Ángel Espina Barrio (Coord.)

LOS INVIERNOS DE LA MÁQUINA

MEMORIA SOCIAL DE LOS FRACASOS ALGORÍTMICOS

Prólogo de Juan Manuel Corchado Rodríguez
Rector de la Universidad de Salamanca

MEMORIAS MÍNIMAS DEL BARRO Y EL SILICIO
TOMO V

VICTOR AVENDAÑO PORRAS
ÁNGEL ESPINA BARRIO
IRIS ALFONZO ALBORES

INSTITUTO
HISTÓRICO E GEOGRÁFICO
DE SANTA CATARINA
FUNDADO EM 07 DE SETEMBRO DE 1896

MEMORIAS MÍNIMAS DEL BARRO Y EL SILICIO
TOMO V

LOS INVIERNOS DE LA MÁQUINA

MEMORIA SOCIAL DE LOS FRACASOS ALGORÍTMICOS

Ángel Espina Barrio (Coord.)

VICTOR AVENDAÑO
ÁNGEL ESPINA BARRIO
IRIS ALFONZO ALBORES

INSTITUTO
HISTÓRICO E GEOGRÁFICO
DE SANTA CATARINA
FUNDADO EM 07 DE SETEMBRO DE 1896

Víctor del Carmen Avendaño Porras; Ángel Espina Barrio e Iris Alfonzo Albores. Los inviernos de la máquina/ coordinación de Ángel Espina Barrio. – 1a ed. – Santa Catarina, Brasil: Instituto Histórico e Geográfico de Santa Catarina, 2025. (Colección: Memorias mínimas del barro y el silicio, tomo V)
Libro digital, PDF

Archivo digital: descarga y online
ISBN: 978-65-989177-2-2

1. Antología. 2. Inteligencia artificial—Aspectos culturales. 3. Tecnología—Historia y crítica.

Otros descriptores asignados por la editorial
Inteligencia artificial / Mitos / Cultura digital / Historia de la tecnología / Antropología de la tecnología / América Latina / Pluriverso / Transiciones tecnoculturales

MEMORIAS MÍNIMAS DEL BARRO Y EL SILICIO
TOMO V

LOS INVIERNOS DE LA MÁQUINA

MEMORIA SOCIAL DE LOS FRACASOS ALGORÍTMICOS

Ángel Espina Barrio (Coord.)

VICTOR AVENDAÑO
ÁNGEL ESPINA BARRIO
IRIS ALFONZO ALBORES

INSTITUTO HISTÓRICO E GEOGRÁFICO DE SANTA CATARINA
FUNDADO EM 07 DE SETEMBRO DE 1896

CÁTEDRA DE ANTROPOLOGÍA SOCIAL Y EDUCACIÓN TRANSFORMADORA 'Ángel Baldomero Espina Barrio'

ILED
INNOVATIO EX SCIENTIA

VNiVERSiDAD D SALAMANCA

INSTITUTO
HISTÓRICO E GEOGRÁFICO
DE SANTA CATARINA
FUNDADO EM 07 DE SETEMBRO DE 1896

Instituto Histórico e Geográfico de Santa Catarina

Ángel Espina Barrio – Coordinador General de la Colección
Víctor del Carmen Avendaño Porras – Autor y Editor académico
Iris Alfonzo Albores– Autora

Equipo Editorial
Corrección y revisión de estilo: Ana Carolina Moraes
Diagramación y diseño: Daniela Freitas / Marcos Tavares
Diseño de tapa: Rodrigo Silva
Traducción y adaptación: Equipo Cefeo

CONOCIMIENTO ABIERTO, CONOCIMIENTO LIBRE
Los libros de la colección Memorias mínimas del barro y el silicio pueden descargarse libremente en formato digital o adquirirse en versión impresa desde cualquier lugar del mundo ingresando a:
https://www.ihgsc.org/

Avendaño Porras, Víctor del Carmen; Espina Barrio, Ángel y Alfonzo Albores, Iris.
Los inviernos de la máquina: Memoria social de los fracasos algorítmicos
(Coordinación de Ángel Espina Barrio).

Santa Catarina, Brasil: Instituto Histórico e Geográfico de Santa Catarina, 2025.
(Colección Memorias mínimas del barro y el silicio, tomo V)

ISBN: 978-65-989177-2-2

Instituto Histórico e Geográfico de Santa Catarina
Casa José Boiteux - Av. Hercílio Luz 523 - Centro Florianópolis.
Santa Catarina - Brasil.

Este material/producción forma parte de la colección Memorias mínimas del barro y el silicio, una iniciativa editorial del Instituto Histórico e Geográfico de Santa Catarina en colaboración con Cefeo.

Prólogo

Hay libros que no se limitan a describir, sino que interrogan. Que no solo explican el mundo, sino que invitan a pensarlo de nuevo. La colección Memorias mínimas del barro y el silicio pertenece a esa rara estirpe. Ocho volúmenes breves, pero de largo aliento, que recorren las huellas profundas del deseo humano por crear inteligencia. No la inteligencia abstracta de los tratados técnicos, sino la que brota del barro y del verbo, del ritual y del engranaje, del mito y del cálculo.

El primer volumen, Autómatas y espíritus, nos conduce a los orígenes simbólicos del artificio: desde golems de barro y talismanes animados hasta la certeza de que la materia, bajo ciertas manos y ciertas palabras, puede despertar. Allí donde se esculpe un cuerpo, late una pregunta sobre el

alma. Le sigue De engranajes y fórmulas, que narra cómo el pensamiento mecánico emergió no de una ruptura con lo sagrado, sino de la lenta traducción del asombro en medida. El mundo se reveló, no menos poético, al descubrirse repetible. La guerra y el código escarba los archivos del siglo XX para mostrarnos que la inteligencia artificial no nació en un laboratorio, sino en la urgencia de la guerra, donde descifrar un mensaje podía salvar mil vidas. El código se volvió sistema, el secreto, forma de saber.

En La tribu de Dartmouth, asistimos al nacimiento de una comunidad. No un grupo de técnicos aislados, sino una cofradía con sus rituales, lenguajes y herencias. De aquel encuentro de 1956 en New Hampshire surgió algo más que una disciplina: una visión compartida del futuro. Tuve el privilegio, décadas más tarde, de encontrarme con algunos de sus pioneros, cuando, becado por la AAAI y el IEEE, asistí a congresos donde aquella intuición inicial se había transformado en ciencia. Sus palabras, su generosidad intelectual y su lucidez me guiaron en mis primeros pasos por un territorio aún incierto. En aquellos años, el invierno de la IA se sentía largo y frío, pero la memoria de los que soñaron antes que nosotros hacía menos áspero el camino. Quizá por eso, este libro no se limita a narrar un origen: lo revive. Y al hacerlo, recuerda que toda comunidad científica nace también del afecto, de la transmisión callada, del gesto compartido de creer que pensar, aun en soledad, tiene sentido.

El invierno llega con Los inviernos de la máquina. No como metáfora romántica, sino como la memoria concreta de cada fracaso, de cada entusiasmo que se quebró ante el límite técnico, el mercado o la simple condición humana. Aquel frío, que conocí en carne propia, no paraliza: ense-

ña. En la pausa impuesta por la desilusión, uno aprende a distinguir lo necesario de lo urgente, lo que permanece de lo que deslumbra. Este volumen rescata esas lecciones calladas, las que no suelen figurar en los congresos ni en los titulares, y les da forma, como si cada tropiezo fuera también una forma de avanzar con más verdad.

Jugadas imposibles nos sumerge en los tableros donde humanos y máquinas midieron su ingenio. No para decidir un vencedor, sino para entender que la creatividad no es patrimonio exclusivo de uno u otro. Cuando una máquina juega lo impensado, el espejo se quiebra, y vemos algo nuevo. Las Culturas de la red delinean el nuevo paisaje. Algoritmos que deciden qué leer, con quién hablar, qué creer. La red ya no es herramienta; es entorno. Y su lenguaje, feeds, métricas, optimización, se filtra en nuestras instituciones, nuestras emociones, nuestros gestos cotidianos. Por último, Chatbots y voces sintéticas pone el oído en la revolución más íntima: la palabra. Máquinas que no solo nos obedecen, sino que nos

escuchan y responden. No es fácil distinguir si conversan o simulan, pero la frontera entre presencia y artificio ya no está donde solía estar.

En esta colección resuenan, de otro modo, preguntas que ya habitaban en la obra de los doctores salmantinos. ¿Qué es el entendimiento humano? ¿Cómo se forma la voluntad? ¿Dónde empieza la libertad? Así como aquellos pensadores, Francisco de Vitoria, Domingo de Soto, Luis de León, Francisco Suárez, trataron de comprender al ser humano en su apertura a Dios, a la ley y al otro, estos libros exploran el reverso moderno: ¿cómo pensar al ser humano cuando ha creado una inteligencia que le imita? La Escuela de Salamanca inauguró una forma de pensamiento que, sin

desligarse de su tiempo, aspiraba a principios universales; que analizaba el comercio, la guerra, el poder, la conciencia, no desde la utilidad sino desde la justicia. Esa misma exigencia de rigor moral y claridad conceptual que animó la cátedra salmantina se despliega aquí, no como doctrina, sino como horizonte que obliga. Leer estas obras es, en cierto modo, continuar ese diálogo. No con la teología, sino con la técnica. No con el alma, sino con su reflejo algorítmico. Pero la pregunta sigue siendo la misma: ¿qué significa ser humano en un mundo que nos excede?

La Universidad de Salamanca, en su doble vocación humanista y científica, sabe que las máquinas no se piensan solas. Como nos enseñaron Vitoria, Soto y Suárez, toda construcción técnica lleva implícita una visión del ser humano y de su lugar en el mundo. La inteligencia artificial no escapa a esa regla: es, antes que una herramienta, una antropología encarnada. Por eso estas memorias mínimas no son notas al pie de una tecnología futura, sino capítulos de una ética en curso. No es un catálogo lo que se ofrece al lector, sino un mapa. No una respuesta, sino un conjunto de claves para orientarse en una época donde la inteligencia ya no es sólo biológica, ni la imaginación patrimonio de lo humano. Estos libros invitan a pensar la IA no como una amenaza o una solución, sino como una continuidad de gestos milenarios: animar, ordenar, prever, hablar, fracasar, volver a intentar.

En tiempos donde las máquinas aprenden a conversar, escribir, decidir; donde los algoritmos no sólo procesan datos, sino que modelan mundos; resulta urgente una reflexión que no se quede en lo técnico ni se pierda en la alarma. Esta colección es esa pausa lúcida. Un intento, bello y necesario, de comprender qué decimos realmente cuando

hablamos de inteligencia artificial. Y también, por qué no, de preguntarnos qué decimos cuando hablamos de nosotros. Frente al vértigo que producen los cambios tecnológicos, la Escuela de Salamanca nos legó una virtud olvidada: la paciencia intelectual. El coraje de pensar despacio. De detenerse en los matices, en las consecuencias remotas, en las preguntas sin respuesta inmediata. Esta colección participa de ese espíritu. No teme al artificio, pero tampoco lo idolatra. Pregunta, compara, recuerda. En tiempos donde todo parece acelerarse, leer estos libros es una forma de volver a mirar con profundidad aquello que estamos construyendo sin darnos cuenta: una nueva imagen del ser humano.

El lector que se adentre en estas páginas no solo encontrará ideas, sino un espejo. Y acaso, una brújula.

Juan Manuel Corchado Rodríguez
Rector de la Universidad de Salamanca

Introducción

El "invierno de la inteligencia artificial" fue un periodo en el que las máquinas "inteligentes" que parecían tan prometedoras dejaron de cumplir las expectativas y se quedaron sin el combustible que las mantenía en marcha. El término nació en los años 80 dentro de la comunidad técnica, cuando investigadores que habían vivido recortes masivos de fondos comenzaron a comparar esos tiempos duros con las estaciones más frías del año. No era solo una metáfora poética: describía una realidad cruda donde laboratorios cerraban, proyectos se cancelaban y el optimismo se congelaba hasta volverse escepticismo.

Los inviernos de la IA siguen un patrón tan predecible como las estaciones climáticas. Primero llega la primavera del entusiasmo: alguien demuestra que una máquina puede hacer algo que antes solo hacían los humanos. Luego viene el verano de la inversión masiva: gobiernos, empresas y

universidades abren la billetera esperando una revolución inmediata. Durante este tiempo cálido, las promesas se inflan como globos: "en cinco años tendremos traducción perfecta", "las máquinas serán médicos mejores que los humanos", "los robots harán todo el trabajo pesado". Pero llega el otoño de la realidad: los sistemas fallan en situaciones imprevistas, los costos se disparan, los plazos se incumplen. Finalmente, el invierno de los recortes: inversores decepcionados retiran fondos, medios de comunicación hablan de fracaso, y la investigación se contrae durante años. Es como lanzar un restaurante prometiendo que serás el mejor del mundo sin haber probado siquiera las recetas.

¿Por qué se repite este ciclo? No es culpa de la tecnología, sino de cómo los humanos la gestionamos. La ambición es una fuerza poderosa: queremos resolver problemas grandes y complejos, y cuando vemos una herramienta nueva, imaginamos que puede hacerlo todo de inmediato. Los investigadores sienten presión para conseguir fondos y apoyo, así que a veces presentan sus trabajos de forma más optimista de lo que la realidad permite. Los inversores y políticos necesitan resultados rápidos para justificar el dinero gastado. Los periodistas buscan historias emocionantes que capturen la atención del público. Y todos nosotros, como sociedad, tenemos hambre de soluciones mágicas a problemas difíciles. Es comprensible, pero crea un cóctel peligroso: expectativas infladas que chocan inevitablemente contra límites técnicos, económicos y humanos que nadie quiso ver durante los días de euforia.

Sin embargo, los inviernos de la IA no son apocalipsis. Durante estos períodos fríos, la vida no se detiene: simplemente cambia de ritmo y estrategia. Pequeños equipos de

investigadores siguen trabajando con presupuestos ajustados, pero a menudo con mayor libertad para explorar ideas arriesgadas sin la presión de entregar resultados comerciales inmediatos. Las universidades continúan formando estudiantes, acumulando conocimiento técnico que será valioso cuando llegue la próxima primavera. Los fracasos se estudian con calma, identificando qué funcionó y qué no. Se consolidan lecciones importantes sobre límites, costos y mejores prácticas. Es como un agricultor que usa el invierno para reparar herramientas, planificar la próxima siembra y aprender de los errores de la cosecha anterior. El invierno no es muerte; es una pausa necesaria para crecer de forma más sensata y sostenible.

En este tomo introducimos un concepto clave: la "memoria de invierno". Es el conjunto de registros, análisis y prácticas que documentan qué salió mal durante los períodos de recortes y cómo se puede evitar repetir los mismos errores. Incluye informes técnicos que explican por qué fallaron ciertos sistemas, testimonios de investigadores que vivieron esos tiempos, análisis económicos de por qué se perdió tanto dinero, y propuestas de mejores formas de comunicar avances científicos al público. La memoria de invierno es como el manual de mantenimiento de una máquina compleja: te dice qué piezas se rompen más frecuentemente, bajo qué condiciones, y cómo detectar problemas antes de que se vuelvan catastróficos. Es memoria colectiva que nos ayuda a tropezar menos veces con la misma piedra, o al menos a caer de forma menos dolorosa cuando el tropiezo sea inevitable.

Mirar los fracasos no es masoquismo intelectual: es una forma inteligente de cuidado preventivo. En otras disciplinas esto se entiende bien. La ingeniería civil estudia minu-

ciosamente cada puente que se cae para diseñar estructuras más seguras. La medicina organiza comités que analizan errores quirúrgicos para evitar que se repitan. La aviación ha convertido el análisis de accidentes en una ciencia precisa que ha hecho que volar sea extraordinariamente seguro. Cada sector maduro tiene sus "libros negros" donde se registran fallas y se extraen lecciones. La inteligencia artificial, como campo relativamente joven, todavía está aprendiendo esta disciplina. Fallar no es vergonzoso; lo vergonzoso es fallar de la misma manera una y otra vez sin aprender nada. Los inviernos de la IA contienen información valiosa sobre límites humanos y técnicos que podemos usar para construir mejor el futuro.

La historia de la inteligencia artificial ha vivido al menos dos inviernos completos y está navegando los riesgos de un tercero. El primer gran frío llegó en los años 70, después de que prometedoras técnicas de los 50 y 60 no lograran cumplir expectativas infladas sobre traducción automática y resolución de problemas complejos. El segundo invierno azotó los 90, cuando los "sistemas expertos" que prometían reemplazar el conocimiento humano resultaron demasiado frágiles y costosos de mantener. Entre estos períodos hubo avances significativos, pero lentos y menos publicitados. Desde mediados de los 2000, el "aprendizaje profundo" ha generado un nuevo verano de inversión y expectativas. Hoy vemos logros impresionantes, pero también señales de alarma: sistemas que fallan de maneras impredecibles, costos energéticos enormes, sesgos que perpetúan injusticias. Cada invierno ha enseñado algo distinto: el primero sobre límites técnicos, el segundo sobre gestión de expectativas y costos, el tercero (si llega) podría enseñarnos sobre responsabilidad social y sostenibilidad.

Un ejemplo clásico es el informe ALPAC de 1966, que congeló la investigación en traducción automática durante más de una década. Durante los años 50 y principios de los 60, gobiernos de Estados Unidos y la Unión Soviética invirtieron millones en crear máquinas que tradujeran idiomas instantáneamente. Las promesas eran grandiosas: comunicación global sin barreras, espionaje facilitado, diplomacia más fluida. Pero las primeras máquinas producían textos confusos, a menudo cómicos. La famosa frase "el espíritu está dispuesto, pero la carne es débil" supuestamente se tradujo al ruso y de vuelta al inglés como "el vodka es bueno, pero la carne está podrida". El comité ALPAC concluyó que la traducción automática era más cara y menos precisa que usar traductores humanos profesionales. Los fondos se cortaron drásticamente. La lección central: medir bien la calidad antes de hacer promesas públicas, y comparar siempre el costo de la solución tecnológica con alternativas humanas existentes.

El informe Lighthill de 1973 marcó el fin de la inocencia británica sobre la inteligencia artificial. Sir James Lighthill, matemático respetado, fue encargado por el gobierno del Reino Unido de evaluar el progreso en IA después de años de inversión generosa. Su veredicto fue devastador: criticó lo que llamó "jugueteo" sin resultados prácticos medibles. Según Lighthill, los investigadores estaban resolviendo problemas artificiales en ambientes controlados, pero sus sistemas colapsaban cuando enfrentaban la complejidad del mundo real. El informe llevó a recortes masivos: universidades perdieron fondos, laboratorios cerraron, carreras académicas se truncaron. Lighthill no negaba el valor de la investigación exploratoria, pero exigía honestidad sobre tiempos y aplicabilidad. Su lección sigue vigente: hay

que separar claramente entre investigación básica (que explora posibilidades a largo plazo) y desarrollo aplicado (que promete soluciones concretas en plazos específicos). Mezclar ambos en la comunicación pública genera expectativas imposibles de cumplir.

Los sistemas expertos vivieron su propio ciclo dramático entre los años 80 y 90. Estas herramientas prometían capturar el conocimiento de especialistas humanos (médicos, ingenieros, abogados) y ponerlo a disposición de cualquiera a través de una computadora. Durante los 80, empresas como Teknowledge y IntelliCorp recibieron millones en inversión. Gobiernos y corporaciones compraron sistemas que diagnosticaban enfermedades, diseñaban circuitos o evaluaban riesgos financieros. Algunos funcionaron bien en nichos muy específicos, pero la mayoría resultó frágil: un pequeño cambio en las condiciones y el sistema daba respuestas absurdas. Además, mantenerlos costaba fortunas: cada vez que un experto humano aprendía algo nuevo, había que reprogramar laboriosamente las reglas del sistema. Muchas empresas del sector quebraron en los 90. La lección fue clara: el conocimiento humano es más fluido y adaptable de lo que parece, y codificarlo a mano en reglas rígidas es un trabajo de Sísifo.

Los equipos que sobrevivieron a los inviernos desarrollaron prácticas de supervivencia que siguen siendo válidas hoy. Aprendieron a hacer pruebas pequeñas antes de promesas grandes: si el sistema funciona bien traduciendo menús de restaurante, tal vez pueda intentar textos técnicos, pero no prometas literatura compleja desde el primer día. Adoptaron comunicación honesta sobre límites: "esto funciona el 85% de las veces en condiciones controladas, pero necesita supervisión humana". Desarrollaron presu-

puestos realistas que incluían no solo investigación inicial, sino mantenimiento a largo plazo, corrección de errores y formación de usuarios. Crearon métricas claras para medir progreso: en lugar de proclamar "inteligencia general", medían tareas específicas como "precisión en diagnóstico de neumonia en radiografías de adultos sanos". Estas prácticas no son glamorosas, pero construyen confianza sostenible entre investigadores, inversores y usuarios.

Desde mediados de los 2000, el "aprendizaje profundo" ha generado avances impresionantes: sistemas que reconocen imágenes mejor que humanos, que traducen con fluidez sorprendente, que generan textos convincentes. Pero también vemos señales de riesgo que recuerdan a inviernos anteriores. Los sistemas más potentes consumen energía equivalente a ciudades pequeñas. Los datos necesarios para entrenarlos a menudo se obtienen sin consentimiento claro de los usuarios. Los modelos más grandes son tan complejos que ni sus creadores entienden completamente cómo toman decisiones. Producen "alucinaciones": respuestas que suenan convincentes pero son completamente falsas. Amplifican sesgos presentes en sus datos de entrenamiento, perpetuando injusticias sociales. Los costos de desarrollo se han vuelto astronómicos: solo unas pocas empresas gigantes pueden permitirse entrenar los modelos más avanzados. ¿Estamos ante un nuevo invierno? Depende de cómo gestionemos estos riesgos en los próximos años.

Afortunadamente, hoy tenemos herramientas sofisticadas para prevenir desastres que no existían en inviernos anteriores. El "red-teaming" consiste en formar equipos especializados cuyo trabajo es intentar romper o engañar sistemas de IA para descubrir vulnerabilidades antes de que

causen daño real. Las auditorías algorítmicas revisan sistemas desplegados para detectar sesgos, errores o comportamientos inesperados. Los "informes de incidentes" documentan fallas de manera sistemática, como los informes de accidentes en aviación. Algunos laboratorios mantienen "libros de fallos" donde registran experimentos que no funcionaron y por qué, creando memoria institucional valiosa. Organizaciones como Partnership on AI o AI Now Institute estudian impactos sociales y proponen mejores prácticas. Estas son formas modernas de "memoria de invierno": sistemas organizados para capturar lecciones y prevenir errores repetidos.

Este tomo está escrito para cualquier persona curiosa sobre cómo funciona realmente la inteligencia artificial más allá de los titulares. No hace falta ser ingeniero, matemático o programador para entender estas historias. Queremos que estudiantes, periodistas, funcionarios públicos, empresarios y usuarios preocupados puedan comprender qué ha pasado históricamente y qué podemos hacer hoy para evitar repetir errores costosos. Los inviernos de la IA no son solo anécdotas técnicas: son historias sobre ambición humana, gestión de recursos públicos, comunicación entre ciencia y sociedad, y responsabilidad sobre herramientas poderosas. Cada ciudadano tiene derecho a entender estas tecnologías que afectan su trabajo, privacidad y futuro. Y cada persona que toma decisiones sobre inversión, regulación o uso de IA necesita conocer los patrones históricos para navegar mejor los riesgos y oportunidades actuales.

Primeras heladas

Como ya se vio en el tomo anterior el verano de 1956, un grupo de investigadores se reunió en el Dartmouth College de New Hampshire con una ambición audaz: organizar un taller de dos meses para explorar si las máquinas podían "simular cada aspecto del aprendizaje o cualquier otra característica de la inteligencia". John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon redactaron la propuesta que daría nombre al campo: "inteligencia artificial". El optimismo era contagioso. McCarthy escribió que esperaban avances significativos en tan solo un verano. Los participantes imaginaban máquinas que jugarían ajedrez a nivel de maestros, comprenderían lenguaje natural, resolverían teoremas matemáticos complejos y aprenderían de la experiencia como lo hacen los niños. Era una primavera de esperanzas enormes, plantada en el suelo fértil de las primeras computado-

ras digitales y el entusiasmo de posguerra por la ciencia y la tecnología.

Los primeros años trajeron logros que alimentaron ese entusiasmo. En 1952, Arthur Samuel creó un programa que jugaba a las damas y mejoraba con la práctica, demostrando que las máquinas podían "aprender" en algún sentido. Allen Newell y Herbert Simon desarrollaron el "Logic Theorist", un programa que demostró teoremas de lógica matemática de manera que sorprendió a los propios matemáticos. En 1958, John McCarthy inventó el lenguaje de programación LISP, que se convertiría en la lengua materna de la IA durante décadas. Frank Rosenblatt presentó el "perceptrón", un dispositivo que podía reconocer patrones visuales simples y que prometía ser el embrión de cerebros artificiales. Cada logro generaba titulares emocionantes: "Las máquinas que piensan", "Cerebros electrónicos", "El futuro ya está aquí". Gobiernos y universidades abrieron líneas de financiación generosas. Todo parecía posible.

