativa y normativa. Un trabajo reciente distingue tres dimensiones a lo largo de las cuales debe juzgarse la validez de los resultados de la IA —la corrección técnica, la validez interpretativa (si los resultados pueden entenderse y aplicarse de manera significativa) y la validez normativa (la equidad y el respeto por los estándares éticos) (Hakim et al., 2026). La «validez científica» se refiere característicamente solo a la dimensión técnica, y el rigor técnico no purifica el resultado. La literatura sobre IA responsable hace el mismo punto en negativo: las conclusiones de la IA descansan sobre estadística inferencial y aprendizaje automático que producen conocimiento probabilístico e incierto y son, por tanto, falibles; la evidencia puede ser inconcluyente, y los sistemas de caja negra inescrutables rompen la conexión entre los datos usados, el modo en que se usaron y la conclusión generada (Trocin et al., 2021). Cuando la derivación es inescrutable, la solidez no puede verificarse sino solo suponerse. Incluso la «precisión» —el valor que la inferencia

presume neutral y absoluto— conlleva un componente axiológico y pragmático constitutivo, de modo que la cuestión operativa es siempre la precisión *para quién*, y para qué propósito (Russo et al., 2023).

La dimensión ética se sigue directamente. Afirmar que toda interpretación es válida porque el método es científico convierte un resultado demostrable y relativo al modelo en una certificación dirigida al mundo, induciendo a error a usuarios, reguladores y público sobre la confiabilidad de los sistemas. El sesgo y el error en los sistemas de IA no meramente distorsionan resultados individuales; socavan la validez y la confiabilidad del conocimiento científico mismo, con consecuencias potencialmente desastrosas para la salud pública, la seguridad y la política social —y la ética de la investigación científica exige que tales errores se identifiquen, describan y corrijan en lugar de disculparlos por el rigor del método (Resnik & Hosseini, 2024). La misma retórica borra las decisiones humanas —qué datos, qué categorías, qué granularidad, qué valores— que en realidad determinaron el resultado, y disuelve así la asignación rastreable de responsabilidad a lo largo del ciclo de vida que la rendición de cuentas exige (Novelli et al., 2023). Cuando «el sistema lo resolvió correctamente» se afirma sin calificaciones, los diseñadores, implementadores y usuarios que eligieron las premisas quedan relevados de la rendición de cuentas, reproduciendo la brecha de responsabilidad en la que un sistema autónomo actúa sin que nadie sea responsable de la acción y sus consecuencias (de Sio & Mecacci, 2021; Königs, 2022). La alternativa defendible es tratar la validez no como una propiedad que un sistema posee de una vez y para siempre, sino como un juicio relativo a propósitos, concedido solo para los usos previstos y mantenido abierto a revisión —una práctica responsable y reflexiva más que un certificado de autoridad científica.