La traducción automática fue uno de los sueños más tentadores y, finalmente, uno de los fracasos más aleccionadores. Durante la Guerra Fría, Estados Unidos necesitaba desesperadamente traducir documentos científicos y militares soviéticos. La promesa era simple: si las computadoras podían hacer cálculos complejos, seguramente podían buscar palabras en un diccionario y reorganizarlas según reglas gramaticales. En 1954, un equipo de IBM y Georgetown University presentó una demostración que traducía sesenta frases del ruso al inglés. La prensa lo proclamó como el inicio de una revolución. El gobierno estadounidense invirtió millones de dólares en varios proyectos de traducción automática. Investigadores prometieron sistemas funcionales en tres a cinco años. Parecía un problema técnico senci-

llo: vocabulario más gramática igual a traducción. Nadie imaginaba la complejidad oculta del lenguaje humano.

Pero el lenguaje resultó ser una bestia más compleja de lo esperado. Las palabras tienen múltiples significados que dependen del contexto. "Banco" puede ser una institución financiera o un asiento en el parque. "Volar" puede significar desplazarse por el aire o moverse muy rápido. La gramática tiene excepciones laberínticas. Los humanos usamos referencias implícitas, sarcasmo, metáforas, conocimiento cultural compartido. Los primeros sistemas de traducción producían resultados que oscilaban entre lo incomprensible y lo cómico. Se hicieron famosas anécdotas como la frase bíblica "el espíritu está dispuesto, pero la carne es débil", que supuestamente se tradujo como "el vodka es bueno, pero la carne está podrida". Aunque muchas de estas historias fueron exageradas o inventadas, capturaban una verdad incómoda: las máquinas no entendían realmente nada de lo que traducían, solo manipulaban símbolos siguiendo reglas rígidas.

En 1964, el gobierno estadounidense decidió evaluar seriamente si valía la pena seguir invirtiendo en traducción automática. Formó el Comité Asesor sobre Procesamiento Automático del Lenguaje, conocido por sus siglas en inglés: ALPAC (Automatic Language Processing Advisory Committee). El comité estaba compuesto por lingüistas, matemáticos e ingenieros respetados. Durante dos años estudiaron el estado del arte: visitaron laboratorios, probaron sistemas, compararon costos y calidad. En 1966 publicaron su informe, y fue demoledor. Concluyeron que la traducción automática era más lenta, menos precisa y más costosa que contratar traductores humanos profesionales. No veían progreso significativo en el horizonte cercano. Recomenda-

ron redirigir fondos hacia investigación lingüística fundamental en lugar de aplicaciones prematuras. Fue como si un médico te dijera que tu tratamiento costoso no funciona y deberías intentar algo completamente distinto.

El impacto del informe ALPAC fue inmediato y severo. Los fondos federales para traducción automática se cortaron drásticamente. Proyectos que habían empleado decenas de investigadores durante años se cancelaron. Estudiantes de posgrado que preparaban tesis sobre el tema tuvieron que cambiar de área. Algunas universidades cerraron laboratorios completos. La confianza pública en la "inteligencia de las máquinas" sufrió un golpe duro. Durante más de una década, la traducción automática fue considerada un callejón sin salida, un ejemplo de promesas tecnológicas infladas. Solo en los años 80, con nuevas aproximaciones basadas en estadística en lugar de reglas gramaticales codificadas a mano, el campo comenzó a recuperarse lentamente. La lección fundamental del episodio ALPAC fue doble: primero, medir cuidadosamente la calidad antes de hacer promesas públicas; segundo, comparar siempre el desempeño de la solución tecnológica con alternativas humanas existentes, incluyendo costos totales realistas.

Mientras Estados Unidos sufría el golpe de ALPAC, el Reino Unido preparaba su propia evaluación crítica de la inteligencia artificial. En 1972, el Consejo de Investigación Científica británico encargó a Sir James Lighthill, un matemático de gran prestigio, que revisara el estado de la IA en el país después de años de financiación generosa. Lighthill no era un enemigo de la ciencia ni un tecnófobo: era un experto en matemática aplicada con reputación impecable. Visitó los principales laboratorios de IA en universidades británicas, habló con investigadores líderes, estudió publi-

caciones y evaluó resultados concretos. Su informe, publicado en 1973, fue cortés en la forma pero devastador en el fondo. Criticó lo que llamó "inteligencia artificial combinatoria" (la aproximación dominante en esa época, que intentaba resolver problemas buscando sistemáticamente entre posibles soluciones) por ser demasiado lenta y frágil cuando se enfrentaba a problemas del mundo real.

Lighthill argumentó que los investigadores de IA estaban resolviendo "problemas juguete" en ambientes artificiales muy controlados, pero sus sistemas colapsaban cuando enfrentaban la complejidad desordenada del mundo real. Un programa podía jugar ajedrez porque las reglas son fijas y el tablero es pequeño, pero no podía reconocer una taza de café en una cocina real porque las tazas vienen en mil formas, colores y contextos diferentes. Lighthill no negaba el valor de la investigación exploratoria a largo plazo, pero criticaba duramente la confusión entre investigación básica y promesas de aplicaciones prácticas inminentes. Según él, los investigadores habían sido poco honestos con financiadores y público sobre los tiempos reales necesarios para lograr sistemas útiles. Su recomendación fue clara: reducir fondos para proyectos de IA grandilocuentes y redirigir recursos hacia investigación más modesta y honesta en áreas como robótica y procesamiento de señales.

Las consecuencias del informe Lighthill fueron brutales para la comunidad británica de IA. El gobierno cortó fondos drásticamente. Universidades que habían sido líderes mundiales en el campo perdieron posiciones académicas. Jóvenes investigadores brillantes abandonaron la IA para trabajar en otros temas. Algunos emigraron a Estados Unidos buscando mejores oportunidades. La confianza entre la comunidad científica y los financiadores gubernamentales

quedó herida durante años. Donald Michie, uno de los pioneros británicos de la IA y figura central en la Universidad de Edimburgo, criticó amargamente el informe y defendió el valor de la investigación a largo plazo. Pero el daño estaba hecho. El episodio Lighthill enseñó una lección dolorosa pero importante: cuando pides dinero público para investigación, tienes la obligación de comunicar honestamente qué es exploración sin garantías y qué es desarrollo con plazos concretos. Mezclar ambas cosas en la comunicación genera expectativas imposibles.

Otro golpe importante llegó desde el propio corazón de la comunidad de IA: el libro "Perceptrones" de Marvin Minsky y Seymour Papert, publicado en 1969. Los perceptrones eran modelos matemáticos simples inspirados en cómo funcionan las neuronas del cerebro. Frank Rosenblatt había presentado el perceptrón en 1958 con gran fanfarria, prometiendo que era el embrión de máquinas que pensarían como humanos. Durante los años 60, el enfoque "conexionista" (construir inteligencia conectando muchas unidades simples) compitió con el enfoque "simbólico" (construir inteligencia manipulando símbolos lógicos según reglas). Minsky y Papert, figuras dominantes del MIT y defensores del enfoque simbólico, decidieron examinar rigurosamente qué podían y qué no podían hacer los perceptrones. Su análisis matemático demostró que los perceptrones simples tenían limitaciones fundamentales: no podían aprender a resolver ciertos problemas básicos, como determinar si una imagen está compuesta por una sola pieza conectada o varias piezas separadas.

El libro "Perceptrones" fue interpretado por muchos como una sentencia de muerte para el enfoque conexionista. Aunque Minsky y Papert mencionaban que perceptro-

nes más complejos (con múltiples "capas" de unidades) podrían superar algunas limitaciones, enfatizaban que nadie sabía cómo entrenar eficientemente esas redes más complejas. Los fondos para investigación conexionista se secaron. Estudiantes interesados en redes neuronales fueron desalentados por sus profesores. Durante casi dos décadas, el enfoque simbólico dominó completamente el campo. Solo en los años 80, cuando investigadores como Geoffrey Hinton, David Rumelhart y Ronald Williams desarrollaron técnicas eficientes para entrenar redes multicapa (el famoso algoritmo de "retropropagación"), el conexionismo resurgió. El episodio de los perceptrones enseña algo sutil: los argumentos técnicos legítimos sobre limitaciones pueden ser utilizados para cerrar prematuramente líneas de investigación que podrían ser valiosas si se les diera tiempo de madurar. La ciencia necesita crítica rigurosa, pero también necesita paciencia y diversidad de enfoques.

Pero ¿cuáles eran las causas de fondo que explican este primer invierno de la IA? Una causa central era el costo y la debilidad del cómputo disponible. Las computadoras de los años 60 y 70 eran máquinas enormes y lentas que costaban millones de dólares. Una computadora de universidad típica tenía menos poder de procesamiento que un teléfono móvil básico de hoy. Los programas de IA requerían buscar entre millones de posibilidades, pero las máquinas disponibles tardaban horas o días en tareas que ahora tomarían segundos. Los investigadores pasaban más tiempo optimizando código para caber en memorias diminutas que explorando ideas nuevas. Era como intentar construir un rascacielos con herramientas de carpintero del siglo XIX: la ambición era válida, pero las herramientas simplemente no estaban listas. Los pioneros subestimaron masivamente

cuánto poder computacional se necesitaría para hacer que sus ideas funcionaran en práctica.

Otra causa fundamental era la escasez de datos. Los sistemas de IA necesitan aprender de ejemplos. Para entrenar un sistema que reconozca gatos en fotos, necesitas miles de fotos etiquetadas como "gato" o "no gato". Para entrenar un traductor, necesitas millones de frases traducidas por humanos profesionales. En los años 60 y 70, estos conjuntos de datos simplemente no existían. No había internet donde recolectar información masivamente. Digitalizar textos, imágenes o sonidos era un proceso manual, lento y costoso. Los investigadores trabajaban con conjuntos de datos minúsculos: decenas o cientos de ejemplos en lugar de los millones que usan los sistemas modernos. Era como intentar aprender a cocinar comida china habiendo probado solo tres platos. Los pioneros no podían anticipar que décadas después, la explosión de datos digitales sería uno de los factores clave que permitiría el renacimiento de la IA.

Una tercera causa era la falta de objetivos claros y métricas precisas. Muchos proyectos de IA temprana tenían metas vagas como "simular la inteligencia humana" o "resolver problemas complejos". Pero ¿cómo sabes si has logrado eso? ¿Qué significa exactamente "inteligencia"? Sin formas claras de medir progreso, era difícil saber si un proyecto iba por buen camino o estaba perdido. Los investigadores a menudo demostraban sistemas en escenarios cuidadosamente elegidos donde funcionaban bien, pero evitaban probarlos en situaciones más difíciles donde fallarían. Esto no era necesariamente deshonestidad: era falta de cultura de evaluación rigurosa. Hoy, el campo de IA usa "benchmarks" (conjuntos de pruebas estándar) para comparar sistemas objetivamente: precisión en reconocer imágenes,

velocidad en traducir texto, tasa de error en diagnosticar enfermedades. En los años 60 y 70, esta disciplina evaluativa apenas existía, y su ausencia facilitaba el autoengaño.

La mala comunicación entre investigadores, financiadores y público fue otra causa crucial del primer invierno. Los científicos, entusiasmados por sus descubrimientos, a menudo exageraban el significado de resultados preliminares cuando hablaban con periodistas o pedían fondos. Los periodistas, buscando historias emocionantes, amplificaban esas exageraciones con titulares sensacionalistas. Los políticos y administradores, presionados para justificar inversiones públicas, prometían aplicaciones prácticas en plazos irrealistas. Los inversores privados buscaban retornos rápidos y no entendían la diferencia entre un prototipo de laboratorio y un producto comercial robusto. Nadie tenía incentivos claros para comunicar honestamente los límites y las incertidumbres. Esta dinámica creó una burbuja de expectativas infladas que inevitablemente estalló cuando los resultados concretos no llegaron a tiempo. La lección es que comunicar ciencia requiere no solo explicar lo que funciona, sino también lo que no funciona y por qué, junto con estimaciones honestas de tiempo y recursos necesarios.

¿Cómo vivieron este invierno los investigadores en los laboratorios? Para muchos, especialmente los más jóvenes, fue un tiempo de incertidumbre y frustración. Estudiantes de doctorado que habían dedicado años a tesis sobre IA vieron cómo sus temas de investigación perdían prestigio. Profesores junior que habían apostado su carrera a la IA encontraban difícil publicar artículos o conseguir financiación. Algunos abandonaron el campo completamente y se movieron a áreas más "respetables" como sistemas operati-

vos, bases de datos o teoría de algoritmos. Otros persistieron con presupuestos mínimos, trabajando en universidades pequeñas o con fondos de proyectos no directamente etiquetados como "IA". Hubo sensación de injusticia: muchos sentían que su trabajo era sólido y valioso, pero estaban pagando el precio de promesas exageradas hechas por otros. El ambiente académico se volvió más conservador: proponer ideas arriesgadas sobre inteligencia de máquinas era mal visto.

Sin embargo, no todo fue negativo durante estos años fríos. Algunos investigadores usaron la pausa obligada para consolidar fundamentos teóricos que resultarían valiosos más tarde. Se escribieron libros de texto importantes que organizaban el conocimiento acumulado. Se desarrollaron lenguajes de programación y herramientas de software que seguirían usándose durante décadas. Se exploraron aplicaciones modestas y específicas que sí funcionaban: sistemas para diseñar circuitos electrónicos simples, programas que ayudaban a químicos a identificar moléculas, algoritmos de búsqueda que se usaban en planificación logística. Estos logros no generaban titulares emocionantes, pero construían conocimiento práctico. Algunos laboratorios aprendieron a comunicar mejor: dejaron de prometer "inteligencia general" y empezaron a hablar de "herramientas especializadas para problemas específicos". Era un discurso menos glamoroso, pero más honesto y sostenible.

La prensa jugó un papel ambiguo durante el primer invierno. Durante los años de euforia (1956-1966), los periódicos y revistas habían publicado historias entusiastas sobre "cerebros electrónicos" que pronto pensarían como humanos. Estos artículos raramente mencionaban límites o incertidumbres. Cuando los fracasos se hicieron evidentes,

algunos medios publicaron piezas muy críticas, a veces con tono de burla: "Las máquinas tontas", "El futuro que nunca llegó". Estos artículos tampoco eran equilibrados: pintaban todo el campo como un fraude o una fantasía. Pocos periodistas intentaron la difícil tarea de explicar matizadamente qué había funcionado, qué no, por qué, y qué podría lograrse con más tiempo y recursos. Esta cobertura simplista (primero hype exagerado, luego condena total) dificultó que el público entendiera la naturaleza real del progreso científico: lento, incremental, con muchos fracasos productivos en el camino.

Algunos investigadores reflexionaron públicamente sobre las lecciones del primer invierno y propusieron buenas prácticas para evitar repetirlo. Allen Newell y Herbert Simon, dos pioneros respetados, escribieron sobre la importancia de definir problemas claramente y medir progreso rigurosamente. John McCarthy argumentó que la comunidad de IA necesitaba mayor rigor matemático y menos demostraciones espectaculares de corto plazo. Algunos propusieron separar institucionalmente la investigación exploratoria (financiada con paciencia, sin promesas de aplicaciones inmediatas) de la investigación aplicada (con plazos claros y métricas de éxito concretas). Otros enfatizaron la necesidad de colaboración interdisciplinaria: los informáticos solos no podían resolver problemas de inteligencia sin ayuda de psicólogos, lingüistas, neurocientíficos y filósofos. Estas reflexiones no siempre se implementaron ampliamente, pero sembraron ideas que germinarían más tarde.

Una lección práctica importante fue la necesidad de "pruebas de concepto" pequeñas antes de grandes inversiones. En lugar de proponer inmediatamente sistemas que traducirían todos los idiomas o resolverían todos los pro-

blemas, los investigadores aprendieron a proponer: "Demostraremos que podemos traducir manuales técnicos de física del ruso al inglés con 70% de precisión en dos años, usando este presupuesto". Ese tipo de objetivo es verificable: al final de los dos años, puedes medir si se logró o no. Si funciona, puedes proponer expandir a otros tipos de texto. Si no funciona, puedes analizar por qué y ajustar. Este enfoque incremental y medible no es tan emocionante como prometer "inteligencia general", pero construye confianza sostenible. Es como renovar una casa: empiezas por una habitación, verificas que quedó bien, aprendes del proceso, y luego continúas con la siguiente.

Otra práctica valiosa fue documentar no solo éxitos sino también fracasos. Algunos laboratorios comenzaron a mantener registros detallados de experimentos que no funcionaron: qué se intentó, qué salió mal, hipótesis sobre las causas. Esto creaba memoria institucional valiosa: cuando un estudiante nuevo proponía una idea, se podía revisar si ya se había intentado antes, ahorrar tiempo evitando callejones sin salida conocidos, o intentar la idea con modificaciones informadas por intentos anteriores. En ciencia, los experimentos "negativos" (que no confirman la hipótesis) son tan valiosos como los positivos, porque delimitan qué no funciona y por qué. Sin embargo, las revistas académicas preferían publicar resultados positivos, y los investigadores tenían poco incentivo para compartir fracasos. Algunos laboratorios intentaron cambiar esta cultura internamente, aunque el problema persiste hasta hoy.

Los presupuestos también necesitaban volverse más realistas y completos. Durante los años de euforia, muchos proyectos solicitaban fondos solo para investigación inicial: contratar investigadores, comprar computadoras, desarro-

llar algoritmos. Pero no incluían costos de mantenimiento a largo plazo, corrección de errores, actualización cuando cambiaban las necesidades, formación de usuarios, documentación técnica. Cuando estos costos "ocultos" aparecían, los proyectos se quedaban sin dinero a medio camino. Los investigadores que sobrevivieron al invierno aprendieron a presentar presupuestos más honestos que incluían todas las fases: diseño, implementación, prueba, corrección, despliegue, mantenimiento. Esto hacía que los proyectos parecieran más caros inicialmente, pero evitaba sorpresas dolorosas más tarde. Es la diferencia entre comprar un auto considerando solo el precio de compra versus considerar también seguro, combustible, mantenimiento y reparaciones.

La gestión de expectativas con inversores y público se volvió otro foco de atención. Algunos investigadores desarrollaron el hábito de comunicar siempre en tres niveles: primero, qué funciona hoy con confianza; segundo, qué podría funcionar en el futuro cercano (tres a cinco años) con esfuerzo e inversión razonables; tercero, qué es especulación a largo plazo sin garantías. Esta estructura ayudaba a financiadores y periodistas a entender la diferencia entre resultados consolidados, objetivos realistas y sueños distantes. Por ejemplo: "Hoy podemos traducir frases simples con 60% de precisión. En cinco años, con más datos y mejor hardware, esperamos alcanzar 85% en textos técnicos. Algún día quizás logremos traducir literatura preservando estilo y sutileza, pero no sabemos cuándo ni si es posible". Ese tipo de comunicación matizada no genera titulares espectaculares, pero construye confianza duradera.

Las métricas de evaluación se volvieron más sofisticadas y honestas. En lugar de demostrar sistemas en los pocos

casos donde funcionaban bien, se empezó a usar conjuntos de prueba diversos y desafiantes. Se desarrolló el concepto de "validación cruzada": dividir tus datos en dos grupos, entrenar el sistema con uno y probarlo con el otro (datos que el sistema nunca vio durante entrenamiento). Esto previene el autoengaño: un sistema puede memorizar ejemplos de entrenamiento sin realmente aprender patrones generales. También se comenzó a comparar sistemas de IA no solo entre sí, sino contra alternativas humanas y contra la opción de no hacer nada. Por ejemplo, si tu sistema de diagnóstico médico acierta 70% de las veces, ¿cuánto acierta un médico humano promedio? ¿Cuánto acertarías simplemente adivinando la enfermedad más común? Estos puntos de referencia revelan si el sistema realmente aporta valor.

Algunos investigadores también reflexionaron sobre los límites fundamentales de las aproximaciones de esa época. Hubert Dreyfus, filósofo del MIT y crítico famoso de la IA, argumentó en su libro "What Computers Can't Do" (1972) que la inteligencia humana no es principalmente manipulación de símbolos según reglas lógicas, sino una capacidad encarnada que depende de tener un cuerpo, emociones, y estar inmerso en un contexto social y cultural. Dreyfus era polémico y muchos investigadores lo detestaban, pero planteaba preguntas difíciles: ¿Puede una máquina sin cuerpo entender realmente qué es "cansancio" o "hambre"? ¿Puede un sistema de reglas lógicas capturar la intuición que desarrolla un maestro de ajedrez después de miles de partidas? Estas preguntas filosóficas, aunque incómodas, ayudaron a algunos investigadores a reconocer que ciertas aproximaciones tenían límites inherentes y que hacían falta ideas radicalmente nuevas.

Durante el primer invierno también surgió mayor conciencia sobre la importancia de la interdisciplinariedad. Los primeros proyectos de IA habían sido dominados por informáticos y matemáticos. Pero resolver problemas de inteligencia requiere entender cómo funciona realmente la cognición humana. Los psicólogos estudian cómo las personas aprenden, recuerdan y resuelven problemas. Los lingüistas estudian cómo funciona el lenguaje en uso real. Los neurocientíficos estudian cómo el cerebro procesa información. Los filósofos estudian qué significa "conocimiento" o "comprensión". Algunos investigadores de IA comenzaron a colaborar seriamente con expertos de estos campos, en lugar de simplemente asumir que podían resolver todo con matemáticas y programación. Nacieron centros de investigación interdisciplinarios, como el Center for Cognitive Science del MIT, donde informáticos, psicólogos, lingüistas y neurocientíficos trabajaban juntos. Este espíritu colaborativo produciría frutos importantes en las décadas siguientes.

Algunos investigadores empezaron a enfocarse en aplicaciones modestas pero útiles, en lugar de perseguir "inteligencia general". Edward Feigenbaum, químico y científico computacional de Stanford, trabajó en sistemas que ayudaban a identificar la estructura molecular de compuestos químicos a partir de datos de espectrometría de masas. No era glamoroso ni pretendía simular la inteligencia humana completa, pero resolvía un problema real y valioso para químicos profesionales. El sistema DENDRAL, desarrollado entre 1965 y 1977, fue considerado uno de los primeros éxitos genuinos de la IA aplicada. Funcionaba porque el problema era bien definido, los datos eran estructurados, y había expertos humanos disponibles para validar y mejorar

el sistema. Esta línea de trabajo evolucionaría hacia los "sistemas expertos" de los años 80, que vivirían su propio ciclo de auge y caída.

El campo de la robótica también avanzó durante estos años, aunque con expectativas más modestas que otras áreas de IA. Laboratorios como el de MIT y Stanford desarrollaron brazos robóticos que podían apilar bloques, robots móviles que navegaban entornos simples usando cámaras y sensores. Shakey, un robot desarrollado en el Stanford Research Institute entre 1966 y 1972, podía moverse por habitaciones, evitar obstáculos y mover objetos simples. Era lento, torpe y funcionaba solo en entornos cuidadosamente preparados, pero demostraba que integrar percepción, planificación y acción en un sistema físico era posible. Los investigadores de robótica tendían a ser más cautelosos en sus promesas que otros investigadores de IA, quizás porque trabajar con hardware hace que los límites y dificultades sean inmediatamente obvios. Nadie podía fingir que un robot funcionaba bien cuando caía al intentar subir una escalera.

Algunos avances teóricos importantes ocurrieron durante el primer invierno, aunque no generaron aplicaciones inmediatas. En 1974, Paul Werbos, en su tesis doctoral, propuso el algoritmo de "retropropagación" (backpropagation) para entrenar redes neuronales multicapa. Este algoritmo resolvía el problema que Minsky y Papert habían señalado: cómo ajustar las conexiones internas de redes complejas para mejorar su desempeño. Sin embargo, la tesis de Werbos pasó casi desapercibida durante años. Las computadoras de la época eran demasiado lentas para entrenar redes grandes, los datos disponibles eran escasos, y el enfoque conexionista estaba desacreditado. Solo una dé-

cada más tarde, cuando otros investigadores redescubrieron independientemente el algoritmo y demostraron aplicaciones interesantes, la retropropagación se volvió central. Este caso ilustra que ideas valiosas pueden llegar "antes de tiempo" y quedar dormidas hasta que las condiciones (hardware, datos, clima intelectual) sean propicias.

El desarrollo de lenguajes de programación especializados fue otro logro menos visible pero duradero. LISP, creado por John McCarthy en 1958, se volvió el lenguaje estándar para investigación en IA durante décadas. Permitía manipular símbolos y estructuras de datos complejas de forma flexible, facilitando la experimentación rápida con ideas nuevas. PROLOG, desarrollado en Francia en 1972, ofrecía una aproximación basada en lógica formal: describes hechos y reglas, y el sistema deduce conclusiones automáticamente. Aunque estos lenguajes no resolvieron los problemas duros de inteligencia, crearon herramientas que aceleraban la investigación. Es como la diferencia entre construir una casa con herramientas modernas versus con herramientas primitivas: el desafío arquitectónico es el mismo, pero las herramientas mejores te permiten experimentar y corregir más rápido.

Durante estos años fríos también hubo reflexión sobre la relación entre IA y empleo humano. Algunos investigadores y sindicalistas temían que las "máquinas inteligentes" desplazaran masivamente trabajadores. En 1964, un grupo de científicos y activistas escribió la "Triple Revolution", un manifiesto que advertía sobre desempleo tecnológico causado por automatización. Sin embargo, durante el primer invierno, estas preocupaciones parecieron prematuras: las máquinas apenas podían hacer tareas simples, y estaba claro que reemplazar trabajadores humanos en la mayoría de

empleos estaba muy lejos. Aún así, la discusión planteó preguntas importantes: si algún día la IA funciona bien, ¿quién se beneficia económicamente? ¿Cómo se distribuyen ganancias y costos sociales? ¿Qué responsabilidad tienen investigadores y empresas hacia trabajadores afectados? Estas preguntas volverían con fuerza décadas después.

Hubo también reflexión ética sobre qué deberían o no deberían hacer las máquinas "inteligentes". Joseph Weizenbaum, científico del MIT, creó ELIZA en 1966, un programa simple que simulaba conversación psicoterapéutica repitiendo frases del usuario en forma de preguntas. Era un truco técnico básico, pero algunas personas que interactuaban con ELIZA se sentían emocionalmente conectadas y compartían problemas personales profundos, incluso después de saber que era solo un programa. Weizenbaum quedó perturbado por esto. Escribió el libro "Computer Power and Human Reason" (1976) argumentando que hay ciertos roles humanos (terapeuta, juez, maestro) que no deberían delegarse a máquinas, incluso si técnicamente fuera posible, porque requieren empatía genuina, responsabilidad moral y comprensión del contexto humano. Su reflexión abrió un debate ético que continúa hoy: ¿qué tareas son apropiadas para automatizar y cuáles no?

Una lección importante del primer invierno fue la necesidad de humildad intelectual. Los pioneros de la IA eran personas brillantes, pero subestimaron masivamente la complejidad de la inteligencia. En 1958, Herbert Simon declaró que en diez años una computadora sería campeón mundial de ajedrez (esto sucedió realmente, pero en 1997, casi cuarenta años después). En 1965, Simon predijo que en veinte años las máquinas podrían hacer cualquier trabajo que un humano puede hacer (todavía esperamos esto en

2025). Estas predicciones no eran tonterías: se basaban en razonamientos lógicos sobre lo que parecía posible. Pero subestimaban lo difícil que es capturar sentido común, contexto, adaptabilidad y las mil pequeñas habilidades que los humanos usamos sin pensar. La lección es que resolver problemas inteligentes es más difícil de lo que parece, y debemos ser modestos al predecir cuándo lograremos capacidades complejas.

El primer invierno también enseñó que la ciencia necesita tiempo. Algunos de los problemas que parecían imposibles en los años 70 (reconocimiento de voz, de imágenes, traducción) se volvieron viables décadas después, no porque aparecieran ideas radicalmente nuevas, sino porque mejoraron tres factores: hardware más rápido y barato, disponibilidad de datos masivos, y refinamiento paciente de algoritmos. Cada generación de investigadores aprendía de los errores de la anterior y mejoraba gradualmente las técnicas. Pero este progreso incremental es lento y requiere financiación estable a largo plazo. El sistema de fondos públicos y privados, sin embargo, tiende a exigir resultados rápidos. Este desajuste entre los tiempos de la ciencia y los tiempos de la inversión sigue siendo una tensión central en el campo. Una sociedad madura necesita balancear financiación para exploración paciente sin garantías de éxito versus financiación para aplicaciones con plazos claros.

Finalmente, el primer invierno dejó claro que los fracasos técnicos son oportunidades de aprendizaje, no catástrofes morales. Muchos de los proyectos que "fracasaron" en los años 60 y 70 generaron ideas, técnicas, herramientas y conocimiento que resultaron valiosos más tarde. LISP sigue influenciando lenguajes de programación modernos. Los algoritmos de búsqueda desarrollados para juegos se

usan hoy en planificación logística. Las reflexiones sobre representación del conocimiento informan el diseño de bases de datos y sistemas de información. Los estudiantes que se formaron en esos laboratorios "fracasados" llevaron ese conocimiento a otros campos y contextos. En ciencia, el fracaso es información: te dice qué no funciona, por qué, y te ayuda a refinar tu comprensión del problema. Una cultura científica sana celebra tanto los éxitos como los fracasos productivos, y construye memoria colectiva que permita a cada generación pararse sobre los hombros (y aprender de los tropiezos) de la anterior.

Auge y caída de las promesas expertas

A principios de los años 80, la inteligencia artificial resurgió de su primer invierno con una propuesta renovada: si las máquinas no podían tener "inteligencia general", al menos podrían capturar el conocimiento especializado de expertos humanos en campos específicos. Nacieron los "sistemas expertos": programas de computadora diseñados para imitar el razonamiento de médicos, ingenieros, geólogos o abogados en áreas muy concretas. La idea era entrevistar a especialistas humanos, extraer sus reglas de decisión ("si el paciente tiene fiebre alta y tos seca, entonces considerar neumonía"), codificarlas en un programa, y así democratizar el acceso a conocimiento valioso. Edward Feigenbaum, pionero en Stanford, prometió que los sistemas expertos revolucionarían la eco-

nomía: cada empresa podría tener acceso al mejor experto del mundo en cualquier tema, disponible las veinticuatro horas, sin cansancio ni vacaciones. El entusiasmo volvió a crecer, y con él, la inversión.

Los primeros sistemas expertos mostraron resultados prometedores en dominios muy específicos. MYCIN, desarrollado en Stanford entre 1972 y 1980, diagnosticaba infecciones bacterianas de la sangre y recomendaba antibióticos. En pruebas controladas, MYCIN acertaba tanto o más que médicos humanos promedio. DENDRAL, otro sistema de Stanford, ayudaba a químicos a identificar estructuras moleculares a partir de datos de espectrometría de masas. XCON, desarrollado por Digital Equipment Corporation, configuraba órdenes de sistemas informáticos complejos, una tarea que requería expertos humanos altamente especializados. Estos éxitos generaron entusiasmo genuino: si funcionaba para diagnóstico médico y diseño de sistemas, ¿por qué no para derecho, finanzas, ingeniería civil, educación? Empresas especializadas como Teknowledge, IntelliCorp y Carnegie Group recibieron millones en inversión de capital de riesgo. Corporaciones grandes compraron sistemas expertos o contrataron equipos para desarrollar los suyos propios.

El gobierno japonés anunció en 1981 el proyecto de "Quinta Generación de Computadoras", una iniciativa ambiciosa de diez años con cientos de millones de dólares para desarrollar máquinas que razonaran usando lógica y conocimiento, en lugar de simplemente procesar datos. El objetivo era crear computadoras que entendieran lenguaje natural, reconocieran imágenes y sonidos, y tomaran decisiones inteligentes usando bases de conocimiento vastas. El proyecto eligió PROLOG, un lenguaje de programación ba-

sado en lógica formal, como su fundamento técnico. El anuncio causó alarma en Estados Unidos y Europa: si Japón lograba dominar la siguiente generación de computación inteligente, obtendría ventaja económica y estratégica enorme. En respuesta, Estados Unidos lanzó iniciativas como la Microelectronics and Computer Technology Corporation (MCC) y aumentó fondos DARPA para IA. Europa respondió con el programa ESPRIT. Una nueva carrera de inteligencia artificial había comenzado, impulsada por competencia geopolítica.

Durante los años 80, la industria de sistemas expertos creció explosivamente. Empresas especializadas vendían no solo sistemas completos, sino también "shells" (estructuras vacías que permitían a las empresas construir sus propios sistemas expertos sin programar desde cero). Universidades ofrecían cursos y posgrados en ingeniería del conocimiento. Consultores se especializaban en "extraer" conocimiento de expertos humanos y codificarlo en reglas. Revistas técnicas publicaban cientos de artículos sobre nuevas aplicaciones: sistemas que evaluaban riesgo crediticio, que diagnosticaban fallas en maquinaria industrial, que diseñaban circuitos electrónicos, que planificaban rutas de transporte. Para mediados de los 80, parecía que los sistemas expertos eran la aplicación comercial exitosa que la IA había estado buscando durante treinta años. Las expectativas alcanzaron niveles similares a los del periodo previo al primer invierno. Y como antes, las grietas comenzaron a aparecer.

El primer problema serio era la fragilidad del conocimiento codificado. Los sistemas expertos funcionaban bien dentro de los límites estrechos para los que fueron diseñados, pero colapsaban espectacularmente cuando enfrenta-

ban situaciones ligeramente fuera de su dominio. Un sistema experto en diagnóstico de infecciones bacterianas no podía ayudar con infecciones virales. Si le preguntabas sobre un síntoma que no estaba en su base de conocimiento, daba respuestas absurdas o simplemente se rendía. Los expertos humanos, en contraste, pueden razonar por analogía, hacer suposiciones informadas, reconocer cuándo están fuera de su área de especialidad y buscar ayuda. Los sistemas expertos carecían de esta flexibilidad. Eran como estudiantes que memorizaron respuestas para un examen específico pero no entienden realmente el tema: funcionan perfectamente si les haces las preguntas exactas que memorizaron, pero fallan miserablemente con cualquier variación.

El segundo problema era el costo astronómico de mantener y actualizar estos sistemas. Extraer conocimiento de expertos humanos era un proceso lento y doloroso. Los "ingenieros del conocimiento" pasaban meses entrevistando especialistas, observándolos trabajar, tratando de convertir su intuición en reglas explícitas. Pero el conocimiento humano no es estático: los expertos aprenden constantemente de nuevos casos, las mejores prácticas evolucionan, aparecen nuevas tecnologías y descubrimientos. Cada vez que el conocimiento del dominio cambiaba, había que actualizar manualmente todas las reglas relevantes del sistema experto. Esto requería más sesiones con expertos humanos, más programación, más pruebas. Algunos sistemas terminaron requiriendo equipos permanentes de mantenimiento que costaban tanto como simplemente contratar a los expertos humanos que supuestamente estaban siendo reemplazados. Era como tener un auto que requiere

un mecánico dedicado trabajando tiempo completo solo para mantenerlo funcionando.

Un tercer problema era que el conocimiento experto resultó ser mucho más tácito e intuitivo de lo que los investigadores anticipaban. Cuando entrevistas a un médico experimentado sobre cómo diagnostica, puede explicar algunas reglas explícitas: "si hay fiebre y rigidez en el cuello, sospechar meningitis". Pero gran parte de su habilidad es intuición desarrollada después de ver miles de pacientes: reconoce patrones sutiles, nota inconsistencias entre síntomas, siente cuándo algo "no cuadra" incluso si no puede articular exactamente por qué. Los sistemas expertos solo podían capturar la parte explícita del conocimiento, perdiendo la intuición tácita que a menudo es la más valiosa. El filósofo Michael Polanyi había identificado este fenómeno décadas antes: "sabemos más de lo que podemos decir". Los sistemas expertos chocaron dolorosamente contra esta verdad: reducir experiencia humana a reglas explícitas pierde información crucial.

Mientras los sistemas expertos luchaban con estos problemas prácticos, el ambicioso proyecto japonés de Quinta Generación enfrentaba dificultades técnicas fundamentales. PROLOG, el lenguaje elegido como base, resultó menos adecuado de lo esperado para construir sistemas complejos de razonamiento. Las máquinas especializadas diseñadas para ejecutar PROLOG eficientemente eran caras y difíciles de programar. El objetivo de crear computadoras que razonaran con conocimiento de sentido común resultó tan difícil en los 80 como había sido en los 60: nadie sabía cómo codificar el vasto conocimiento implícito que los humanos usan para entender el mundo. El proyecto produjo algunos avances técnicos y formó investigadores talentosos,

pero no logró los objetivos revolucionarios prometidos. Para principios de los 90, estaba claro que la Quinta Generación no daría el salto cualitativo esperado. Japón había invertido cientos de millones sin obtener la ventaja estratégica buscada.

Paralelamente, el enfoque conexionista (redes neuronales) experimentó un renacimiento en los años 80 después de haber sido marginado por la crítica de Minsky y Papert. En 1986, David Rumelhart, Geoffrey Hinton y Ronald Williams popularizaron el algoritmo de "retropropagación" (que Paul Werbos había propuesto en 1974 pero había pasado desapercibido). Este algoritmo permitía entrenar redes neuronales con múltiples capas de manera eficiente, superando las limitaciones de los perceptrones simples. Los investigadores demostraron que estas redes podían aprender patrones complejos a partir de datos, sin necesidad de codificar reglas explícitas a mano. Esto ofrecía una alternativa atractiva a los sistemas expertos: en lugar de entrevistar expertos y codificar su conocimiento, simplemente alimenta la red con miles de ejemplos y déjala aprender los patrones automáticamente. Sin embargo, las redes neuronales tenían sus propios problemas: requerían grandes cantidades de datos, mucho poder computacional, y producían decisiones difíciles de interpretar.

Para finales de los 80, el mercado de sistemas expertos comenzó a contraerse. Muchas empresas que habían comprado sistemas descubrieron que los costos de mantenimiento eran prohibitivos. Los sistemas quedaban obsoletos rápidamente y requerían actualizaciones constantes. Los resultados en el mundo real eran menos impresionantes que en demostraciones controladas. Además, los sistemas expertos requerían hardware especializado (las llamadas

"máquinas LISP") que era muy caro. Cuando aparecieron computadoras personales y estaciones de trabajo potentes y baratas que usaban lenguajes de programación convencionales, el hardware especializado de IA se volvió económicamente insostenible. Empresas que habían invertido millones en sistemas expertos comenzaron a abandonarlos. La burbuja de inversión en IA comercial comenzó a desinflarse. En 1987, el mercado de hardware especializado para IA colapsó, un evento que algunos llaman el "colapso de las máquinas LISP".

El segundo invierno de la IA llegó a principios de los 90. Empresas especializadas como Teknowledge e IntelliCorp vieron sus ingresos caer drásticamente; algunas quebraron. DARPA, la agencia de investigación militar estadounidense que había sido un financiador clave de IA durante décadas, recortó fondos dramáticamente en 1988 bajo presión del Congreso, que cuestionaba los retornos de la inversión. El Strategic Computing Initiative, un programa DARPA que había invertido cientos de millones en aplicaciones de IA para defensa, fue cancelado parcialmente. Universidades que habían construido laboratorios grandes de IA enfrentaron recortes presupuestarios. Estudiantes brillantes evitaban posgrados en IA, prefiriendo áreas más "seguras" como bases de datos, redes de computadoras o interfaces de usuario. La prensa, que había celebrado los sistemas expertos años antes, ahora publicaba artículos sobre "el fracaso de la inteligencia artificial" y "promesas rotas".

¿Qué lecciones dejó el auge y caída de los sistemas expertos? Primera lección: separar claramente prototipos de productos. Muchos sistemas expertos funcionaban bien como prototipos de investigación en entornos controlados, pero no estaban listos para ser productos comerciales ro-

bustos. Los investigadores, presionados por inversores y administradores, exageraron la madurez de la tecnología. Convertir un prototipo que funciona el 80% del tiempo en laboratorio en un producto que funciona el 99.9% del tiempo en condiciones reales requiere esfuerzo, tiempo y dinero exponencialmente mayores. Esta brecha entre prototipo y producto sigue siendo una fuente frecuente de decepciones en tecnología. Las demostraciones impresionantes en conferencias no garantizan sistemas confiables en hospitales, fábricas o bancos. La ingeniería robusta es mucho más difícil y costosa que la investigación exploratoria.

Segunda lección: gobernar expectativas requiere comunicación honesta y diferenciada. Cuando hablas con inversores, medios o público general, es crucial distinguir entre: (1) lo que funciona hoy de manera confiable, (2) lo que podría funcionar en tres a cinco años con inversión y esfuerzo razonables, (3) lo que es investigación exploratoria a largo plazo sin garantías de éxito, y (4) lo que es especulación teórica. Mezclar estos niveles en el discurso público crea confusión y expectativas infladas. Los investigadores de sistemas expertos a menudo hablaban de éxitos de nivel 1 (sistemas que funcionaban en nichos estrechos) como si fueran nivel 2 o 3 (sistemas que pronto resolverían problemas generales). Cuando la realidad se hizo evidente, la credibilidad del campo sufrió. Una comunicación madura reconoce límites explícitamente y actualiza estimaciones cuando nueva evidencia lo requiere.

Tercera lección: diseñar métricas útiles es tan importante como desarrollar tecnología. ¿Cómo mides si un sistema experto es "exitoso"? Precisión (porcentaje de diagnósticos correctos) es importante, pero no suficiente. También importa: ¿qué tan confiable es en casos difíciles o atípicos?

¿Qué tan bien explica sus recomendaciones? ¿Cuánto tiempo toma usarlo comparado con métodos tradicionales? ¿Cuánto cuesta desarrollar, desplegar y mantener? ¿Qué pasa cuando el sistema se equivoca? Durante los 80, muchos sistemas expertos se evaluaban con métricas simples (precisión en casos de prueba) que no capturaban estas dimensiones más complejas. Solo cuando los sistemas se desplegaban en uso real, las limitaciones se volvían evidentes. Hoy sabemos que evaluar tecnología requiere métricas multidimensionales que capturen costos, beneficios, riesgos y contextos de uso realistas.

Cuarta lección: presupuestos que contemplan mantenimiento son esenciales para sostenibilidad. Muchos proyectos de sistemas expertos solicitaban fondos para desarrollo inicial (dos o tres años de investigación y programación), pero no incluían presupuesto para las décadas siguientes de mantenimiento, actualización y soporte. Esto creaba una dinámica perversa: el sistema se construía con fondos generosos, se desplegaba con fanfarria, y luego se dejaba morir lentamente porque no había dinero para mantenerlo vivo. Los usuarios se frustraban cuando el sistema quedaba obsoleto o dejaba de funcionar correctamente. Esta lección aplica a toda tecnología compleja: el costo del ciclo de vida completo (desarrollo + despliegue + mantenimiento + eventual retiro) es mucho mayor que el costo de desarrollo inicial, y debe planificarse desde el principio. Un sistema que no puede mantenerse de forma sostenible es un experimento, no una solución.

Durante el segundo invierno, algunos investigadores persistieron trabajando en problemas específicos con expectativas modestas. Stuart Russell y Peter Norvig escribieron "Artificial Intelligence: A Modern Approach" (1995), un libro

de texto que organizó el conocimiento acumulado del campo y se convirtió en estándar mundial. Judea Pearl desarrolló teoría matemática rigurosa sobre razonamiento probabilístico y causalidad, que se volvería fundamental para aplicaciones futuras. Yann LeCun y otros mejoraron redes neuronales convolucionales para reconocimiento de imágenes, aunque el impacto de este trabajo no sería evidente hasta una década después. Estos investigadores no prometían revoluciones inminentes; simplemente hacían ciencia sólida y paciente. Su trabajo creó los fundamentos teóricos y técnicos que permitirían el renacimiento de la IA en los 2000s. Es un recordatorio de que durante los inviernos, el progreso continúa en forma menos visible pero crucialmente importante.

Paradójicamente, algunos legados importantes de la era de sistemas expertos sobrevivieron y se integraron silenciosamente en tecnología cotidiana. Muchas empresas aprendieron que incluso si los sistemas expertos no podían reemplazar completamente a expertos humanos, podían ser herramientas útiles de apoyo. Sistemas de soporte a decisiones que ayudan (pero no reemplazan) a médicos, ingenieros o analistas financieros se volvieron comunes. Bases de conocimiento y ontologías (formas estructuradas de organizar información sobre dominios complejos) desarrolladas en los 80 siguieron usándose. Motores de inferencia (software que aplica reglas lógicas a datos) se incorporaron en software empresarial. Estos usos modestos no generan titulares, pero crean valor real. La lección es que tecnologías que "fracasan" en cumplir promesas grandiosas pueden aún ser útiles en aplicaciones más humildes y realistas.

Durante los 90, mientras la IA académica atravesaba su invierno, algunas aplicaciones específicas comenzaron a

funcionar bien en silencio. Sistemas de reconocimiento de voz mejoraron gradualmente, usados primero en aplicaciones telefónicas automatizadas. Sistemas de recomendación (como los que sugieren películas o productos) se volvieron valiosos para comercio electrónico. Algoritmos de aprendizaje automático se usaban en detección de fraude con tarjetas de crédito, filtrado de spam en correo electrónico, y optimización de rutas logísticas. Estas aplicaciones eran menos glamorosas que "crear inteligencia general", pero resolvían problemas reales y generaban valor económico. Muchas empresas tecnológicas usaban técnicas de IA sin llamarlas así, evitando el estigma del término. Solo cuando funcionaban bien, se revelaba que "inteligencia artificial" estaba involucrada.

Algunos investigadores reflexionaron críticamente sobre la cultura y estructura institucional de la investigación en IA. Douglas Lenat, creador del proyecto Cyc (un intento masivo de codificar sentido común humano en forma de millones de reglas lógicas), reconoció públicamente que el proyecto era mucho más difícil y lento de lo anticipado. Otros investigadores argumentaron que el campo había sufrido por falta de rigor experimental: muchos sistemas se demostraban en ejemplos cuidadosamente elegidos pero no se evaluaban rigurosamente en condiciones realistas. Se propusieron competencias y benchmarks estándar donde diferentes equipos podían probar sus sistemas en las mismas tareas usando las mismas métricas. Por ejemplo, la competencia TREC para sistemas de búsqueda de información, iniciada en 1992, permitía comparar objetivamente diferentes aproximaciones. Esta cultura de evaluación rigurosa y comparativa ayudaría al campo a madurar.

Un desarrollo crucial pero poco publicitado durante los 90 fue la acumulación silenciosa de datos digitales. Internet comenzó a crecer exponencialmente. Empresas digitalizaban documentos, transacciones, imágenes. Sensores en dispositivos generaban datos continuamente. Esta explosión de datos digitales creaba las condiciones para un renacimiento de técnicas de aprendizaje automático que requerían grandes cantidades de ejemplos. Mientras los sistemas expertos basados en reglas codificadas a mano perdían favor, las aproximaciones basadas en aprendizaje desde datos ganaban viabilidad. A finales de los 90, algunos investigadores comenzaron a demostrar que algoritmos relativamente simples alimentados con enormes cantidades de datos podían superar sistemas complejos basados en conocimiento codificado manualmente. Este cambio de "ingeniería de conocimiento" a "aprendizaje desde datos masivos" transformaría el campo en la década siguiente.

Un momento simbólico llegó en 1997 cuando Deep Blue, un sistema de IBM, venció al campeón mundial de ajedrez Garry Kasparov. Era un logro técnico impresionante: el ajedrez había sido durante décadas un símbolo de inteligencia humana. Pero Deep Blue no era "inteligente" en sentido general: era una máquina especializada que evaluaba millones de posiciones por segundo usando hardware masivamente paralelo. No podía hacer nada excepto jugar ajedrez. Kasparov se quejó de que enfrentaba no un oponente individual sino un equipo entero de programadores e ingenieros que ajustaban el sistema entre partidas. El evento generó cobertura mediática enorme pero también reflexión: ¿qué significa realmente "inteligencia"? ¿Es suficiente con superar humanos en una tarea específica? Deep Blue mostraba tanto el poder de la computación especializada como

las limitaciones de equiparar capacidad en una tarea estrecha con inteligencia general.

A finales de los 90 y principios de los 2000, las condiciones para un nuevo verano comenzaron a alinearse. Primero, el hardware se volvió exponencialmente más potente y barato siguiendo la Ley de Moore. Segundo, internet proporcionó acceso a cantidades masivas de datos digitales. Tercero, algoritmos de aprendizaje automático (especialmente "máquinas de vectores de soporte" y versiones mejoradas de redes neuronales) mostraron resultados impresionantes en competencias académicas. Cuarto, empresas tecnológicas como Google, Microsoft y Yahoo necesitaban desesperadamente IA práctica para búsqueda web, publicidad dirigida y reconocimiento de voz. Quinto, agencias militares y de inteligencia renovaron inversión en IA para análisis de datos masivos y vigilancia. Esta confluencia de necesidad, capacidad técnica y recursos crearía las condiciones para el explosivo crecimiento de la IA en los años 2000s y 2010s, particularmente alrededor del "aprendizaje profundo".

Pero antes de cerrar este capítulo sobre el segundo invierno, vale sintetizar las buenas prácticas que emergieron de esa experiencia dolorosa. Primera: siempre comparar el sistema propuesto contra alternativas realistas (incluyendo no hacer nada, o usar métodos tradicionales más simples). Segunda: evaluar en condiciones representativas del uso real, no solo en casos de prueba idealizados. Tercera: calcular costos totales del ciclo de vida (desarrollo + despliegue + mantenimiento + retiro), no solo desarrollo inicial. Cuarta: comunicar diferenciadamente entre resultados consolidados, objetivos realistas a mediano plazo, y especulación a largo plazo. Quinta: documentar no solo éxitos sino tam-

bién fracasos y limitaciones conocidas. Sexta: construir equipos interdisciplinarios que incluyan expertos del dominio de aplicación, no solo informáticos. Séptima: planificar mantenimiento y actualización desde el inicio, no como reflexión tardía. Estas prácticas no garantizan éxito, pero reducen el riesgo de fracasos costosos y decepciones evitables.

Finalmente, el segundo invierno enseñó humildad sobre lo que significa "capturar conocimiento humano". Los sistemas expertos asumían que el conocimiento es principalmente un conjunto de reglas explícitas que pueden extraerse de expertos y codificarse en software. Esta visión resultó ser ingenua. Gran parte del conocimiento experto es tácito, intuitivo, contextual y difícil de articular. Se adquiere a través de años de práctica y experiencia, no solo memorizando reglas. Los expertos humanos constantemente adaptan su conocimiento a situaciones nuevas, reconocen analogías, hacen juicios matizados que consideran múltiples factores sutiles. Reducir esto a reglas explícitas pierde información crucial. Las aproximaciones modernas de aprendizaje automático, que aprenden patrones implícitos desde datos masivos sin requerir reglas explícitas, abordan este problema de manera diferente. Pero también traen nuevos desafíos, como veremos en el siguiente capítulo.

El segundo invierno terminó no con un momento dramático, sino gradualmente, a medida que nuevas técnicas, nuevos datos y nuevo hardware creaban nuevas posibilidades. Para mediados de los 2000s, quedaba claro que el aprendizaje automático basado en datos masivos era más prometedor que sistemas basados en conocimiento codificado a mano. Las redes neuronales profundas comenzaban

a mostrar resultados sorprendentes en reconocimiento de imágenes y voz. Empresas tecnológicas invertían agresivamente en IA práctica. Una nueva primavera llegaba, trayendo entusiasmo renovado pero también, inevitablemente, nuevos riesgos de expectativas infladas. La pregunta clave era: ¿había aprendido el campo lo suficiente de dos inviernos anteriores para evitar un tercero? O ¿los patrones de euforia, promesas exageradas, decepción y recortes se repetirían una vez más? Esa es la historia del próximo capítulo.

Manual para no congelarse hoy

En 2012, un equipo de investigadores liderado por Geoffrey Hinton sorprendió a la comunidad científica ganando la competencia ImageNet de reconocimiento de imágenes por un margen enorme. Usaron una red neuronal profunda entrenada con millones de imágenes etiquetadas, aprovechando GPUs (procesadores gráficos) que aceleraban los cálculos masivamente. El error del sistema era casi la mitad del segundo lugar. Este momento marcó el inicio de la "revolución del aprendizaje profundo": redes neuronales con muchas capas, alimentadas con datos masivos y entrenadas con hardware potente, comenzaron a superar métodos tradicionales en tarea tras tarea. En pocos años, el aprendizaje profundo transformó reconocimiento de voz, traducción automática, diagnóstico médico, con-

ducción autónoma, generación de texto e imágenes. Empresas tecnológicas invirtieron miles de millones. Universidades lanzaron programas especializados. La IA volvía a estar en todas partes, prometiendo revolucionar todo. El ciclo familiar de entusiasmo comenzaba de nuevo.

Los logros del aprendizaje profundo son genuinamente impresionantes. Sistemas de reconocimiento de voz como los de Google, Amazon y Apple entienden lenguaje natural con precisión sorprendente. Traductores automáticos como DeepL y Google Translate producen textos fluidos en decenas de idiomas. Sistemas de diagnóstico médico detectan cáncer en imágenes con precisión comparable o superior a radiólogos humanos. Vehículos autónomos conducen millones de kilómetros usando visión por computadora. Modelos de lenguaje como GPT generan textos convincentes, responden preguntas complejas, escriben código funcional. AlphaGo venció al campeón mundial de Go, un juego que se consideraba demasiado complejo para máquinas. AlphaFold revolucionó biología molecular prediciendo estructuras de proteínas con precisión extraordinaria. Estas no son promesas: son aplicaciones funcionando hoy, usadas por millones de personas. El progreso desde 2012 ha sido más rápido y amplio que en todas las décadas anteriores combinadas.

Sin embargo, junto a estos éxitos aparecen señales de alarma que recuerdan a inviernos anteriores. Primera señal: consumo energético astronómico. Entrenar un modelo de lenguaje grande consume electricidad equivalente a cientos de hogares durante meses. Los centros de datos de IA requieren enfriamiento masivo. GPT-3, un modelo de OpenAI, supuestamente consumió energía equivalente a 552 toneladas de emisiones de CO2 solo en entrenamiento.

Esto plantea preguntas urgentes sobre sostenibilidad ambiental. Si cada mejora requiere exponencialmente más energía, ¿cuánto tiempo es viable este crecimiento? Segunda señal: concentración de poder. Solo unas pocas empresas (Google, Microsoft, Meta, OpenAI) pueden permitirse entrenar modelos de frontera. Esto crea dependencia y reduce diversidad de aproximaciones. Los laboratorios universitarios, que tradicionalmente lideraban investigación fundamental, no pueden competir en escala de recursos.

Tercera señal de alarma: los datos usados para entrenar modelos a menudo provienen de fuentes éticamente cuestionables. Modelos de lenguaje se entrenan "raspando" internet: recolectando textos de sitios web, redes sociales, libros digitalizados, a menudo sin consentimiento explícito de autores. Modelos de generación de imágenes usan millones de fotografías e ilustraciones descargadas sin permiso de artistas. Esto genera conflictos legales y éticos sobre propiedad intelectual y derechos de creadores. Cuarta señal: opacidad algorítmica. Los modelos de aprendizaje profundo son "cajas negras": ni siquiera sus creadores entienden completamente cómo toman decisiones. Una red con cientos de millones de parámetros ajusta valores internos de formas imposibles de interpretar. Esto hace difícil detectar errores, sesgos o vulnerabilidades. Cuando un sistema rechaza injustamente una solicitud de préstamo o recomienda un tratamiento médico equivocado, es casi imposible explicar por qué.

Quinta señal preocupante: las "alucinaciones" de modelos de lenguaje. Estos sistemas generan textos que suenan convincentes pero son completamente falsos. Un modelo puede inventar referencias bibliográficas que no existen, citar estadísticas fabricadas, atribuir frases a personas que

nunca las dijeron, describir eventos históricos imaginarios. Para usuarios sin conocimiento del tema, distinguir información correcta de invención es difícil. Esto plantea riesgos serios si estos sistemas se usan en contextos donde la precisión es crucial: investigación académica, periodismo, medicina, derecho. El problema no es fácil de resolver: los modelos no "saben" qué es verdadero, solo predicen qué palabras son estadísticamente probables basándose en sus datos de entrenamiento. No entienden significado ni verifican hechos. Sexta señal: amplificación de sesgos sociales. Si entrenas un modelo con textos de internet que contienen prejuicios raciales, de género o culturales, el modelo aprenderá y reproducirá esos sesgos.

Ejemplos documentados de sesgos son abundantes y preocupantes. Sistemas de reconocimiento facial funcionan peor con personas de piel oscura, llevando a identificaciones erróneas que han resultado en arrestos injustos. Algoritmos de contratación discriminan contra mujeres porque fueron entrenados con datos históricos de empresas que contrataban principalmente hombres. Sistemas de evaluación de riesgo criminal usados en tribunales estadounidenses califican injustamente a personas afroamericanas como más propensas a reincidir. Traductores automáticos asignan género estereotípicamente: "el doctor" y "la enfermera" incluso cuando el idioma original es neutral. Generadores de imágenes, cuando se les pide crear imágenes de "CEO" o "ingeniero", producen mayoritariamente hombres blancos. Estos sesgos no son malicia intencional sino reflejo de desigualdades presentes en datos de entrenamiento. Pero el resultado es que sistemas supuestamente "objetivos" perpetúan y amplifican injusticias sociales existentes.

Un caso emblemático de fracaso es el de vehículos autónomos. Alrededor de 2015, varias empresas prometieron coches completamente autónomos para 2020. Tesla, Uber, Waymo y otras invirtieron miles de millones. Las promesas eran audaces: "en cinco años nadie necesitará licencia de conducir", "los accidentes de tráfico se reducirán 90%", "los coches autónomos serán más seguros que humanos desde el primer día". Pero la realidad resultó más compleja. Los sistemas funcionan bien en condiciones ideales (clima bueno, carreteras bien marcadas, tráfico predecible) pero fallan de maneras impredecibles en situaciones atípicas: nieve que cubre las líneas de la carretera, construcciones que cambian rutas habituales, peatones que se comportan inesperadamente, señales de tráfico vandalizadas. Han ocurrido accidentes fatales. Para 2024, ninguna empresa ha logrado vehículos autónomos completamente confiables en todas las condiciones. Las promesas se han retrasado repetidamente.

Otro tropiezo notable involucra chatbots corporativos. En 2016, Microsoft lanzó Tay, un chatbot en Twitter diseñado para aprender de conversaciones con usuarios. En menos de 24 horas, Tay comenzó a publicar mensajes racistas, sexistas y ofensivos. ¿Qué pasó? Usuarios maliciosos deliberadamente "entrenaron" a Tay alimentándola con contenido tóxico, y el sistema lo aprendió sin discriminar. Microsoft tuvo que desactivar Tay rápidamente. En 2023, cuando empresas lanzaron chatbots basados en modelos de lenguaje grandes (como Bing Chat de Microsoft o Bard de Google), aparecieron problemas similares: los sistemas producían respuestas inapropiadas, agresivas o completamente falsas. Usuarios descubrieron formas de engañar sistemas ("jailbreaking") para que ignoraran sus restricciones

de seguridad. Estos incidentes muestran que desplegar sistemas de IA en el mundo real, donde interactúan con usuarios impredecibles, es mucho más difícil que demostrarlos en ambientes controlados de laboratorio.

Un fracaso particularmente costoso ocurrió con IBM Watson Health. Después de que Watson ganara el concurso Jeopardy en 2011, IBM invirtió miles de millones desarrollando aplicaciones médicas de IA. Watson for Oncology prometía revolucionar tratamiento de cáncer, recomendando terapias personalizadas basadas en literatura médica masiva. Hospitales de todo el mundo compraron el sistema por millones de dólares. Pero investigaciones independientes revelaron que Watson frecuentemente hacía recomendaciones incorrectas o inseguras. El sistema no entendía realmente medicina; solo buscaba patrones en datos. Los médicos descubrieron que era más lento y menos útil que sus propios procesos de decisión. Para 2022, IBM había vendido o cerrado la mayoría de proyectos de Watson Health, después de invertir más de cuatro mil millones de dólares. Es un recordatorio doloroso de que resultados impresionantes en demostraciones no garantizan utilidad en aplicaciones complejas del mundo real.

Afortunadamente, la comunidad de IA ha desarrollado herramientas y prácticas para gestionar estos riesgos, aprendiendo de inviernos anteriores. Primera herramienta: "red-teaming". Consiste en contratar equipos especializados cuyo trabajo es intentar romper, engañar o hacer fallar sistemas de IA antes de desplegarlos públicamente. OpenAI, por ejemplo, contrató docenas de expertos en seguridad para atacar GPT-4 durante meses antes de lanzarlo, buscando vulnerabilidades. Estos equipos prueban si el sistema puede generar contenido dañino, revelar información

sensible, ser manipulado para comportarse incorrectamente. El red-teaming permite identificar problemas en entorno controlado donde pueden corregirse sin causar daño real. Es como contratar hackers éticos para probar la seguridad de un sistema bancario antes de que criminales reales lo intenten. Esta práctica, importada de ciberseguridad y operaciones militares, se está volviendo estándar en desarrollo responsable de IA.

Segunda herramienta: auditorías algorítmicas sistemáticas. Organizaciones independientes revisan sistemas de IA desplegados para detectar sesgos, errores o comportamientos inesperados. Por ejemplo, auditorías de sistemas de reconocimiento facial han revelado tasas de error significativamente más altas para personas de piel oscura y mujeres. Auditorías de algoritmos de contratación han identificado discriminación por género o edad. Estas auditorías usan técnicas estadísticas rigurosas: prueban el sistema con conjuntos de datos diversos, miden diferencias de desempeño entre grupos demográficos, identifican casos donde el sistema falla sistemáticamente. Idealmente, las auditorías son realizadas por terceros independientes sin conflicto de interés, y sus resultados se publican transparentemente. Algunos gobiernos están comenzando a requerir auditorías regulares de sistemas de IA usados en contextos de alto riesgo (contratación, crédito, justicia criminal, educación). Es similar a inspecciones de seguridad que se requieren para edificios, puentes o medicamentos.

Tercera herramienta: informes de incidentes estructurados. Inspirados en la aviación, donde cada accidente se investiga meticulosamente y se publica un informe detallado, algunos investigadores proponen hacer lo mismo con fallas de IA. El AI Incident Database, lanzado en 2020, recopila y

documenta casos donde sistemas de IA causaron daño: discriminación, accidentes, violaciones de privacidad, desinformación. Cada incidente se describe con detalle: qué sistema falló, en qué contexto, qué daño causó, qué factores contribuyeron, qué se aprendió. Esta base de datos permite identificar patrones: qué tipos de sistemas fallan más frecuentemente, qué contextos son más riesgosos, qué salvaguardas funcionan. Ayuda a que desarrolladores, reguladores y usuarios aprendan de errores pasados en lugar de repetirlos. Es memoria institucional colectiva, exactamente la "memoria de invierno" que este tomo defiende. El desafío es que reportar incidentes debe ser obligatorio y sin represalias, para que las empresas no oculten fallas por miedo a dañar su reputación.

Cuarta herramienta: "model cards" y "datasheets". Margaret Mitchell, Timnit Gebru y colegas propusieron que cada modelo de IA publicado debería venir con una "tarjeta" (model card) que documente: para qué fue entrenado, con qué datos, qué limitaciones conocidas tiene, en qué contextos funciona bien y dónde falla, qué sesgos se detectaron, qué pruebas se realizaron. Similarmente, cada conjunto de datos usado para entrenar IA debería tener una "hoja de datos" (datasheet) que explique: cómo se recopiló, quién lo recopiló y por qué, qué población representa, qué sesgos podría contener, quién tiene derechos sobre él. Estas herramientas de documentación son como las etiquetas nutricionales en alimentos o los prospectos de medicamentos: permiten que usuarios informados tomen decisiones basadas en transparencia. También crean incentivos para que desarrolladores consideren cuidadosamente implicaciones éticas desde el inicio, no como reflexión tardía. Mu-

chas organizaciones de investigación ahora requieren model cards y datasheets como parte de publicaciones.

Quinta herramienta: benchmarks bien diseñados que miden no solo precisión sino también robustez, equidad, interpretabilidad y eficiencia. Durante años, los sistemas de IA se evaluaban principalmente por precisión promedio: porcentaje de respuestas correctas en un conjunto de prueba. Pero esto oculta problemas importantes. Un sistema puede tener 95% de precisión global pero solo 70% en ciertos subgrupos demográficos. Puede funcionar bien con datos limpios de laboratorio pero colapsar con datos ruidosos del mundo real. Puede ser preciso pero consumir energía insostenible. Los benchmarks modernos miden múltiples dimensiones: precisión desglosada por subgrupos, robustez ante perturbaciones, capacidad de reconocer cuando no sabe la respuesta, tiempo de cómputo, consumo energético, explicabilidad de decisiones. Por ejemplo, el benchmark HELM (Holistic Evaluation of Language Models) evalúa modelos en docenas de tareas y métricas diferentes, proporcionando una vista completa de capacidades y limitaciones en lugar de un solo número simplista.

Sexta herramienta: evaluación por escenarios de uso real. En lugar de medir solo en datos de prueba estándar, algunos investigadores desarrollan "escenarios de estrés" que prueban sistemas en condiciones desafiantes que reflejan el mundo real. Por ejemplo, probar un sistema de reconocimiento de voz no solo con audio grabado en estudio silencioso, sino también con ruido de fondo, hablantes con acentos diversos, personas con discapacidades del habla. Probar un sistema de diagnóstico médico no solo con casos claros y típicos, sino también con enfermedades raras, síntomas ambiguos, pacientes con múltiples condiciones si-

multáneas. Esta evaluación realista revela fragilidades que métricas promedio ocultan. También permite identificar para qué contextos específicos el sistema es confiable y para cuáles no, información crucial para despliegue responsable. Es como probar un auto no solo en carretera perfecta con clima ideal, sino también en nieve, lluvia, caminos sin pavimentar y tráfico denso.

Las buenas prácticas de comunicación también han evolucionado. Organizaciones responsables ahora publican "limitaciones conocidas" explícitamente cuando lanzan sistemas. OpenAI, al lanzar GPT-4, publicó un documento técnico de casi cien páginas detallando no solo capacidades sino también fallas: tipos de preguntas que el modelo responde incorrectamente, sesgos identificados, situaciones donde alucina, vulnerabilidades de seguridad conocidas. Esta transparencia permite que usuarios, investigadores y reguladores evalúen riesgos informadamente. Contrasta radicalmente con prácticas anteriores donde limitaciones se mencionaban solo en letra pequeña o se ignoraban completamente. Algunas empresas van más allá: publican "avisos de uso" que especifican para qué contextos el sistema es apropiado y para cuáles no. Por ejemplo: "Este sistema es apropiado para sugerir diagnósticos preliminares pero no para tomar decisiones finales de tratamiento sin supervisión médica profesional". Comunicación honesta construye confianza duradera.

Otra práctica valiosa es diseñar sistemas que "fallen de forma segura". En ingeniería de seguridad crítica (aeronáutica, medicina, energía nuclear), se asume que todo sistema eventualmente fallará, y se diseña para que cuando falle, lo haga de manera que minimice daño. Aplicado a IA: un sistema de diagnóstico médico que no está seguro debería de-

cir "no sé, consulte a un especialista" en lugar de adivinar. Un vehículo autónomo que enfrenta situación que no entiende debería detenerse de forma segura y pedir control humano en lugar de intentar improvisar. Un chatbot que detecta que va a generar información falsa debería decir "no tengo información confiable sobre esto" en lugar de inventar. Diseñar incertidumbre explícita en sistemas de IA es técnicamente difícil pero crucialmente importante. Los humanos son buenos reconociendo los límites de su conocimiento; necesitamos que las máquinas aprendan esta humildad epistémica también.

La medición de costos energéticos y ambientales se está volviendo práctica estándar. Investigadores publican cada vez más el consumo energético y la huella de carbono de entrenar modelos grandes. Algunas conferencias académicas ahora requieren que autores reporten el costo computacional de sus experimentos. Esto crea conciencia sobre sostenibilidad y presión para desarrollar técnicas más eficientes. Aparecen investigaciones sobre "IA verde": cómo lograr resultados comparables con menos cómputo. Por ejemplo, "destilación de conocimiento" entrena modelos pequeños que imitan modelos grandes pero consumen mucho menos energía. "Poda de redes" elimina conexiones innecesarias reduciendo tamaño sin perder precisión significativa. "Búsqueda de arquitectura neural eficiente" diseña modelos optimizados para hardware específico. Estas técnicas no solo reducen impacto ambiental sino también hacen IA accesible a organizaciones con presupuestos limitados que no pueden permitirse entrenar modelos gigantes.

La gobernanza regulatoria también está emergiendo. La Unión Europea propuso el AI Act, una ley que clasifica sistemas de IA por nivel de riesgo y requiere salvaguardas

proporcionales. Sistemas de "riesgo inaceptable" (como scoring social masivo) quedarían prohibidos. Sistemas de "alto riesgo" (contratación, crédito, justicia criminal, educación, medicina) requerirían auditorías rigurosas, documentación transparente, supervisión humana, y sistemas de apelación cuando decisiones afecten personas. Estados Unidos está desarrollando regulaciones sectoriales. China ha implementado regulaciones sobre algoritmos de recomendación y generación de contenido. Aunque los detalles varían, emerge consenso global de que IA poderosa requiere supervisión, especialmente en contextos de alto impacto social. El desafío es regular lo suficiente para prevenir daños serios sin ahogar innovación legítima. Es un equilibrio delicado que requiere diálogo continuo entre tecnólogos, reguladores, sociedad civil y usuarios afectados.

Convertir "memoria de invierno" en políticas concretas requiere instituciones dedicadas. Algunos países están creando agencias gubernamentales especializadas en IA. Reino Unido estableció la Fundación para Modelos de IA (AI Foundation Models) para evaluar riesgos de sistemas avanzados. Estados Unidos está fortaleciendo capacidades del National Institute of Standards and Technology (NIST) para desarrollar estándares de IA. Canadá, Singapur y otros países crean centros de excelencia en IA responsable. Universidades lanzan programas interdisciplinarios que combinan informática, ética, derecho y ciencias sociales. Organizaciones no gubernamentales como Partnership on AI, AI Now Institute, y Ada Lovelace Institute documentan impactos sociales y proponen mejores prácticas. Estas instituciones crean infraestructura social necesaria para gestionar tecnología poderosa de forma responsable, aprendiendo de experiencias pasadas y anticipando desafíos futuros.

En educación, aparecen currículos que enseñan no solo técnicas de IA sino también implicaciones éticas y sociales. Estudiantes de informática aprenden sobre sesgos algorítmicos, privacidad de datos, explicabilidad, equidad, impacto ambiental y responsabilidad profesional. Algunas universidades requieren cursos de ética como parte de programas de IA. Stanford, MIT y otras instituciones desarrollan casos de estudio basados en fallas reales para que estudiantes aprendan de errores pasados. Organizaciones profesionales como ACM y IEEE publican códigos de ética para ingenieros de IA. Estas iniciativas educativas son "memoria de invierno" aplicada: transmitir a nuevas generaciones lecciones aprendidas dolorosamente por generaciones anteriores. Un ingeniero que conoce la historia de ALPAC, Lighthill y Watson Health está mejor preparado para evitar repetir esos errores en su propia carrera.

Finalmente, las prácticas contractuales y de adquisición están cambiando. Gobiernos y grandes organizaciones que compran sistemas de IA comienzan a exigir garantías específicas en contratos: documentación completa de limitaciones, auditorías independientes, explicabilidad de decisiones, procedimientos de apelación, responsabilidad legal clara cuando el sistema falla. Algunos contratos requieren "cláusulas de salida": capacidad de terminar uso del sistema y migrar a alternativas si no funciona según lo prometido, evitando dependencia tecnológica peligrosa. Otros exigen que proveedores mantengan y actualicen sistemas durante plazos específicos, no solo entregar software y desaparecer. Estas prácticas contractuales traducen lecciones de inviernos anteriores en protecciones legales concretas. Son memoria de invierno codificada en contratos: compradores y vendedores acuerdan explícitamente expectativas realistas,

métricas de éxito, responsabilidades mutuas y procedimientos cuando las cosas no funcionan según lo planeado. Es madurez institucional aplicada a tecnología poderosa y compleja.

Conclusión

Hemos viajado a través de setenta años de ambiciones, promesas, fracasos y aprendizajes. Desde el optimismo desbordante de Dartmouth en 1956 hasta los dilemas actuales del aprendizaje profundo, la inteligencia artificial ha vivido ciclos de entusiasmo y decepción que siguen patrones predecibles. Cada invierno llegó después de promesas exageradas que chocaron contra límites técnicos, económicos o humanos. Cada invierno trajo dolor: proyectos cancelados, carreras truncadas, confianza pública dañada, inversiones perdidas. Pero cada invierno también generó lecciones valiosas sobre cómo gestionar tecnología poderosa de forma más sensata y sostenible. La pregunta central de este tomo ha sido: ¿podemos aprender de estos ciclos para evitar repetirlos? ¿Podemos construir memoria colectiva que nos ayude a navegar mejor la próxima primavera, el próximo verano, el próximo

otoño? La respuesta es sí, pero solo si tomamos en serio las lecciones históricas y las traducimos en prácticas concretas.

El primer invierno (años 70) nos enseñó que resolver problemas inteligentes es más difícil de lo que parece. El lenguaje humano, el sentido común, la adaptabilidad a situaciones nuevas, todas estas capacidades que los humanos ejercemos sin pensar resultan extraordinariamente complejas de capturar en programas. Las computadoras de esa época eran lentas y caras, los datos escasos, las métricas de evaluación pobres. Pero más allá de limitaciones técnicas, el primer invierno reveló problemas humanos: ambición desmedida, comunicación exagerada, presión por resultados rápidos, confusión entre investigación exploratoria y aplicaciones listas para usar. Los informes ALPAC y Lighthill fueron dolorosos pero necesarios: obligaron al campo a confrontar honestamente qué funcionaba y qué no. De ese invierno emergieron prácticas más sensatas: objetivos medibles, pruebas pequeñas antes de grandes inversiones, presupuestos realistas, comunicación diferenciada entre lo consolidado y lo especulativo.

El segundo invierno (años 90) nos enseñó que el conocimiento humano es más fluido, tácito y contextual de lo que los sistemas expertos asumían. Reducir experiencia de décadas a reglas explícitas pierde información crucial. Mantener sistemas basados en conocimiento codificado a mano resultó prohibitivamente costoso. El hardware especializado de IA no podía competir económicamente con computadoras de propósito general cada vez más potentes. El proyecto japonés de Quinta Generación mostró que inversión masiva sin claridad técnica no garantiza éxito. Pero el segundo invierno también trajo aprendizajes valiosos: la

importancia de separar prototipos de productos, de calcular costos de ciclo de vida completo, de evaluar en condiciones realistas, de construir equipos interdisciplinarios. Durante ese invierno, avances menos visibles (mejora de algoritmos de aprendizaje automático, acumulación de datos digitales, refinamiento de teoría) crearon fundamentos para el renacimiento posterior.

El periodo actual (desde 2012) ha traído logros genuinamente impresionantes: reconocimiento de voz y visión casi humanos, traducción fluida, diagnóstico médico preciso, generación de texto convincente. Pero también vemos señales de alarma: consumo energético insostenible, concentración de poder en pocas empresas, sesgos que perpetúan injusticias, alucinaciones que producen desinformación, opacidad que dificulta responsabilidad, fallos en contextos reales que cuestan vidas y fortunas. ¿Estamos ante un tercer invierno? Depende de cómo gestionemos estos riesgos en los próximos años. La buena noticia es que hoy tenemos herramientas, instituciones y prácticas más sofisticadas que en inviernos anteriores: red-teaming, auditorías, informes de incidentes, model cards, benchmarks holísticos, regulación emergente, educación ética. Pero estas herramientas solo funcionan si las usamos sistemáticamente, no como decoración voluntaria.

La "memoria de invierno" que este tomo defiende no es nostalgia ni pesimismo. Es memoria práctica: registros organizados de qué salió mal, por qué, y cómo evitarlo. Es similar a cómo la aviación mantiene bases de datos de accidentes, cómo la medicina organiza comités de morbimortalidad, cómo la ingeniería civil estudia puentes caídos. No se trata de avergonzar a nadie ni de frenar innovación, sino de aprender sistemáticamente de errores para no repetirlos.

Una memoria de invierno robusta incluye: documentación técnica de fallas (qué sistema, qué contexto, qué falló), análisis de causas profundas (técnicas, organizacionales, humanas), propuestas de corrección (qué cambios prevendrían fallas similares), diseminación amplia (publicar lecciones para que otros aprendan), e incorporación institucional (traducir lecciones en políticas, contratos, currículos, regulaciones). Es un círculo completo de aprendizaje colectivo.

Destilemos todo en un decálogo práctico: diez reglas simples para evitar repetir los errores más costosos de la historia de la IA. Estas reglas no garantizan éxito, pero reducen significativamente el riesgo de fracasos evitables. No requieren genialidad técnica: requieren disciplina, honestidad y humildad. Son aplicables tanto para investigadores en laboratorios como para empresas desplegando productos, para gobiernos comprando sistemas, para periodistas cubriendo tecnología, para usuarios decidiendo si confiar en un sistema. El decálogo sintetiza setenta años de experiencia colectiva. Ignorarlo es arrogancia costosa; aplicarlo es sabiduría práctica. Las reglas están formuladas en lenguaje sencillo, sin jerga, accesibles para cualquier persona involucrada en decisiones sobre IA. Son memoria de invierno condensada en principios accionables.

Primera regla: medir bien antes de prometer. Define objetivos específicos y medibles, no vagos. "Mejorar diagnóstico de neumonía en radiografías de tórax con precisión de 90% en adultos" es medible. "Revolucionar la medicina" no lo es. Usa métricas múltiples que capturen precisión, robustez, equidad, costo, consumo energético. Compara contra alternativas realistas (humanos, métodos tradicionales, no hacer nada). Prueba en datos diversos que representan

el mundo real, no solo en casos ideales de laboratorio. Reconoce cuándo tu métrica es insuficiente: precisión promedio puede ocultar fallas graves en subgrupos específicos. Medir bien es fundamento de honestidad: te obliga a confrontar qué tan bien funciona realmente tu sistema, no qué tan bien quisieras que funcione. Es termómetro que previene fiebres de expectativas infladas.

Segunda regla: comunicar sin humo. Distingue explícitamente entre lo que funciona hoy de manera confiable, lo que podría funcionar en tres a cinco años con esfuerzo razonable, lo que es investigación exploratoria a largo plazo sin garantías, y lo que es especulación teórica. Comunica limitaciones conocidas con la misma claridad que capacidades. Actualiza estimaciones cuando nueva evidencia cambia el panorama. Evita lenguaje sensacionalista que exagera logros o minimiza riesgos. Recuerda que audiencias diferentes (inversores, medios, usuarios, reguladores) interpretan mensajes de formas diferentes; adapta tu comunicación pero mantén honestidad central. Comunicar sin humo no es aburrido ni destruye entusiasmo genuino: construye confianza sostenible. La gente respeta más la honestad que la exageración. Los inviernos llegaron precisamente cuando la brecha entre promesas y realidad se volvió insostenible.

Tercera regla: cuidar datos y trabajo humano. Los datos que alimentan sistemas de IA no aparecen por magia: alguien los creó, los recopiló, los etiquetó. Respeta derechos de creadores y privacidad de usuarios. Obtén consentimiento explícito cuando uses datos personales. Compensa justamente a trabajadores que etiquetan datos, a menudo en condiciones precarias. Documenta cómo se recopilaron datos, qué sesgos podrían contener, qué poblaciones repre-

sentan o excluyen. Reconoce que "datos grandes" no son necesariamente "datos buenos": calidad importa más que cantidad. Datos sesgados producen sistemas sesgados; datos de baja calidad producen sistemas poco confiables. Cuidar datos también significa cuidar las personas cuyo trabajo, creaciones y privacidad se convierten en combustible de IA. Es ética práctica traducida en gestión responsable de recursos fundamentales.

Cuarta regla: planificar mantenimiento desde el inicio. El costo real de un sistema no es solo desarrollo inicial sino desarrollo más despliegue más mantenimiento más actualización más eventual retiro. Planifica presupuesto, personal y procesos para todo el ciclo de vida. ¿Qué pasa cuando el sistema encuentra situaciones que sus creadores no anticiparon? ¿Quién lo monitoreará, detectará errores, implementará correcciones? ¿Cómo se actualiza cuando cambian contextos (nuevas enfermedades, nuevas regulaciones, nuevos patrones de fraude)? ¿Quién responde cuando algo falla? ¿Cómo se retira el sistema de forma segura si deja de ser útil? Un sistema sin plan de mantenimiento es un experimento, no una solución. Los sistemas expertos de los años 80 quebraron empresas porque nadie presupuestó el costo astronómico de mantenerlos actualizados. No repitas ese error.

Quinta regla: evaluar energía y sostenibilidad. El progreso técnico que consume recursos insostenibles no es progreso real. Mide y reporta consumo energético de entrenar y ejecutar modelos. Busca eficiencia: cómo lograr resultados comparables con menos cómputo. Considera impacto ambiental completo: electricidad, enfriamiento, fabricación de hardware, obsolescencia acelerada. Pregunta si la aplicación justifica el costo energético: ¿vale la pena con-

sumir energía de una ciudad pequeña durante meses para mejorar recomendaciones de películas en 2%? Investiga técnicas eficientes: destilación, poda, arquitecturas optimizadas. Apoya investigación en "IA verde". La sostenibilidad no es lujo moral: es requisito práctico para viabilidad a largo plazo. Un campo que consume energía insosteniblemente eventualmente enfrentará límites políticos, económicos o físicos que frenarán su crecimiento.

Sexta regla: auditar sistemáticamente. No asumas que tu sistema funciona como esperas: verifica. Prueba con usuarios diversos, en contextos variados, bajo condiciones de estrés. Contrata auditores independientes sin conflicto de interés. Busca específicamente sesgos, errores en subgrupos, vulnerabilidades de seguridad, casos donde el sistema falla silenciosamente. Usa técnicas como red-teaming: paga gente inteligente para intentar romper tu sistema. Repite auditorías regularmente: un sistema que funcionaba bien hace seis meses puede degradarse cuando cambia el contexto. Publica resultados de auditorías transparentemente. Las auditorías no son castigo sino mantenimiento preventivo. Son como inspecciones de seguridad para edificios o puentes: detectan problemas antes de que causen tragedias. Un sistema no auditado es un riesgo invisible.

Séptima regla: registrar fallos sistemáticamente. Crea "libros de fallos" donde documentes: qué experimento o despliegue falló, en qué contexto, qué daño causó, qué factores contribuyeron, qué aprendiste, qué correcciones implementaste. Trata los fallos como información valiosa, no como vergüenza a ocultar. Incentiva reportar errores sin represalias: crea culturas organizacionales donde admitir "esto no funcionó" es respetado, no castigado. Comparte lecciones ampliamente: publica casos de estudio, contribu-

ye a bases de datos de incidentes como AI Incident Database. Estudia fallos de otros: lee informes de incidentes, analiza qué salió mal, pregunta si tus propios sistemas tienen vulnerabilidades similares. Registrar fallos es construcción de memoria colectiva. Cada fallo documentado y analizado es una lección que puede prevenir tragedias futuras.

Octava regla: compartir aprendizajes abiertamente. Publica no solo éxitos sino también limitaciones, fracasos, lecciones. Escribe "model cards" que documenten capacidades y límites de sistemas. Crea "datasheets" que expliquen cómo se recopilaron datos y qué sesgos pueden contener. Participa en conferencias, publicaciones académicas, foros públicos donde se discuten mejores prácticas. Contribuye a estándares técnicos y éticos del campo. Enseña: ofrece cursos, tutoriales, materiales educativos que transmitan conocimiento a nuevas generaciones. Colabora con investigadores de otros campos (ética, derecho, ciencias sociales, políticas públicas). El conocimiento que se guarda en secreto beneficia solo a quien lo tiene; el conocimiento compartido beneficia a toda la sociedad. Compartir es inversión en salud colectiva del campo.

Novena regla: proteger a usuarios siempre. Las personas afectadas por sistemas de IA merecen protección, información y poder. Informa claramente qué hace un sistema, cómo toma decisiones, qué datos usa. Diseña sistemas que "fallen de forma segura": si no están seguros, deben reconocerlo y escalar a supervisión humana. Proporciona mecanismos de apelación: si un sistema toma decisión que afecta negativamente a alguien (rechaza préstamo, niega empleo, recomienda tratamiento equivocado), esa persona debe poder cuestionar la decisión y obtener revisión humana. Respeta privacidad y consentimiento. No desplie-

gues sistemas de alto riesgo (medicina, justicia, educación) sin pruebas exhaustivas y supervisión continua. Pregunta a usuarios afectados qué protecciones consideran importantes. Proteger usuarios no es altruismo: es responsabilidad profesional básica y requisito para confianza social sostenible.

Décima regla: mantener curiosidad con prudencia. No dejes que el miedo al fracaso mate la exploración, pero no confundas exploración arriesgada con despliegue prematuro. Investigación exploratoria (probar ideas nuevas sin garantías de éxito) es vital para progreso científico. Pero debe hacerse con salvaguardas apropiadas: en laboratorios con controles éticos, con revisión de pares, con transparencia sobre incertidumbres. Separar claramente investigación exploratoria de aplicaciones listas para uso real. Celebra tanto éxitos como fracasos productivos en investigación, pero aplica estándares más rigurosos para sistemas que afectan vidas reales. La curiosidad impulsa descubrimiento; la prudencia previene daños evitables. Ambas son necesarias. Los mejores científicos e ingenieros son audaces en imaginar posibilidades pero humildes en reconocer límites y riesgos.

Este decálogo no es exhaustivo ni final. Es punto de partida, destilación de lecciones históricas en principios prácticos. Cada campo específico (medicina, finanzas, educación, justicia) necesitará adaptar y extender estas reglas a sus contextos particulares. Cada organización necesitará traducirlas en políticas, contratos, currículos, procesos internos concretos. El decálogo tampoco es sustituto de pensamiento crítico: habrá situaciones donde las reglas entren en tensión (por ejemplo, transparencia versus privacidad) y se requieran juicios matizados. Pero estas diez reglas cap-

turan el núcleo de lo que setenta años de historia nos enseñan: medir honestamente, comunicar claramente, cuidar recursos humanos, planificar sostenibilidad, auditar continuamente, aprender de fallos, compartir conocimiento, proteger usuarios, y balancear audacia con humildad. Son brújula para navegar territorios complejos e inciertos.

Los inviernos de la IA no son castigos divinos ni accidentes aleatorios. Son consecuencias predecibles de patrones humanos recurrentes: ambición que ignora límites, comunicación que exagera logros, inversión que busca retornos imposibles, instituciones que no aprenden de errores pasados. Estos patrones no son exclusivos de IA: aparecen en toda tecnología poderosa y en muchas empresas humanas (burbujas financieras, manías médicas, modas educativas). La diferencia es que algunos campos han desarrollado mejores mecanismos de aprendizaje institucional. La aviación aprendió a analizar accidentes sistemáticamente después de décadas de tragedias. La medicina aprendió a exigir ensayos clínicos rigurosos después de desastres farmacológicos. La ingeniería civil aprendió a estudiar estructuras colapsadas después de puentes caídos. La IA está aprendiendo ahora, quizás más rápido que campos anteriores porque tiene esos modelos como referencia. Pero aprender requiere voluntad colectiva de confrontar errores honestamente.

Cada lector de este tomo tiene un papel en construir memoria de invierno y aplicarla. Si eres investigador, documenta tus fracasos además de éxitos, publica limitaciones claramente, colabora interdisciplinariamente. Si eres ingeniero desplegando sistemas, audita rigurosamente, protege usuarios, planifica mantenimiento. Si eres inversor o ejecutivo, exige honestidad sobre límites, financia sostenibilidad a largo plazo, no solo demostraciones espectacu-

lares. Si eres regulador o funcionario, desarrolla políticas informadas por historia, requiere transparencia y responsabilidad. Si eres periodista, cubre tanto promesas como riesgos, explica matices sin sensacionalismo. Si eres educador, enseña ética junto con técnicas, usa casos históricos como lecciones. Si eres usuario, pregunta cómo funcionan sistemas que te afectan, exige explicaciones, participa en debates públicos. La memoria de invierno es responsabilidad colectiva, no solo de expertos.

Cerramos con una invitación a ver los inviernos no como fracasos definitivos sino como pausas necesarias para crecer de forma más inteligente y humana. El invierno no es muerte: es tiempo de consolidación, reflexión, reparación. Durante los inviernos, las raíces se fortalecen bajo tierra mientras la superficie parece dormida. Los árboles que sobreviven inviernos crecen más fuertes y profundos. Lo mismo aplica a campos de conocimiento: los inviernos de la IA, aunque dolorosos, generaron lecciones, prácticas, instituciones y humildad que hoy nos sirven. Si aplicamos esas lecciones conscientemente, podemos hacer que la próxima primavera sea más sostenible, el próximo verano más equitativo, el próximo otoño más preparado. Podemos construir inteligencia artificial que sirva genuinamente a la humanidad, que respete límites ecológicos y sociales, que reconozca incertidumbres, que proteja a los más vulnerables. Ese futuro no está garantizado, pero es posible si elegimos recordar, aprender y actuar con sabiduría colectiva. La memoria de invierno es nuestra herramienta más poderosa para construirlo.

Prefácio

Há livros que não se limitam a descrever, mas a interrogar. Que não apenas explicam o mundo, mas convidam a repensá-lo. A coleção *Memórias mínimas do barro e do silício* pertence a essa rara estirpe. Oito volumes breves, mas de fôlego longo, que percorrem as marcas profundas do desejo humano de criar inteligência. Não a inteligência abstrata dos tratados técnicos, mas aquela que brota do barro e da palavra, do ritual e da engrenagem, do mito e do cálculo.

O primeiro volume, *Autômatos e espíritos*, conduz-nos às origens simbólicas do artifício: desde os golens de barro e talismãs animados até a certeza de que a matéria, sob certas mãos e certas palavras, pode despertar. Onde se esculpe um corpo, pulsa uma pergunta sobre a alma. Segue-se *De*

engrenagens e fórmulas, que narra como o pensamento mecânico emergiu não de uma ruptura com o sagrado, mas da lenta tradução do assombro em medida. O mundo revelou-se, não menos poético, ao descobrir-se repetível. *A guerra e o código* mergulha nos arquivos do século XX para mostrar-nos que a inteligência artificial não nasceu num laboratório, mas na urgência da guerra, quando decifrar uma mensagem podia salvar mil vidas. O código tornou-se sistema; o segredo, forma de saber.

Em *A tribo de Dartmouth*, assistimos ao nascimento de uma comunidade. Não um grupo de técnicos isolados, mas uma confraria com seus rituais, linguagens e heranças. Daquele encontro de 1956, em New Hampshire, surgiu algo mais do que uma disciplina: uma visão compartilhada do futuro. Tive o privilégio, décadas depois, de encontrar-me com alguns de seus pioneiros, quando, bolsista da AAAI e do IEEE, participei de congressos onde aquela intuição inicial já se havia transformado em ciência. Suas palavras, sua generosidade intelectual e sua lucidez guiaram-me nos primeiros passos por um território ainda incerto. Naqueles anos, o inverno da IA parecia longo e frio, mas a memória dos que sonharam antes de nós tornava o caminho menos árduo. Talvez por isso este livro não se limite a narrar uma origem: ele a revive. E, ao fazê-lo, recorda que toda comunidade científica nasce também do afeto, da transmissão silenciosa, do gesto compartilhado de acreditar que pensar, mesmo na solidão, tem sentido.

O inverno chega com *Os invernos da máquina*. Não como metáfora romântica, mas como a memória concreta de cada fracasso, de cada entusiasmo que se quebrou diante do limite técnico, do mercado ou da simples condição humana. Esse frio, que conheci em carne própria, não parali-

sa: ensina. Na pausa imposta pela desilusão, aprende-se a distinguir o necessário do urgente, o que permanece do que apenas deslumbra. Este volume resgata essas lições silenciosas –as que raramente figuram nos congressos ou nas manchetes– e lhes dá forma, como se cada tropeço fosse também uma maneira de avançar com mais verdade.

Jogadas impossíveis mergulha nos tabuleiros onde humanos e máquinas mediram o seu engenho. Não para decidir um vencedor, mas para compreender que a criatividade não é patrimônio exclusivo de um ou de outro. Quando uma máquina joga o impensado, o espelho se quebra, e vislumbramos algo novo. *As culturas da rede* delineiam o novo cenário: algoritmos que decidem o que ler, com quem falar, em que acreditar. A rede já não é ferramenta; é ambiente. E sua linguagem –feeds, métricas, otimização– infiltra-se em nossas instituições, emoções e gestos cotidianos. Por fim, *Chatbots e vozes sintéticas* põem o ouvido na revolução mais íntima: a palavra. Máquinas que não apenas obedecem, mas nos escutam e respondem. Nem sempre é fácil distinguir se conversam ou simulam, mas a fronteira entre presença e artifício já não está onde costumava estar.

Nesta coleção ressoam, de outro modo, perguntas que já habitavam a obra dos doutores salmantinos. O que é o entendimento humano? Como se forma a vontade? Onde começa a liberdade? Assim como aqueles pensadores –Francisco de Vitoria, Domingo de Soto, Luís de León, Francisco Suárez– buscaram compreender o ser humano em sua abertura a Deus, à lei e ao outro, estes livros exploram o reverso moderno: como pensar o humano quando criou uma inteligência que o imita? A Escola de Salamanca inaugurou uma forma de pensamento que, sem desligar-se do seu tempo, aspirava a princípios universais; que analisava o

comércio, a guerra, o poder, a consciência, não a partir da utilidade, mas da justiça. Essa mesma exigência de rigor moral e clareza conceitual que animou a cátedra salmantina manifesta-se aqui, não como doutrina, mas como horizonte que obriga. Ler estas obras é, de certo modo, continuar esse diálogo –não com a teologia, mas com a técnica; não com a alma, mas com o seu reflexo algorítmico. Contudo, a pergunta continua a mesma: o que significa ser humano num mundo que nos excede?

A Universidade de Salamanca, em sua dupla vocação humanista e científica, sabe que as máquinas não se pensam sozinhas. Como nos ensinaram Vitoria, Soto e Suárez, toda construção técnica traz implícita uma visão do ser humano e do seu lugar no mundo. A inteligência artificial não escapa a essa regra: é, antes de uma ferramenta, uma antropologia encarnada. Por isso, estas *memórias mínimas* não são notas de rodapé de uma tecnologia futura, mas capítulos de uma ética em curso. Não é um catálogo o que se oferece ao leitor, mas um mapa. Não uma resposta, mas um conjunto de chaves para orientar-se numa época em que a inteligência já não é apenas biológica, nem a imaginação, privilégio do humano. Estes livros convidam a pensar a IA não como ameaça ou solução, mas como continuidade de gestos milenares: animar, ordenar, prever, falar, fracassar, tentar de novo.

Em tempos em que as máquinas aprendem a conversar, escrever, decidir; em que os algoritmos não apenas processam dados, mas modelam mundos; torna-se urgente uma reflexão que não se detenha no técnico nem se perca no alarme. Esta coleção é essa pausa lúcida: um intento, belo e necessário, de compreender o que realmente dizemos quando falamos de inteligência artificial. E também –por

que não?– de perguntar o que dizemos quando falamos de nós mesmos. Diante do vértigo provocado pelas transformações tecnológicas, a Escola de Salamanca legou-nos uma virtude esquecida: a paciência intelectual, a coragem de pensar devagar, de deter-se nos matizes, nas consequências remotas, nas perguntas sem resposta imediata. Esta coleção participa desse espírito. Não teme o artifício, mas também não o idolatra. Pergunta, compara, recorda. Em tempos em que tudo parece acelerar-se, ler estes livros é uma forma de voltar a olhar, com profundidade, aquilo que estamos construindo sem perceber: uma nova imagem do ser humano.

O leitor que se aventurar por estas páginas não encontrará apenas ideias, mas um espelho. E, talvez, uma bússola.

Juan Manuel Corchado Rodríguez
Reitor da Universidade de Salamanca

Introdução

O "invierno de la inteligencia artificial" foi um período em que as máquinas "inteligentes" que pareciam tão promissoras deixaram de corresponder às expectativas e ficaram sem o combustível que as mantinha em funcionamento. O termo nasceu nos anos 1980 dentro da comunidade técnica, quando pesquisadoras e pesquisadores que haviam vivido cortes maciços de verbas começaram a comparar aqueles tempos difíceis às estações mais frias do ano. Não era apenas uma metáfora poética: descrevia uma realidade crua em que laboratórios fechavam, projetos eram cancelados e o otimismo se congelava até virar ceticismo.

Os invernos da IA seguem um padrão tão previsível quanto as estações climáticas. Primeiro chega a primavera do entusiasmo: alguém demonstra que uma máquina pode fazer algo que antes só humanos faziam. Em seguida vem o

verão do investimento maciço: governos, empresas e universidades abrem a carteira esperando uma revolução imediata. Durante esse período quente, as promessas se inflam como balões: "en cinco años tendremos traducción perfecta", "las máquinas serán médicos mejores que los humanos", "los robots harán todo el trabajo pesado". Mas chega o outono da realidade: os sistemas falham em situações imprevistas, os custos disparam, os prazos são descumpridos. Por fim, o inverno dos cortes: investidores decepcionados retiram recursos, meios de comunicação falam em fracasso e a pesquisa se contrai por anos. É como abrir um restaurante prometendo ser o melhor do mundo sem sequer ter testado as receitas.

Por que esse ciclo se repete? Não é culpa da tecnologia, e sim de como nós, humanos, a gerimos. A ambição é uma força poderosa: queremos resolver problemas grandes e complexos e, quando vemos uma ferramenta nova, imaginamos que ela pode fazer tudo de imediato. Pesquisadores e pesquisadoras sentem pressão para obter financiamento e apoio, e por isso às vezes apresentam seus trabalhos de forma mais otimista do que a realidade permite. Investidores e agentes públicos precisam de resultados rápidos para justificar o dinheiro gasto. Jornalistas buscam histórias empolgantes que captem a atenção do público. E todos nós, como sociedade, temos fome de soluções mágicas para problemas difíceis. É compreensível, mas cria um coquetel perigoso: expectativas infladas que inevitavelmente colidem com limites técnicos, econômicos e humanos que ninguém quis ver durante os dias de euforia.

No entanto, os invernos da IA não são apocalipses. Nesses períodos frios, a vida não para: apenas muda de ritmo e estratégia. Pequenas equipes de pesquisa continuam tra-

balhando com orçamentos apertados, mas muitas vezes com maior liberdade para explorar ideias arriscadas sem a pressão de entregar resultados comerciais imediatos. As universidades seguem formando estudantes, acumulando conhecimento técnico que será valioso quando chegar a próxima primavera. Os fracassos são estudados com calma, identificando o que funcionou e o que não funcionou. Consolidam-se lições importantes sobre limites, custos e melhores práticas. É como um agricultor que usa o inverno para consertar ferramentas, planejar o próximo plantio e aprender com os erros da safra anterior. O inverno não é morte; é uma pausa necessária para crescer de modo mais sensato e sustentável.

Neste tomo introduzimos um conceito chave: a "memoria de invierno". É o conjunto de registros, análises e práticas que documentam o que deu errado durante períodos de cortes e como evitar repetir os mesmos erros. Inclui relatórios técnicos que explicam por que certos sistemas falharam, depoimentos de pesquisadoras e pesquisadores que viveram aqueles tempos, análises econômicas sobre por que se perdeu tanto dinheiro e propostas de melhores formas de comunicar avanços científicos ao público. A "memoria de invierno" é como o manual de manutenção de uma máquina complexa: indica quais peças se quebram com mais frequência, em que condições e como detectar problemas antes que se tornem catastróficos. É memória coletiva que nos ajuda a tropeçar menos vezes na mesma pedra – ou, ao menos, a cair de modo menos doloroso quando o tropeço for inevitável.

Observar os fracassos não é masoquismo intelectual: é uma forma inteligente de cuidado preventivo. Em outras disciplinas isso é bem compreendido. A engenharia civil

estuda minuciosamente cada ponte que cai para projetar estruturas mais seguras. A medicina organiza comitês que analisam erros cirúrgicos para evitar sua repetição. A aviação transformou a análise de acidentes em uma ciência precisa que tornou voar extraordinariamente seguro. Todo setor maduro tem seus "livros negros" onde se registram falhas e se extraem lições. A inteligência artificial, como campo relativamente jovem, ainda está aprendendo essa disciplina. Fracassar não é vergonhoso; vergonhoso é fracassar do mesmo modo repetidas vezes sem aprender nada. Os invernos da IA contêm informações valiosas sobre limites humanos e técnicos que podemos usar para construir melhor o futuro.

A história da inteligência artificial viveu ao menos dois invernos completos e navega os riscos de um terceiro. O primeiro grande frio chegou nos anos 1970, depois que técnicas promissoras dos anos 1950 e 1960 não conseguiram corresponder a expectativas infladas sobre tradução automática e resolução de problemas complexos. O segundo inverno abateu os anos 1990, quando os "sistemas especialistas" que prometiam substituir o conhecimento humano se revelaram frágeis e caros de manter. Entre esses períodos houve avanços significativos, porém lentos e menos publicizados. Desde meados dos anos 2000, o "aprendizaje profundo" gerou um novo verão de investimentos e expectativas. Hoje vemos feitos impressionantes, mas também sinais de alerta: sistemas que falham de maneiras imprevisíveis, custos energéticos enormes, vieses que perpetuam injustiças. Cada inverno ensinou algo distinto: o primeiro sobre limites técnicos; o segundo sobre gestão de expectativas e custos; o terceiro (se vier) poderá nos ensinar sobre responsabilidade social e sustentabilidade.

Um exemplo clássico é o relatório ALPAC de 1966, que congelou a pesquisa em tradução automática por mais de uma década. Durante os anos 1950 e início dos 1960, governos dos Estados Unidos e da União Soviética investiram milhões para criar máquinas que traduzissem idiomas instantaneamente. As promessas eram grandiosas: comunicação global sem barreiras, espionagem facilitada, diplomacia mais fluida. Mas as primeiras máquinas produziam textos confusos, por vezes cômicos. A célebre frase "el espíritu está dispuesto, pero la carne es débil" teria sido traduzida para o russo e de volta ao inglês como "el vodka es bueno, pero la carne está podrida". O comitê ALPAC concluiu que a tradução automática era mais cara e menos precisa do que usar tradutores humanos profissionais. Os fundos foram cortados drasticamente. A lição central: medir bem a qualidade antes de fazer promessas públicas e comparar sempre o custo da solução tecnológica com alternativas humanas existentes.

O relatório Lighthill de 1973 marcou o fim da inocência britânica sobre a inteligência artificial. Sir James Lighthill, matemático respeitado, foi encarregado pelo governo do Reino Unido de avaliar o progresso em IA após anos de investimento generoso. Seu veredito foi devastador: criticou o que chamou de "brincadeira" sem resultados práticos mensuráveis. Segundo Lighthill, pesquisadoras e pesquisadores resolviam problemas artificiais em ambientes controlados, mas seus sistemas colapsavam diante da complexidade do mundo real. O relatório levou a cortes maciços: universidades perderam verbas, laboratórios fecharam, carreiras acadêmicas foram truncadas. Lighthill não negava o valor da pesquisa exploratória, mas exigia honestidade quanto a prazos e aplicabilidade. Sua lição permanece

atual: é preciso separar claramente a pesquisa básica (que explora possibilidades de longo prazo) do desenvolvimento aplicado (que promete soluções concretas em prazos específicos). Misturar ambos na comunicação pública gera expectativas impossíveis de cumprir.

Os sistemas especialistas viveram seu próprio ciclo dramático entre os anos 1980 e 1990. Essas ferramentas prometiam capturar o conhecimento de especialistas humanos (médicos, engenheiros, advogados) e colocá-lo à disposição de qualquer pessoa por meio de um computador. Nos anos 1980, empresas como Teknowledge e IntelliCorp receberam milhões em investimentos. Governos e corporações compraram sistemas que diagnosticavam doenças, projetavam circuitos ou avaliavam riscos financeiros. Alguns funcionaram bem em nichos muito específicos, mas a maioria mostrou-se frágil: uma pequena mudança nas condições e o sistema dava respostas absurdas. Além disso, mantê-los custava fortunas: sempre que um especialista humano aprendia algo novo, era preciso reprogramar laboriosamente as regras do sistema. Muitas empresas do setor quebraram nos anos 1990. A lição foi clara: o conhecimento humano é mais fluido e adaptável do que parece, e codificá-lo manualmente em regras rígidas é um trabalho de Sísifo.

As equipes que sobreviveram aos invernos desenvolveram práticas de sobrevivência que seguem válidas hoje. Aprenderam a fazer testes pequenos antes de promessas grandes: se o sistema funciona bem traduzindo cardápios de restaurante, talvez possa tentar textos técnicos, mas não se deve prometer literatura complexa desde o primeiro dia. Adotaram comunicação honesta sobre limites: "esto funciona el 85% de las veces en condiciones controladas, pero necesita supervisión humana". Desenvolveram orçamentos realistas

que incluíam não apenas a pesquisa inicial, mas manutenção de longo prazo, correção de erros e capacitação de usuárias e usuários. Criaram métricas claras para medir progresso: em vez de proclamar "inteligencia general", mediam tarefas específicas como "precisión en diagnóstico de neumonia en radiografías de adultos sanos". Essas práticas não são glamorosas, mas constroem confiança sustentável entre pesquisadoras/es, investidores e usuárias/os.

Desde meados dos anos 2000, o "aprendizaje profundo" tem gerado avanços impressionantes: sistemas que reconhecem imagens melhor que humanos, que traduzem com fluidez surpreendente, que geram textos convincentes. Mas também vemos sinais de risco que lembram invernos anteriores. Os sistemas mais potentes consomem energia equivalente à de cidades pequenas. Os dados necessários para treiná-los muitas vezes são obtidos sem consentimento claro das pessoas usuárias. Os modelos maiores são tão complexos que nem suas próprias criadoras e criadores compreendem completamente como tomam decisões. Produzem "alucinações": respostas que soam convincentes, mas são completamente falsas. Amplificam vieses presentes nos dados de treinamento, perpetuando injustiças sociais. Os custos de desenvolvimento tornaram-se astronômicos: apenas algumas gigantes conseguem arcar com o treinamento dos modelos mais avançados. Estamos diante de um novo inverno? Depende de como geriremos esses riscos nos próximos anos.

Felizmente, hoje dispomos de ferramentas sofisticadas para prevenir desastres que não existiam em invernos anteriores. O "red-teaming" consiste em formar equipes especializadas cujo trabalho é tentar quebrar ou enganar sistemas de IA para descobrir vulnerabilidades antes que causem

danos reais. Auditorias algorítmicas revisam sistemas implantados para detectar vieses, erros ou comportamentos inesperados. Os "informes de incidentes" documentam falhas de maneira sistemática, como os relatórios de acidentes na aviação. Alguns laboratórios mantêm "libros de fallos" onde registram experimentos que não funcionaram e por quê, criando memória institucional valiosa. Organizações como Partnership on AI ou AI Now Institute estudam impactos sociais e propõem melhores práticas. Estas são formas modernas de "memoria de invierno": sistemas organizados para capturar lições e prevenir erros repetidos.

Este tomo foi escrito para qualquer pessoa curiosa sobre como a inteligência artificial realmente funciona para além das manchetes. Não é preciso ser engenheira/o, matemática/o ou programadora/or para entender essas histórias. Queremos que estudantes, jornalistas, servidoras e servidores públicos, empresárias/os e usuárias/os preocupados possam compreender o que aconteceu historicamente e o que podemos fazer hoje para evitar repetir erros custosos. Os invernos da IA não são apenas anedotas técnicas: são histórias sobre ambição humana, gestão de recursos públicos, comunicação entre ciência e sociedade e responsabilidade sobre ferramentas poderosas. Cada cidadã e cidadão tem o direito de entender essas tecnologias que afetam seu trabalho, sua privacidade e seu futuro. E cada pessoa que toma decisões sobre investimento, regulação ou uso de IA precisa conhecer os padrões históricos para navegar melhor os riscos e oportunidades atuais.

Primeiras geadas

Como já se viu no tomo anterior, no verão de 1956 um grupo de pesquisadores se reuniu no Dartmouth College, em New Hampshire, com uma ambição audaz: organizar um workshop de dois meses para explorar se as máquinas podiam "simular cada aspecto del aprendizaje o cualquier otra característica de la inteligencia". John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon redigiram a proposta que daria nome ao campo: "inteligencia artificial". O otimismo era contagioso. McCarthy escreveu que esperavam avanços significativos em apenas um verão. As participantes e os participantes imaginavam máquinas que jogariam xadrez em nível de mestres, compreenderiam linguagem natural, resolveriam teoremas matemáticos complexos e aprenderiam com a experiência como fazem as crianças. Era uma primavera de esperanças enormes, plantada no solo fértil dos primeiros

computadores digitais e do entusiasmo do pós-guerra pela ciência e pela tecnologia.

Os primeiros anos trouxeram feitos que alimentaram esse entusiasmo. Em 1952, Arthur Samuel criou um programa que jogava damas e melhorava com a prática, demonstrando que as máquinas podiam "aprender" em algum sentido. Allen Newell e Herbert Simon desenvolveram o "Logic Theorist", um programa que demonstrou teoremas de lógica matemática de modo que surpreendeu as próprias pessoas matemáticas. Em 1958, John McCarthy inventou a linguagem de programação LISP, que se tornaria a língua materna da IA por décadas. Frank Rosenblatt apresentou o "perceptrón", um dispositivo que podia reconhecer padrões visuais simples e que prometia ser o embrião de cérebros artificiais. Cada conquista gerava manchetes entusiasmadas: "Las máquinas que piensan", "Cerebros electrónicos", "El futuro ya está aquí". Governos e universidades abriram linhas generosas de financiamento. Tudo parecia possível.

A tradução automática foi um dos sonhos mais tentadores e, por fim, um dos fracassos mais instrutivos. Durante a Guerra Fria, os Estados Unidos precisavam desesperadamente traduzir documentos científicos e militares soviéticos. A promessa era simples: se os computadores podiam realizar cálculos complexos, certamente poderiam buscar palavras em um dicionário e reorganizá-las segundo regras gramaticais. Em 1954, uma equipe da IBM e da Georgetown University apresentou uma demonstração que traduzia sessenta frases do russo para o inglês. A imprensa proclamou o início de uma revolução. O governo estadunidense investiu milhões de dólares em vários projetos de tradução automática. Pesquisadoras e pesquisadores prometeram sis-

temas funcionais em três a cinco anos. Parecia um problema técnico simples: vocabulário mais gramática igual a tradução. Ninguém imaginava a complexidade oculta da linguagem humana.

Mas a linguagem mostrou-se uma fera mais complexa do que o esperado. As palavras têm múltiplos significados que dependem do contexto. "Banco" pode ser uma instituição financeira ou um assento no parque. "Volar" pode significar deslocar-se pelo ar ou mover-se muito rápido. A gramática tem exceções labirínticas. Usamos referências implícitas, sarcasmo, metáforas, conhecimento cultural compartilhado. Os primeiros sistemas de tradução produziam resultados que variavam do incompreensível ao cômico. Tornaram-se célebres anedotas como a frase bíblica "el espíritu está dispuesto, pero la carne es débil", que supostamente se traduziu como "el vodka es bueno, pero la carne está podrida". Embora muitas dessas histórias tenham sido exageradas ou inventadas, captavam uma verdade incômoda: as máquinas não entendiam de fato o que traduziam; apenas manipulavam símbolos seguindo regras rígidas.

Em 1964, o governo dos Estados Unidos decidiu avaliar seriamente se valia a pena continuar investindo em tradução automática. Formou o Comitê Assessor sobre Processamento Automático da Linguagem, conhecido pela sigla em inglês ALPAC (Automatic Language Processing Advisory Committee). O comitê era composto por linguistas, matemáticas/os e engenheiras/os respeitados. Durante dois anos estudaram o estado da arte: visitaram laboratórios, testaram sistemas, compararam custos e qualidade. Em 1966, publicaram seu relatório – e ele foi demolidor. Concluíram que a tradução automática era mais lenta, menos precisa e mais cara do que contratar tradutoras e traduto-

res humanos profissionais. Não viam progresso significativo no horizonte próximo. Recomendaram redirecionar verbas para pesquisa linguística fundamental em vez de aplicações prematuras. Foi como se uma médica ou um médico lhe dissesse que seu tratamento caro não funciona e que você deveria tentar algo completamente diferente.

O impacto do relatório ALPAC foi imediato e severo. Os fundos federais para tradução automática foram cortados drasticamente. Projetos que haviam empregado dezenas de pesquisadoras/es durante anos foram cancelados. Estudantes de pós-graduação que preparavam teses sobre o tema tiveram de mudar de área. Algumas universidades fecharam laboratórios inteiros. A confiança pública na "inteligência das máquinas" sofreu um duro golpe. Por mais de uma década, a tradução automática foi considerada um beco sem saída, exemplo de promessas tecnológicas infladas. Só nos anos 1980, com novas abordagens baseadas em estatística em vez de regras gramaticais codificadas à mão, o campo começou a se recuperar lentamente. A lição fundamental do episódio ALPAC foi dupla: primeiro, medir cuidadosamente a qualidade antes de fazer promessas públicas; segundo, comparar sempre o desempenho da solução tecnológica com alternativas humanas existentes, incluindo custos totais realistas.

Enquanto os Estados Unidos sofriam o choque do ALPAC, o Reino Unido preparava sua própria avaliação crítica da inteligência artificial. Em 1972, o Conselho Britânico de Pesquisa Científica encarregou Sir James Lighthill, matemático de grande prestígio, de revisar o estado da IA no país após anos de financiamento generoso. Lighthill não era inimigo da ciência nem tecnófobo: era especialista em matemática aplicada com reputação impecável. Visitou os

principais laboratórios de IA em universidades britânicas, conversou com lideranças da pesquisa, estudou publicações e avaliou resultados concretos. Seu relatório, publicado em 1973, foi cortês na forma, mas devastador no conteúdo. Criticou o que chamou de "inteligência artificial combinatória" (a abordagem dominante à época, que tentava resolver problemas buscando sistematicamente entre soluções possíveis) por ser lenta demais e frágil quando confrontada com problemas do mundo real.

Lighthill argumentou que pesquisadoras e pesquisadores de IA resolviam "problemas brinquedo" em ambientes artificiais muito controlados, mas seus sistemas colapsavam ao enfrentar a complexidade desordenada do mundo real. Um programa podia jogar xadrez porque as regras são fixas e o tabuleiro é pequeno, mas não conseguia reconhecer uma xícara de café numa cozinha real porque as xícaras vêm em mil formas, cores e contextos diferentes. Lighthill não negava o valor da pesquisa exploratória de longo prazo, mas criticava duramente a confusão entre pesquisa básica e promessas de aplicações práticas iminentes. Segundo ele, a comunidade havia sido pouco honesta com financiadores e com o público acerca do tempo realmente necessário para obter sistemas úteis. Sua recomendação foi clara: reduzir verbas para projetos grandiloquentes de IA e redirecionar recursos para pesquisas mais modestas e honestas em áreas como robótica e processamento de sinais.

As consequências do relatório Lighthill foram brutais para a comunidade britânica de IA. O governo cortou verbas drasticamente. Universidades que haviam sido líderes mundiais no campo perderam posições acadêmicas. Jovens pesquisadores brilhantes abandonaram a IA para trabalhar em outros temas. Alguns emigraram para os Estados Uni-

dos em busca de melhores oportunidades. A confiança entre a comunidade científica e os financiadores governamentais ficou ferida por anos. Donald Michie, um dos pioneiros britânicos da IA e figura central na Universidade de Edimburgo, criticou amargamente o relatório e defendeu o valor da pesquisa de longo prazo. Mas o dano estava feito. O episódio Lighthill ensinou uma lição dolorosa, porém importante: ao pedir dinheiro público para pesquisa, há a obrigação de comunicar honestamente o que é exploração sem garantias e o que é desenvolvimento com prazos concretos. Misturar as duas coisas na comunicação gera expectativas impossíveis.

Outro golpe importante veio do próprio coração da comunidade de IA: o livro "Perceptrones", de Marvin Minsky e Seymour Papert, publicado em 1969. Os perceptrons eram modelos matemáticos simples, inspirados no funcionamento dos neurônios do cérebro. Frank Rosenblatt apresentara o perceptron em 1958 com grande fanfarra, prometendo que era o embrião de máquinas que pensariam como humanos. Durante os anos 1960, a abordagem "conexonista" (construir inteligência conectando muitas unidades simples) competiu com a abordagem "simbólica" (construir inteligência manipulando símbolos lógicos segundo regras). Minsky e Papert, figuras dominantes no MIT e defensores do enfoque simbólico, decidiram examinar rigorosamente o que os perceptrons podiam ou não fazer. Sua análise matemática demonstrou que perceptrons simples tinham limitações fundamentais: não podiam aprender a resolver certos problemas básicos, como determinar se uma imagem é composta por uma única peça conectada ou várias peças separadas.

O livro "Perceptrones" foi interpretado por muitos como sentença de morte para a abordagem conexonista. Embora Minsky e Papert mencionassem que perceptrons mais complexos (com múltiplas "camadas" de unidades) poderiam superar algumas limitações, enfatizavam que ninguém sabia como treinar eficientemente essas redes mais complexas. As verbas para pesquisa conexonista secaram. Estudantes interessados em redes neurais foram desencorajados por seus professores. Por quase duas décadas, o enfoque simbólico dominou completamente o campo. Só nos anos 1980, quando pesquisadores como Geoffrey Hinton, David Rumelhart e Ronald Williams desenvolveram técnicas eficientes para treinar redes multicamadas (o famoso algoritmo de "retropropagación"), o conexionismo ressurgiu. O episódio dos perceptrons ensina algo sutil: argumentos técnicos legítimos sobre limitações podem ser usados para encerrar prematuramente linhas de pesquisa que poderiam ser valiosas se lhes fosse dado tempo para amadurecer. A ciência precisa de crítica rigorosa, mas também de paciência e diversidade de abordagens.

Quais eram, porém, as causas de fundo que explicam esse primeiro inverno da IA? Uma causa central era o custo e a debilidade da computação disponível. Os computadores dos anos 1960 e 1970 eram máquinas enormes e lentas que custavam milhões de dólares. Um computador típico de universidade tinha menos poder de processamento do que um celular básico de hoje. Os programas de IA exigiam buscas entre milhões de possibilidades, mas as máquinas disponíveis levavam horas ou dias em tarefas que hoje tomariam segundos. Pesquisadoras e pesquisadores passavam mais tempo otimizando código para caber em memórias diminutas do que explorando ideias novas. Era como

tentar construir um arranha-céu com ferramentas de carpinteiro do século XIX: a ambição era válida, mas as ferramentas simplesmente não estavam prontas. As pioneiras e os pioneiros subestimaram massivamente quanto poder computacional seria necessário para fazer suas ideias funcionarem na prática.

Outra causa fundamental era a escassez de dados. Sistemas de IA precisam aprender com exemplos. Para treinar um sistema que reconheça gatos em fotos, são necessárias milhares de imagens rotuladas como "gato" ou "não gato". Para treinar um tradutor, são necessárias milhões de frases traduzidas por profissionais. Nos anos 1960 e 1970, esses conjuntos de dados simplesmente não existiam. Não havia internet para coletar informação em massa. Digitalizar textos, imagens ou sons era um processo manual, lento e caro. Pesquisadoras e pesquisadores trabalhavam com conjuntos minúsculos: dezenas ou centenas de exemplos, em vez dos milhões usados pelos sistemas modernos. Era como tentar aprender a cozinhar comida chinesa tendo provado apenas três pratos. As pioneiras e os pioneiros não podiam antecipar que, décadas depois, a explosão de dados digitais seria um dos fatores chave do renascimento da IA.

Uma terceira causa era a falta de objetivos claros e métricas precisas. Muitos projetos de IA inicial tinham metas vagas como "simular a inteligência humana" ou "resolver problemas complexos". Mas como saber se isso foi alcançado? O que exatamente significa "inteligência"? Sem formas claras de medir progresso, era difícil saber se um projeto estava no caminho certo ou perdido. Frequentemente, apresentavam-se sistemas em cenários cuidadosamente escolhidos, nos quais funcionavam bem, evitando-se testá-los em situações mais difíceis em que falhariam. Isso não era

necessariamente desonestidade: era falta de cultura de avaliação rigorosa. Hoje, o campo de IA usa "benchmarks" (conjuntos de testes padrão) para comparar sistemas objetivamente: precisão no reconhecimento de imagens, velocidade na tradução de texto, taxa de erro no diagnóstico de doenças. Nos anos 1960 e 1970, essa disciplina avaliativa mal existia, e sua ausência facilitava o autoengano.

A má comunicação entre pesquisadoras/es, financiadoras/es e público foi outra causa crucial do primeiro inverno. Cientistas, entusiasmados com suas descobertas, frequentemente exageravam o significado de resultados preliminares ao falar com jornalistas ou ao solicitar verbas. Jornalistas, em busca de histórias empolgantes, amplificavam esses exageros com títulos sensacionalistas. Políticos e administradores, pressionados a justificar investimentos públicos, prometiam aplicações práticas em prazos irreais. Investidores privados buscavam retornos rápidos e não entendiam a diferença entre um protótipo de laboratório e um produto comercial robusto. Ninguém tinha incentivos claros para comunicar honestamente limites e incertezas. Essa dinâmica criou uma bolha de expectativas infladas que inevitavelmente estourou quando os resultados concretos não chegaram a tempo. A lição é que comunicar ciência requer não só explicar o que funciona, mas também o que não funciona e por quê, junto com estimativas honestas de tempo e recursos necessários.

Como viveram esse inverno as pessoas pesquisadoras nos laboratórios? Para muitas, especialmente as mais jovens, foi um tempo de incerteza e frustração. Doutorandas/os que haviam dedicado anos a teses sobre IA viram seus temas perder prestígio. Professoras/es em início de carreira que apostaram sua trajetória na IA tiveram dificuldade para

publicar artigos ou conseguir financiamento. Algumas pessoas abandonaram completamente o campo e migraram para áreas mais "respeitáveis", como sistemas operacionais, bancos de dados ou teoria de algoritmos. Outras persistiram com orçamentos mínimos, trabalhando em universidades pequenas ou com verbas de projetos não diretamente rotulados como "IA". Havia sensação de injustiça: muitas sentiam que seu trabalho era sólido e valioso, mas pagavam o preço por promessas exageradas feitas por outras pessoas. O ambiente acadêmico tornou-se mais conservador: propor ideias arriscadas sobre inteligência de máquinas era malvisto.

Todavia, nem tudo foi negativo nesses anos frios. Algumas pesquisadoras e alguns pesquisadores usaram a pausa obrigatória para consolidar fundamentos teóricos que se mostrariam valiosos mais tarde. Escreveram-se livros-texto importantes que organizaram o conhecimento acumulado. Desenvolveram-se linguagens de programação e ferramentas de software que seguiriam em uso por décadas. Exploraram-se aplicações modestas e específicas que de fato funcionavam: sistemas para projetar circuitos eletrônicos simples, programas que ajudavam químicas/os a identificar moléculas, algoritmos de busca usados em planejamento logístico. Esses feitos não rendiam manchetes empolgantes, mas construíam conhecimento prático. Alguns laboratórios aprenderam a comunicar melhor: deixaram de prometer "inteligencia general" e passaram a falar de "herramientas especializadas para problemas específicos". Era um discurso menos glamoroso, porém mais honesto e sustentável.

A imprensa desempenhou um papel ambíguo durante o primeiro inverno. Nos anos de euforia (1956–1966), jornais e revistas publicaram matérias entusiásticas sobre "cere-

bros electrónicos" que em breve pensariam como humanos. Esses artigos raramente mencionavam limites ou incertezas. Quando os fracassos se tornaram evidentes, alguns meios publicaram peças muito críticas, às vezes em tom de escárnio: "Las máquinas tontas", "El futuro que nunca llegó". Esses textos tampouco eram equilibrados: pintavam todo o campo como fraude ou fantasia. Poucas/poucos jornalistas tentaram a difícil tarefa de explicar, com nuance, o que havia funcionado, o que não, por quê e o que poderia ser alcançado com mais tempo e recursos. Essa cobertura simplista (primeiro hype exagerado, depois condenação total) dificultou que o público entendesse a natureza real do progresso científico: lento, incremental, com muitos fracassos produtivos pelo caminho.

Alguns pesquisadores refletiram publicamente sobre as lições do primeiro inverno e propuseram boas práticas para evitar repeti-lo. Allen Newell e Herbert Simon, dois pioneiros respeitados, escreveram sobre a importância de definir problemas com clareza e medir o progresso de forma rigorosa. John McCarthy argumentou que a comunidade de IA precisava de maior rigor matemático e menos demonstrações espetaculares de curto prazo. Algumas pessoas propuseram separar institucionalmente a pesquisa exploratória (financiada com paciência, sem promessas de aplicações imediatas) da pesquisa aplicada (com prazos claros e métricas concretas de sucesso). Outras enfatizaram a necessidade de colaboração interdisciplinar: cientistas da computação sozinhos não poderiam resolver problemas de inteligência sem ajuda de psicologia, linguística, neurociência e filosofia. Essas reflexões nem sempre foram amplamente implementadas, mas semearam ideias que germinariam mais tarde.

Uma lição prática importante foi a necessidade de "provas de conceito" pequenas antes de grandes investimentos. Em vez de propor imediatamente sistemas que traduziriam todos os idiomas ou resolveriam todos os problemas, pesquisadoras e pesquisadores aprenderam a propor: "Demostraremos que podemos traducir manuales técnicos de física del ruso al inglés con 70% de precisión en dos años, usando este presupuesto". Esse tipo de objetivo é verificável: ao fim de dois anos, é possível medir se foi alcançado. Se funcionar, pode-se propor expansão para outros tipos de texto. Se não funcionar, analisa-se por quê e ajusta-se. Essa abordagem incremental e mensurável não é tão emocionante quanto prometer "inteligencia general", mas constrói confiança sustentável. É como reformar uma casa: começa-se por um cômodo, verifica-se se ficou bom, aprende-se com o processo e só então segue-se para o próximo.

Outra prática valiosa foi documentar não apenas sucessos, mas também fracassos. Alguns laboratórios começaram a manter registros detalhados de experimentos que não funcionaram: o que se tentou, o que deu errado, hipóteses sobre as causas. Isso criava memória institucional valiosa: quando uma/um estudante nova/o propunha uma ideia, era possível verificar se já havia sido tentada, economizar tempo evitando becos conhecidos ou tentar a ideia com modificações informadas por tentativas anteriores. Na ciência, experimentos "negativos" (que não confirmam a hipótese) são tão valiosos quanto os positivos, porque delimitam o que não funciona e por quê. No entanto, periódicos acadêmicos preferiam publicar resultados positivos, e havia pouco incentivo para compartilhar fracassos. Alguns laboratórios tentaram mudar essa cultura internamente, embora o problema persista até hoje.

Os orçamentos também precisavam tornar-se mais realistas e completos. Nos anos de euforia, muitos projetos solicitavam fundos apenas para a pesquisa inicial: contratar pessoal, comprar computadores, desenvolver algoritmos. Mas não incluíam custos de manutenção de longo prazo, correção de erros, atualização quando as necessidades mudavam, capacitação de usuárias/os, documentação técnica. Quando esses custos "ocultos" apareciam, os projetos ficavam sem dinheiro no meio do caminho. Quem sobreviveu ao inverno aprendeu a apresentar orçamentos mais honestos que incluíam todas as fases: concepção, implementação, teste, correção, implantação e manutenção. Isso fazia os projetos parecerem mais caros inicialmente, mas evitava surpresas dolorosas depois. É a diferença entre comprar um carro considerando apenas o preço de compra e considerar também seguro, combustível, manutenção e reparos.

A gestão de expectativas com investidores e com o público tornou-se outro foco de atenção. Algumas pesquisadoras e alguns pesquisadores desenvolveram o hábito de comunicar sempre em três níveis: primeiro, o que funciona hoje com confiança; segundo, o que pode funcionar no futuro próximo (três a cinco anos) com esforço e investimento razoáveis; terceiro, o que é especulação de longo prazo sem garantias. Essa estrutura ajudava financiadores e jornalistas a entender a diferença entre resultados consolidados, objetivos realistas e sonhos distantes. Por exemplo: "Hoy podemos traducir frases simples con 60% de precisión. En cinco años, con más datos y mejor hardware, esperamos alcanzar 85% en textos técnicos. Algún día quizás logremos traducir literatura preservando estilo y sutileza, pero no sabemos cuándo ni si es posible". Esse tipo de comunicação

matizada não rende manchetes espetaculares, mas constrói confiança duradoura.

As métricas de avaliação tornaram-se mais sofisticadas e honestas. Em vez de demonstrar sistemas apenas nos poucos casos em que funcionavam bem, passou-se a usar conjuntos de teste diversos e desafiadores. Desenvolveu-se o conceito de "validação cruzada": dividir os dados em dois grupos, treinar o sistema com um e testá-lo com o outro (dados que o sistema nunca viu durante o treinamento). Isso previne o autoengano: um sistema pode memorizar exemplos de treinamento sem realmente aprender padrões gerais. Também se começou a comparar sistemas de IA não apenas entre si, mas com alternativas humanas e com a opção de nada fazer. Por exemplo, se seu sistema de diagnóstico médico acerta 70% das vezes, quanto acerta um médico ou uma médica média? Quanto se acertaria simplesmente chutando a doença mais comum? Esses referenciais revelam se o sistema realmente agrega valor.

Alguns pesquisadores também refletiram sobre os limites fundamentais das abordagens daquela época. Hubert Dreyfus, filósofo do MIT e crítico famoso da IA, argumentou em seu livro "What Computers Can't Do" (1972) que a inteligência humana não é principalmente manipulação de símbolos segundo regras lógicas, mas uma capacidade encarnada que depende de ter corpo, emoções e estar imerso em um contexto social e cultural. Dreyfus era polêmico e muitas pessoas na comunidade o detestavam, mas levantava questões difíceis: pode uma máquina sem corpo entender realmente o que é "cansaço" ou "fome"? Pode um sistema de regras lógicas capturar a intuição que uma mestra ou um mestre de xadrez desenvolve após milhares de partidas? Essas perguntas filosóficas, embora incômodas, ajudaram

parte da comunidade a reconhecer que certas abordagens tinham limites inerentes e que eram necessárias ideias radicalmente novas.

Durante o primeiro inverno também surgiu maior consciência sobre a importância da interdisciplinaridade. Os primeiros projetos de IA foram dominados por cientistas da computação e matemáticas/os. Mas resolver problemas de inteligência requer compreender como a cognição humana realmente funciona. Psicologia estuda como as pessoas aprendem, lembram e resolvem problemas. Linguística estuda como a linguagem funciona no uso real. Neurociência estuda como o cérebro processa informações. Filosofia estuda o que significa "conhecimento" ou "compreensão". Algumas pesquisadoras e alguns pesquisadores de IA começaram a colaborar seriamente com especialistas dessas áreas, em vez de simplesmente supor que poderiam resolver tudo com matemática e programação. Surgiram centros de pesquisa interdisciplinares, como o Center for Cognitive Science do MIT, onde cientistas da computação, psicólogos/as, linguistas e neurocientistas trabalhavam juntos. Esse espírito colaborativo produziria frutos importantes nas décadas seguintes.

Algumas pessoas começaram a se concentrar em aplicações modestas, porém úteis, em vez de perseguir "inteligencia general". Edward Feigenbaum, químico e cientista da computação de Stanford, trabalhou em sistemas que ajudavam a identificar a estrutura molecular de compostos químicos a partir de dados de espectrometria de massas. Não era glamoroso nem pretendia simular a inteligência humana completa, mas resolvia um problema real e valioso para profissionais da química. O sistema DENDRAL, desenvolvido entre 1965 e 1977, foi considerado um dos primeiros

sucessos genuínos de IA aplicada. Funcionava porque o problema era bem definido, os dados eram estruturados e havia especialistas humanos disponíveis para validar e aprimorar o sistema. Essa linha de trabalho evoluiria para os "sistemas especialistas" dos anos 1980, que viveriam seu próprio ciclo de auge e queda.

O campo da robótica também avançou nesses anos, ainda que com expectativas mais modestas do que outras áreas da IA. Laboratórios como os do MIT e de Stanford desenvolveram braços robóticos capazes de empilhar blocos, robôs móveis que navegavam em ambientes simples usando câmeras e sensores. Shakey, um robô desenvolvido no Stanford Research Institute entre 1966 e 1972, conseguia mover-se por salas, evitar obstáculos e manipular objetos simples. Era lento, desajeitado e funcionava apenas em ambientes cuidadosamente preparados, mas demonstrava que integrar percepção, planejamento e ação em um sistema físico era possível. Pesquisadoras e pesquisadores de robótica tendiam a ser mais cautelosos em suas promessas do que outros de IA, talvez porque trabalhar com hardware torna limites e dificuldades imediatamente óbvios. Ninguém podia fingir que um robô funcionava bem quando ele caía ao tentar subir uma escada.

Alguns avanços teóricos importantes ocorreram durante o primeiro inverno, embora não tenham gerado aplicações imediatas. Em 1974, Paul Werbos, em sua tese de doutorado, propôs o algoritmo de "retropropagación" (backpropagation) para treinar redes neurais multicamadas. Esse algoritmo resolvia o problema apontado por Minsky e Papert: como ajustar as conexões internas de redes complexas para melhorar seu desempenho. Contudo, a tese de Werbos passou quase despercebida por anos. Os computadores da

época eram lentos demais para treinar redes grandes, os dados disponíveis eram escassos e a abordagem conexonista estava desacreditada. Apenas uma década mais tarde, quando outras pessoas redescobriram independentemente o algoritmo e demonstraram aplicações interessantes, a retropropagação tornou-se central. Esse caso ilustra que ideias valiosas podem chegar "antes do tempo" e permanecer adormecidas até que as condições (hardware, dados, clima intelectual) se tornem propícias.

O desenvolvimento de linguagens de programação especializadas foi outro feito menos visível, porém duradouro. LISP, criado por John McCarthy em 1958, tornou-se a linguagem padrão para pesquisa em IA por décadas. Permitira manipular símbolos e estruturas de dados complexas de forma flexível, facilitando a experimentação rápida com ideias novas. PROLOG, desenvolvido na França em 1972, ofereceu uma abordagem baseada em lógica formal: descrevem-se fatos e regras, e o sistema deduz conclusões automaticamente. Embora essas linguagens não tenham resolvido os problemas difíceis da inteligência, criaram ferramentas que aceleraram a pesquisa. É como a diferença entre construir uma casa com ferramentas modernas versus com ferramentas primitivas: o desafio arquitetônico é o mesmo, mas melhores ferramentas permitem experimentar e corrigir mais rapidamente.

Durante esses anos frios também houve reflexão sobre a relação entre IA e emprego humano. Algumas pesquisadoras, alguns pesquisadores e sindicalistas temiam que as "máquinas inteligentes" deslocassem em massa trabalhadoras e trabalhadores. Em 1964, um grupo de cientistas e ativistas escreveu a "Triple Revolution", um manifesto que alertava para o desemprego tecnológico causado pela au-

tomação. No entanto, durante o primeiro inverno, essas preocupações pareceram prematuras: as máquinas mal conseguiam executar tarefas simples e estava claro que substituir trabalhadoras/es na maioria dos empregos estava muito distante. Ainda assim, o debate levantou perguntas importantes: se algum dia a IA funcionar bem, quem se beneficia economicamente? Como se distribuem ganhos e custos sociais? Que responsabilidade têm pesquisadoras/es e empresas para com trabalhadoras/es afetados? Essas questões retornariam com força décadas depois.

Houve também reflexão ética sobre o que as máquinas "inteligentes" deveriam ou não deveriam fazer. Joseph Weizenbaum, cientista do MIT, criou ELIZA em 1966, um programa simples que simulava conversa psicoterapêutica repetindo frases do usuário em forma de perguntas. Era um truque técnico básico, mas algumas pessoas que interagiam com ELIZA sentiam conexão emocional e compartilhavam problemas pessoais profundos, mesmo após saber que era apenas um programa. Weizenbaum ficou perturbado com isso. Escreveu o livro "Computer Power and Human Reason" (1976) argumentando que há certos papéis humanos (terapeuta, juiz, professora/or) que não deveriam ser delegados a máquinas, mesmo que tecnicamente fosse possível, porque exigem empatia genuína, responsabilidade moral e compreensão do contexto humano. Sua reflexão abriu um debate ético que continua hoje: que tarefas são apropriadas para automatizar e quais não são?

Uma lição importante do primeiro inverno foi a necessidade de humildade intelectual. As pioneiras e os pioneiros da IA eram pessoas brilhantes, mas subestimaram massivamente a complexidade da inteligência. Em 1958, Herbert Simon declarou que em dez anos um computador seria

campeão mundial de xadrez (isso de fato aconteceu, mas em 1997, quase quarenta anos depois). Em 1965, Simon previu que em vinte anos as máquinas poderiam realizar qualquer trabalho que um humano pode fazer (ainda aguardamos isso em 2025). Essas previsões não eram tolices: baseavam-se em raciocínios lógicos sobre o que parecia possível. Mas subestimavam a dificuldade de capturar senso comum, contexto, adaptabilidade e as mil pequenas habilidades que humanos usam sem pensar. A lição é que resolver problemas inteligentes é mais difícil do que parece, e devemos ser modestos ao prever quando alcançaremos capacidades complexas.

O primeiro inverno também deixou claro que a ciência precisa de tempo. Alguns dos problemas que pareciam impossíveis nos anos 1970 (reconhecimento de voz, de imagens, tradução) tornaram-se viáveis décadas depois, não porque surgiram ideias radicalmente novas, mas porque três fatores melhoraram: hardware mais rápido e barato, disponibilidade de dados massivos e refinamento paciente de algoritmos. Cada geração de pesquisadoras/es aprendeu com os erros da anterior e melhorou gradualmente as técnicas. Mas esse progresso incremental é lento e requer financiamento estável de longo prazo. O sistema de verbas públicas e privadas, porém, tende a exigir resultados rápidos. Esse descompasso entre os tempos da ciência e os tempos do investimento permanece uma tensão central no campo. Uma sociedade madura precisa equilibrar financiamento para exploração paciente, sem garantias de sucesso, e financiamento para aplicações com prazos claros.

Por fim, o primeiro inverno deixou evidente que fracassos técnicos são oportunidades de aprendizado, não catástrofes morais. Muitos dos projetos que "fracassaram" nos

anos 1960 e 1970 geraram ideias, técnicas, ferramentas e conhecimento que se mostraram valiosos mais tarde. LISP continua influenciando linguagens de programação modernas. Algoritmos de busca desenvolvidos para jogos são usados hoje em planejamento logístico. As reflexões sobre representação do conhecimento informam o desenho de bancos de dados e sistemas de informação. Estudantes formados naqueles laboratórios "fracassados" levaram esse conhecimento a outros campos e contextos. Na ciência, o fracasso é informação: diz o que não funciona, por quê, e ajuda a refinar a compreensão do problema. Uma cultura científica saudável celebra tanto sucessos quanto fracassos produtivos e constrói memória coletiva que permite a cada geração apoiar-se nos ombros (e aprender dos tropeços) da anterior.

Auge e queda das promessas especialistas

No início dos anos 1980, a inteligência artificial ressurgiu de seu primeiro inverno com uma proposta renovada: se as máquinas não podiam ter "inteligência geral", ao menos poderiam capturar o conhecimento especializado de especialistas humanos em campos específicos. Nasceram os "sistemas especialistas": programas de computador concebidos para imitar o raciocínio de médicas/os, engenheiras/os, geólogas/os ou advogadas/os em áreas bem delimitadas. A ideia era entrevistar especialistas humanos, extrair suas regras de decisão ("se o paciente tem febre alta e tosse seca, então considerar pneumonia"), codificá-las em um programa e, assim, democratizar o acesso ao conhecimento valioso. Edward Feigenbaum, pioneiro em Stanford, prometia que os sistemas

especialistas revolucionariam a economia: cada empresa poderia ter acesso ao melhor especialista do mundo em qualquer tema, disponível vinte e quatro horas por dia, sem cansaço nem férias. O entusiasmo voltou a crescer – e com ele, o investimento.

Os primeiros sistemas especialistas mostraram resultados promissores em domínios muito específicos. MYCIN, desenvolvido em Stanford entre 1972 e 1980, diagnosticava infecções bacterianas no sangue e recomendava antibióticos. Em testes controlados, MYCIN acertava tanto quanto – ou mais que – médicas/os humanos medianos. DENDRAL, outro sistema de Stanford, auxiliava químicas/os a identificar estruturas moleculares a partir de dados de espectrometria de massas. XCON, desenvolvido pela Digital Equipment Corporation, configurava pedidos de sistemas informáticos complexos, uma tarefa que exigia especialistas altamente treinados. Esses êxitos geraram entusiasmo genuíno: se funcionava para diagnóstico médico e projeto de sistemas, por que não para direito, finanças, engenharia civil ou educação? Empresas especializadas como Teknowledge, IntelliCorp e Carnegie Group receberam milhões em investimento de capital de risco. Grandes corporações compraram sistemas especialistas ou contrataram equipes para desenvolver os seus próprios.

O governo japonês anunciou em 1981 o projeto da "Quinta Geração de Computadores", uma iniciativa ambiciosa de dez anos com centenas de milhões de dólares destinada a desenvolver máquinas que raciocinassem usando lógica e conhecimento, em vez de apenas processar dados. O objetivo era criar computadores capazes de compreender linguagem natural, reconhecer imagens e sons, e tomar decisões inteligentes baseadas em vastas bases de conhecimen-

to. O projeto escolheu o PROLOG, uma linguagem de programação baseada em lógica formal, como seu fundamento técnico. O anúncio causou alarme nos Estados Unidos e na Europa: se o Japão dominasse a próxima geração de computação inteligente, obteria enorme vantagem econômica e estratégica. Em resposta, os EUA lançaram iniciativas como a Microelectronics and Computer Technology Corporation (MCC) e aumentaram os fundos da DARPA para IA. A Europa respondeu com o programa ESPRIT. Uma nova corrida da inteligência artificial havia começado, impulsionada pela competição geopolítica.

Durante os anos 1980, a indústria de sistemas especialistas cresceu explosivamente. Empresas especializadas vendiam não apenas sistemas completos, mas também *shells* (estruturas vazias que permitiam às empresas construir seus próprios sistemas especialistas sem programar do zero). Universidades ofereciam cursos e pós-graduações em engenharia do conhecimento. Consultoras e consultores se especializavam em "extrair" conhecimento de especialistas humanos e codificá-lo em regras. Revistas técnicas publicavam centenas de artigos sobre novas aplicações: sistemas que avaliavam risco de crédito, diagnosticavam falhas em máquinas industriais, projetavam circuitos eletrônicos ou planejavam rotas de transporte. Em meados da década, parecia que os sistemas especialistas eram a aplicação comercial bem-sucedida que a IA procurava havia trinta anos. As expectativas atingiram níveis comparáveis aos do período anterior ao primeiro inverno. E, como antes, as fissuras começaram a aparecer.

O primeiro problema sério era a fragilidade do conhecimento codificado. Os sistemas especialistas funcionavam bem dentro dos limites estreitos para os quais haviam sido

projetados, mas colapsavam espetacularmente quando enfrentavam situações ligeiramente fora de seu domínio. Um sistema de diagnóstico de infecções bacterianas não podia ajudar com infecções virais. Se lhe perguntassem sobre um sintoma fora de sua base de conhecimento, dava respostas absurdas ou simplesmente desistia. Especialistas humanos, em contraste, raciocinam por analogia, fazem suposições informadas, reconhecem quando estão fora de sua área e buscam ajuda. Os sistemas especialistas careciam dessa flexibilidade. Eram como estudantes que memorizaram respostas para uma prova específica, mas não entendem realmente o conteúdo: funcionam perfeitamente se recebem as perguntas exatas que decoraram, mas falham miseravelmente diante de qualquer variação.

O segundo problema era o custo astronômico de manter e atualizar esses sistemas. Extrair conhecimento de especialistas humanos era um processo lento e doloroso. "Engenheiras e engenheiros do conhecimento" passavam meses entrevistando especialistas, observando seu trabalho, tentando converter sua intuição em regras explícitas. Mas o conhecimento humano não é estático: especialistas aprendem constantemente com novos casos, as boas práticas evoluem, surgem novas tecnologias e descobertas. Cada vez que o conhecimento do domínio mudava, era preciso atualizar manualmente todas as regras relevantes do sistema especialista. Isso exigia mais sessões com especialistas humanos, mais programação, mais testes. Alguns sistemas acabaram precisando de equipes permanentes de manutenção que custavam tanto quanto simplesmente contratar as pessoas especialistas que supostamente estavam sendo substituídas. Era como ter um carro que exige um mecânico em tempo integral apenas para mantê-lo funcionando.

Um terceiro problema era que o conhecimento especializado se revelou muito mais tácito e intuitivo do que as pesquisadoras e os pesquisadores esperavam. Quando se entrevista uma médica experiente sobre como diagnostica, ela pode explicar algumas regras explícitas: "se há febre e rigidez no pescoço, suspeitar de meningite". Mas grande parte de sua habilidade é intuição desenvolvida após ver milhares de pacientes: reconhecer padrões sutis, notar inconsistências entre sintomas, perceber quando algo "não encaixa", mesmo sem saber dizer exatamente por quê. Os sistemas especialistas só conseguiam capturar a parte explícita do conhecimento, perdendo a intuição tácita que muitas vezes é a mais valiosa. O filósofo Michael Polanyi havia identificado esse fenômeno décadas antes: "sabemos mais do que podemos dizer". Os sistemas especialistas chocaram-se dolorosamente com essa verdade: reduzir a experiência humana a regras explícitas elimina informações cruciais.

Enquanto os sistemas especialistas enfrentavam esses problemas práticos, o ambicioso projeto japonês da Quinta Geração encontrava dificuldades técnicas fundamentais. O PROLOG, a linguagem escolhida como base, mostrou-se menos adequado do que o esperado para construir sistemas complexos de raciocínio. As máquinas especializadas projetadas para executar PROLOG eficientemente eram caras e difíceis de programar. O objetivo de criar computadores que raciocinassem com conhecimento de senso comum mostrou-se tão difícil nos anos 1980 quanto havia sido nos 1960: ninguém sabia como codificar o vasto conhecimento implícito que os humanos usam para entender o mundo. O projeto produziu alguns avanços técnicos e formou pesquisadoras/es talentosas/os, mas não alcançou os objetivos re-

volucionários prometidos. No início dos anos 1990, estava claro que a Quinta Geração não daria o salto qualitativo esperado. O Japão havia investido centenas de milhões sem obter a vantagem estratégica buscada.

Paralelamente, a abordagem conexionista (redes neurais) experimentou um renascimento nos anos 1980, após ter sido marginalizada pela crítica de Minsky e Papert. Em 1986, David Rumelhart, Geoffrey Hinton e Ronald Williams popularizaram o algoritmo de "retropropagación" (que Paul Werbos propusera em 1974, mas passara despercebido). Esse algoritmo permitia treinar redes neurais com múltiplas camadas de modo eficiente, superando as limitações dos perceptrons simples. Pesquisadoras e pesquisadores demonstraram que essas redes podiam aprender padrões complexos a partir de dados, sem necessidade de codificar regras à mão. Isso oferecia uma alternativa atraente aos sistemas especialistas: em vez de entrevistar especialistas e codificar seu conhecimento, bastava alimentar a rede com milhares de exemplos e deixá-la aprender automaticamente. No entanto, as redes neurais tinham seus próprios problemas: exigiam grandes volumes de dados, muito poder computacional e produziam decisões difíceis de interpretar.

No final dos anos 1980, o mercado de sistemas especialistas começou a encolher. Muitas empresas que haviam adquirido sistemas descobriram que os custos de manutenção eram proibitivos. Os sistemas tornavam-se obsoletos rapidamente e exigiam atualizações constantes. Os resultados no mundo real eram menos impressionantes do que nas demonstrações controladas. Além disso, os sistemas especialistas requeriam hardware especializado (as chamadas "máquinas LISP"), extremamente caro. Quando surgiram

computadores pessoais e estações de trabalho potentes e baratas usando linguagens convencionais, o hardware especializado de IA tornou-se economicamente inviável. Empresas que haviam investido milhões em sistemas especialistas começaram a abandoná-los. A bolha de investimento em IA comercial começou a esvaziar-se. Em 1987, o mercado de hardware especializado para IA colapsou – um evento que alguns chamam de "colapso das máquinas LISP".

O segundo inverno da IA chegou no início dos anos 1990. Empresas especializadas como Teknowledge e IntelliCorp viram suas receitas despencarem; algumas faliram. A DARPA, agência de pesquisa militar dos EUA que fora financiadora-chave da IA por décadas, cortou verbas drasticamente em 1988 sob pressão do Congresso, que questionava o retorno dos investimentos. O *Strategic Computing Initiative*, programa da DARPA que investira centenas de milhões em aplicações de IA para defesa, foi parcialmente cancelado. Universidades que haviam construído grandes laboratórios de IA enfrentaram cortes orçamentários. Estudantes brilhantes evitaram pós-graduações em IA, preferindo áreas mais "seguras", como bancos de dados, redes de computadores ou interfaces de usuário. A imprensa, que havia celebrado os sistemas especialistas anos antes, agora publicava artigos sobre "o fracasso da inteligência artificial" e "promessas quebradas".

Que lições ficaram do auge e da queda dos sistemas especialistas?

Primeira lição: separar claramente protótipos de produtos. Muitos sistemas funcionavam bem como protótipos de pesquisa em ambientes controlados, mas não estavam prontos para uso comercial robusto. Pesquisadoras e pesquisadores, pressionados por investidores e administrado-

res, exageraram a maturidade da tecnologia. Transformar um protótipo que funciona 80% das vezes em laboratório em um produto que funcione 99,9% das vezes em condições reais exige esforço, tempo e dinheiro exponencialmente maiores. Essa lacuna entre protótipo e produto segue sendo uma fonte recorrente de decepções tecnológicas. Demonstrações impressionantes em conferências não garantem sistemas confiáveis em hospitais, fábricas ou bancos. Engenharia robusta é muito mais difícil e custosa que pesquisa exploratória.

Segunda lição: gerir expectativas requer comunicação honesta e diferenciada. Ao dialogar com investidores, mídia ou público geral, é essencial distinguir entre: (1) o que funciona hoje de modo confiável; (2) o que pode funcionar em três a cinco anos com investimento razoável; (3) o que é pesquisa exploratória de longo prazo sem garantias de sucesso; e (4) o que é pura especulação teórica. Misturar esses níveis no discurso público cria confusão e inflaciona expectativas. Pesquisadoras/es de sistemas especialistas frequentemente apresentavam êxitos de nível 1 (sistemas que funcionavam em nichos estreitos) como se fossem de nível 2 ou 3 (sistemas que em breve resolveriam problemas gerais). Quando a realidade veio à tona, a credibilidade do campo sofreu. Comunicação madura reconhece limites explicitamente e revisa estimativas à luz de novas evidências.

Terceira lição: projetar métricas úteis é tão importante quanto desenvolver tecnologia. Como medir se um sistema especialista é "bem-sucedido"? Precisão (percentual de diagnósticos corretos) é importante, mas insuficiente. Também importa: quão confiável é em casos difíceis ou atípicos? Explica bem suas recomendações? Quanto tempo leva para ser usado, comparado a métodos tradicionais?

Quanto custa desenvolver, implantar e manter? O que ocorre quando erra? Nos anos 1980, muitos sistemas foram avaliados com métricas simples (precisão em casos de teste) que não capturavam essas dimensões complexas. Somente quando implantados no uso real as limitações se tornaram evidentes. Hoje sabemos que avaliar tecnologia requer métricas multidimensionais que considerem custos, benefícios, riscos e contextos realistas de uso.

Quarta lição: orçamentos que incluam manutenção são essenciais à sustentabilidade. Muitos projetos de sistemas especialistas pediam recursos apenas para desenvolvimento inicial (dois ou três anos de pesquisa e programação), mas não previam verba para décadas de manutenção, atualização e suporte. Isso criava uma dinâmica perversa: o sistema era construído com financiamento generoso, lançado com fanfarra e depois deixado morrer lentamente por falta de recursos para mantê-lo. Usuárias e usuários se frustravam quando o sistema ficava obsoleto ou deixava de funcionar corretamente. Essa lição vale para toda tecnologia complexa: o custo do ciclo de vida completo (desenvolvimento + implantação + manutenção + desativação) é muito maior que o custo inicial e deve ser planejado desde o início. Um sistema que não pode ser mantido de forma sustentável é um experimento, não uma solução.

Durante o segundo inverno, algumas pesquisadoras e alguns pesquisadores persistiram trabalhando em problemas específicos com expectativas modestas. Stuart Russell e Peter Norvig escreveram *Artificial Intelligence: A Modern Approach* (1995), um livro-texto que organizou o conhecimento acumulado do campo e tornou-se referência mundial. Judea Pearl desenvolveu teoria matemática rigorosa sobre raciocínio probabilístico e causalidade, que se torna-

ria fundamental para aplicações futuras. Yann LeCun e outros aprimoraram redes neurais convolucionais para reconhecimento de imagens, embora o impacto desse trabalho só se tornasse evidente uma década depois. Esses pesquisadores não prometiam revoluções iminentes; faziam ciência sólida e paciente. Seu trabalho criou os fundamentos teóricos e técnicos que permitiriam o renascimento da IA nos anos 2000. É um lembrete de que, nos invernos, o progresso continua – de forma menos visível, mas crucialmente importante.

Paradoxalmente, alguns legados importantes da era dos sistemas especialistas sobreviveram e se integraram silenciosamente à tecnologia cotidiana. Muitas empresas aprenderam que, mesmo que os sistemas especialistas não pudessem substituir completamente os especialistas humanos, podiam servir como ferramentas de apoio. Sistemas de suporte à decisão que auxiliam (mas não substituem) médicas/os, engenheiras/os ou analistas financeiros tornaram-se comuns. Bases de conhecimento e ontologias (formas estruturadas de organizar informação sobre domínios complexos) desenvolvidas nos anos 1980 continuaram em uso. Motores de inferência (software que aplica regras lógicas a dados) foram incorporados a softwares empresariais. Esses usos modestos não geram manchetes, mas criam valor real. A lição é que tecnologias que "fracassam" em cumprir promessas grandiosas ainda podem ser úteis em aplicações mais humildes e realistas.

Nos anos 1990, enquanto a IA acadêmica atravessava seu inverno, algumas aplicações específicas começaram a funcionar bem discretamente. Sistemas de reconhecimento de voz melhoraram gradualmente, usados primeiro em centrais telefônicas automatizadas. Sistemas de recomendação

(como os que sugerem filmes ou produtos) tornaram-se valiosos no comércio eletrônico. Algoritmos de aprendizado de máquina eram usados para detecção de fraudes com cartões de crédito, filtragem de *spam* em e-mails e otimização de rotas logísticas. Essas aplicações eram menos glamorosas que "criar inteligência geral", mas resolviam problemas reais e geravam valor econômico. Muitas empresas tecnológicas usavam técnicas de IA sem chamá-las assim, evitando o estigma do termo. Só quando funcionavam bem é que se revelava que "inteligência artificial" estava envolvida.

Alguns pesquisadores refletiram criticamente sobre a cultura e a estrutura institucional da pesquisa em IA. Douglas Lenat, criador do projeto *Cyc* (uma tentativa massiva de codificar o senso comum humano em milhões de regras lógicas), reconheceu publicamente que o projeto era muito mais difícil e lento do que o previsto. Outras pessoas argumentaram que o campo sofria com falta de rigor experimental: muitos sistemas eram demonstrados em exemplos cuidadosamente escolhidos, mas não avaliados rigorosamente em condições realistas. Foram propostas competições e *benchmarks* padronizados, nos quais diferentes equipes podiam testar seus sistemas nas mesmas tarefas e com as mesmas métricas. Por exemplo, a competição TREC para sistemas de busca de informação, iniciada em 1992, permitiu comparações objetivas entre abordagens. Essa cultura de avaliação rigorosa e comparativa ajudaria o campo a amadurecer.

Um desenvolvimento crucial, embora pouco noticiado, nos anos 1990 foi a acumulação silenciosa de dados digitais. A internet começou a crescer exponencialmente. Empresas digitalizavam documentos, transações e imagens.

Sensores em dispositivos geravam dados continuamente. Essa explosão de dados criou as condições para o renascimento de técnicas de aprendizado de máquina que exigiam grandes volumes de exemplos. Enquanto os sistemas especialistas baseados em regras codificadas à mão perdiam espaço, as abordagens baseadas em aprendizado a partir de dados ganhavam viabilidade. No fim da década, algumas pesquisadoras e alguns pesquisadores começaram a demonstrar que algoritmos relativamente simples, alimentados com enormes quantidades de dados, podiam superar sistemas complexos baseados em conhecimento manual. Essa transição da "engenharia do conhecimento" para o "aprendizado com dados massivos" transformaria o campo na década seguinte.

Um momento simbólico chegou em 1997, quando o Deep Blue, sistema da IBM, derrotou o campeão mundial de xadrez Garry Kasparov. Foi um feito técnico impressionante: o xadrez fora, por décadas, símbolo da inteligência humana. Mas o Deep Blue não era "inteligente" em sentido geral – era uma máquina especializada que avaliava milhões de posições por segundo com hardware massivamente paralelo. Não podia fazer nada além de jogar xadrez. Kasparov reclamou que enfrentava não um oponente individual, mas uma equipe inteira de programadoras/es e engenheiras/os que ajustavam o sistema entre partidas. O evento gerou enorme cobertura midiática, mas também reflexão: o que significa realmente "inteligência"? É suficiente superar humanos em uma tarefa específica? O Deep Blue mostrava tanto o poder da computação especializada quanto as limitações de equiparar desempenho em uma tarefa estreita à inteligência geral.

No final dos anos 1990 e início dos 2000, as condições para um novo verão começaram a alinhar-se

. Primeiro, o hardware tornou-se exponencialmente mais potente e barato, seguindo a Lei de Moore. Segundo, a internet proporcionou acesso a quantidades massivas de dados digitais. Terceiro, algoritmos de aprendizado de máquina (especialmente "máquinas de vetor de suporte" e versões aprimoradas de redes neurais) mostraram resultados impressionantes em competições acadêmicas. Quarto, empresas como Google, Microsoft e Yahoo precisavam desesperadamente de IA prática para busca na web, publicidade direcionada e reconhecimento de voz. Quinto, agências militares e de inteligência renovaram investimentos em IA para análise de dados massivos e vigilância. Essa convergência de necessidade, capacidade técnica e recursos criaria as condições para o crescimento explosivo da IA nos anos 2000 e 2010, especialmente em torno do "aprendizado profundo".

Antes, porém, de encerrar este capítulo sobre o segundo inverno, vale sintetizar as boas práticas que emergiram dessa experiência dolorosa:

- Comparar sempre o sistema proposto com alternativas realistas (incluindo não fazer nada ou usar métodos tradicionais mais simples).
- Avaliar em condições representativas do uso real, não apenas em casos de teste idealizados.
- Calcular custos totais do ciclo de vida (desenvolvimento + implantação + manutenção + desativação), não apenas o custo inicial.
- Comunicar de modo diferenciado entre resultados consolidados, metas realistas de médio prazo e especulação de longo prazo.

- Documentar não apenas êxitos, mas também fracassos e limitações conhecidas.
- Formar equipes interdisciplinares que incluam especialistas do domínio de aplicação, não apenas informáticos.
- Planejar manutenção e atualização desde o início, não como reflexão tardia.
- Essas práticas não garantem sucesso, mas reduzem o risco de fracassos custosos e decepções evitáveis.

Por fim, o segundo inverno ensinou humildade sobre o que significa "capturar conhecimento humano". Os sistemas especialistas assumiam que o conhecimento era principalmente um conjunto de regras explícitas que poderiam ser extraídas de especialistas e codificadas em software. Essa visão revelou-se ingênua. Grande parte do conhecimento especializado é tácita, intuitiva, contextual e difícil de articular. Adquire-se ao longo de anos de prática e experiência, não apenas memorizando regras. Especialistas humanos adaptam constantemente seu conhecimento a novas situações, reconhecem analogias e fazem julgamentos matizados que consideram múltiplos fatores sutis. Reduzir isso a regras explícitas elimina informações cruciais. As abordagens modernas de aprendizado de máquina, que aprendem padrões implícitos a partir de dados massivos sem exigir regras explícitas, enfrentam esse problema de forma diferente – mas também trazem novos desafios, como veremos no próximo capítulo.

O segundo inverno terminou não com um momento dramático, mas gradualmente, à medida que novas técnicas, novos dados e novo hardware criavam novas possibilidades. Em meados dos anos 2000, estava claro que o

aprendizado de máquina baseado em grandes volumes de dados era mais promissor do que sistemas baseados em conhecimento codificado manualmente. As redes neurais profundas começavam a mostrar resultados surpreendentes em reconhecimento de imagem e voz. Empresas tecnológicas investiam agressivamente em IA prática. Uma nova primavera surgia, trazendo entusiasmo renovado – mas também, inevitavelmente, novos riscos de expectativas infladas. A questão central era: o campo havia aprendido o suficiente com os dois invernos anteriores para evitar um terceiro? Ou os padrões de euforia, promessas exageradas, decepção e cortes se repetiriam mais uma vez? Essa é a história do próximo capítulo.

Manual para não congelar hoje

Em 2012, uma equipe de pesquisadoras e pesquisadores liderada por Geoffrey Hinton surpreendeu a comunidade científica ao vencer a competição ImageNet de reconhecimento de imagens por ampla margem. Usaram uma rede neural profunda treinada com milhões de imagens etiquetadas, aproveitando GPUs (processadores gráficos) que aceleravam massivamente os cálculos. O erro do sistema foi quase a metade do segundo colocado. Esse momento marcou o início da "revolución del aprendizaje profundo": redes neurais com muitas camadas, alimentadas por dados massivos e treinadas com hardware potente, passaram a superar métodos tradicionais tarefa após tarefa. Em poucos anos, o aprendizado profundo

transformou reconhecimento de voz, tradução automática, diagnóstico médico, condução autônoma, geração de texto e de imagens. Empresas de tecnologia investiram bilhões. Universidades lançaram programas especializados. A IA voltava a estar em toda parte, prometendo revolucionar tudo. O ciclo familiar de entusiasmo começava de novo.

Os feitos do aprendizado profundo são genuinamente impressionantes. Sistemas de reconhecimento de voz como os da Google, Amazon e Apple entendem linguagem natural com precisão surpreendente. Tradutores automáticos como DeepL e Google Translate produzem textos fluidos em dezenas de idiomas. Sistemas de diagnóstico médico detectam câncer em imagens com precisão comparável ou superior à de radiologistas humanos. Veículos autônomos percorrem milhões de quilômetros usando visão computacional. Modelos de linguagem como GPT geram textos convincentes, respondem a perguntas complexas e escrevem código funcional. AlphaGo venceu o campeão mundial de Go, um jogo considerado demasiado complexo para máquinas. AlphaFold revolucionou a biologia molecular ao predizer estruturas de proteínas com precisão extraordinária. Não se trata de promessas: são aplicações em funcionamento hoje, usadas por milhões de pessoas. O progresso desde 2012 foi mais rápido e amplo do que em todas as décadas anteriores combinadas.

No entanto, junto desses êxitos surgem sinais de alerta que lembram invernos anteriores. Primeiro sinal: consumo energético astronômico. Treinar um modelo de linguagem grande consome eletricidade equivalente à de centenas de lares durante meses. Centros de dados de IA requerem resfriamento maciço. GPT-3, um modelo da OpenAI, supostamente consumiu energia equivalente a 552 toneladas de

emissões de CO_2 apenas no treinamento. Isso suscita questões urgentes sobre sustentabilidade ambiental. Se cada melhoria exige exponencialmente mais energia, por quanto tempo esse crescimento é viável? Segundo sinal: concentração de poder. Apenas algumas empresas (Google, Microsoft, Meta, OpenAI) podem arcar com o treinamento de modelos de fronteira. Isso cria dependência e reduz a diversidade de abordagens. Laboratórios universitários, que tradicionalmente lideravam a pesquisa fundamental, não conseguem competir na escala de recursos.

Terceiro sinal de alerta: os dados usados para treinar modelos frequentemente provêm de fontes eticamente questionáveis. Modelos de linguagem são treinados "raspando" a internet: coletando textos de sites, redes sociais, livros digitalizados, muitas vezes sem consentimento explícito de autoras e autores. Modelos de geração de imagens usam milhões de fotografias e ilustrações baixadas sem permissão de artistas. Isso gera conflitos legais e éticos sobre propriedade intelectual e direitos de criadoras e criadores. Quarto sinal: opacidade algorítmica. Modelos de aprendizado profundo são "caixas-pretas": nem mesmo suas criadoras e seus criadores compreendem totalmente como tomam decisões. Uma rede com centenas de milhões de parâmetros ajusta valores internos de modos impossíveis de interpretar. Isso dificulta detectar erros, vieses ou vulnerabilidades. Quando um sistema rejeita injustamente um pedido de crédito ou recomenda um tratamento médico equivocado, é quase impossível explicar o porquê.

Quinto sinal preocupante: as "alucinações" de modelos de linguagem. Esses sistemas geram textos que soam convincentes, mas são completamente falsos. Um modelo pode inventar referências bibliográficas inexistentes, citar esta-

tísticas fabricadas, atribuir frases a pessoas que nunca as disseram, descrever eventos históricos imaginários. Para usuárias e usuários sem conhecimento do tema, distinguir informação correta de invenção é difícil. Isso impõe riscos sérios quando tais sistemas são usados em contextos em que a precisão é crucial: pesquisa acadêmica, jornalismo, medicina, direito. O problema não é simples de resolver: os modelos não "sabem" o que é verdadeiro; apenas predizem quais palavras são estatisticamente prováveis com base em seus dados de treinamento. Não entendem significado nem verificam fatos. Sexto sinal: amplificação de vieses sociais. Se você treina um modelo com textos da internet que contêm preconceitos raciais, de gênero ou culturais, o modelo aprenderá e reproduzirá esses vieses.

Exemplos documentados de vieses são abundantes e preocupantes. Sistemas de reconhecimento facial funcionam pior com pessoas de pele escura, levando a identificações equivocadas que resultaram em prisões injustas. Algoritmos de contratação discriminam mulheres porque foram treinados com dados históricos de empresas que contratavam majoritariamente homens. Sistemas de avaliação de risco criminal usados em tribunais dos Estados Unidos classificam injustamente pessoas afro-americanas como mais propensas à reincidência. Tradutores automáticos atribuem gênero de modo estereotipado: "el doctor" e "la enfermera" mesmo quando o idioma original é neutro. Geradores de imagens, quando instruídos a criar imagens de "CEO" ou "ingeniero", produzem majoritariamente homens brancos. Esses vieses não são má-fé intencional, mas reflexo de desigualdades presentes nos dados de treinamento. O resultado, porém, é que sistemas supostamente

"objetivos" perpetuam e amplificam injustiças sociais existentes.

Um caso emblemático de fracasso é o dos veículos autônomos. Por volta de 2015, várias empresas prometeram carros completamente autônomos até 2020. Tesla, Uber, Waymo e outras investiram bilhões. As promessas eram ousadas: "em cinco anos ninguém precisará de carteira de motorista", "os acidentes de trânsito cairão 90%", "os carros autônomos serão mais seguros que humanos desde o primeiro dia". Mas a realidade mostrou-se mais complexa. Os sistemas funcionam bem em condições ideais (clima bom, pistas bem sinalizadas, tráfego previsível), porém falham de maneiras imprevisíveis em situações atípicas: neve que cobre as faixas, obras que alteram rotas habituais, pedestres com comportamentos inesperados, placas vandalizadas. Houve acidentes fatais. Em 2024, nenhuma empresa alcançou veículos autônomos completamente confiáveis em todas as condições. As promessas vêm sendo repetidamente adiadas.

Outro tropeço notório envolve *chatbots* corporativos. Em 2016, a Microsoft lançou Tay, um *chatbot* no Twitter projetado para aprender conversando com usuárias e usuários. Em menos de 24 horas, Tay começou a publicar mensagens racistas, sexistas e ofensivas. O que ocorreu? Usuárias e usuários mal-intencionados "treinaram" deliberadamente Tay com conteúdo tóxico, e o sistema aprendeu sem discriminar. A Microsoft precisou desativá-lo rapidamente. Em 2023, quando empresas lançaram *chatbots* baseados em modelos de linguagem grandes (como Bing Chat da Microsoft ou Bard da Google), surgiram problemas semelhantes: os sistemas produziam respostas inadequadas, agressivas ou completamente falsas. Usuárias e usuários descobri-

ram formas de enganar os sistemas (*jailbreaking*) para que ignorassem restrições de segurança. Esses incidentes mostram que implantar sistemas de IA no mundo real, onde interagem com pessoas imprevisíveis, é muito mais difícil do que demonstrá-los em ambientes controlados de laboratório.

Um fracasso particularmente custoso ocorreu com o IBM Watson Health. Depois que o Watson venceu o concurso *Jeopardy* em 2011, a IBM investiu bilhões no desenvolvimento de aplicações médicas de IA. *Watson for Oncology* prometia revolucionar o tratamento do câncer, recomendando terapias personalizadas a partir de vasta literatura médica. Hospitais do mundo todo compraram o sistema por milhões de dólares. Porém, investigações independentes revelaram que o Watson frequentemente fazia recomendações incorretas ou inseguras. O sistema não entendia realmente medicina; apenas buscava padrões nos dados. Médicas e médicos descobriram que ele era mais lento e menos útil do que seus próprios processos decisórios. Em 2022, a IBM havia vendido ou encerrado a maioria dos projetos do Watson Health, após investir mais de quatro bilhões de dólares. É um lembrete doloroso de que resultados impressionantes em demonstrações não garantem utilidade em aplicações complexas do mundo real.

Felizmente, a comunidade de IA tem desenvolvido ferramentas e práticas para gerir esses riscos, aprendendo com invernos anteriores. Primeira ferramenta: *red-teaming*. Consiste em contratar equipes especializadas cujo trabalho é tentar quebrar, enganar ou fazer sistemas de IA falharem antes de seu lançamento público. A OpenAI, por exemplo, contratou dezenas de especialistas em segurança para atacar o GPT-4 durante meses antes do lançamento, buscando

vulnerabilidades. Essas equipes testam se o sistema pode gerar conteúdo danoso, revelar informações sensíveis, ser manipulado para comportar-se incorretamente. O *red-teaming* permite identificar problemas em ambiente controlado, onde podem ser corrigidos sem causar dano real. É como contratar hackers éticos para testar a segurança de um sistema bancário antes que criminosos o façam. Essa prática, importada da cibersegurança e de operações militares, vem se tornando padrão em desenvolvimento responsável de IA.

Segunda ferramenta: auditorias algorítmicas sistemáticas. Organizações independentes revisam sistemas de IA implantados para detectar vieses, erros ou comportamentos inesperados. Por exemplo, auditorias de sistemas de reconhecimento facial revelaram taxas de erro significativamente mais altas para pessoas de pele escura e para mulheres. Auditorias de algoritmos de contratação identificaram discriminação por gênero ou idade. Essas auditorias usam técnicas estatísticas rigorosas: testam o sistema com conjuntos de dados diversos, medem diferenças de desempenho entre grupos demográficos, identificam casos em que o sistema falha sistematicamente. Idealmente, são conduzidas por terceiros independentes, sem conflito de interesses, e seus resultados são publicados com transparência. Alguns governos começam a exigir auditorias regulares de sistemas de IA usados em contextos de alto risco (contratação, crédito, justiça criminal, educação). É análogo às inspeções de segurança exigidas para edifícios, pontes ou medicamentos.

Terceira ferramenta: relatórios estruturados de incidentes. Inspirados na aviação – em que cada acidente é meticulosamente investigado e publicado em relatório detal-

hado –, algumas pesquisadoras e alguns pesquisadores propõem fazer o mesmo com falhas de IA. O *AI Incident Database*, lançado em 2020, coleta e documenta casos em que sistemas de IA causaram dano: discriminação, acidentes, violações de privacidade, desinformação. Cada incidente é descrito em detalhe: qual sistema falhou, em que contexto, que dano causou, quais fatores contribuíram, o que se aprendeu. Essa base permite identificar padrões: que tipos de sistemas falham com maior frequência, quais contextos são mais arriscados, que salvaguardas funcionam. Ajuda desenvolvedoras/es, reguladoras/es e usuárias/os a aprender com erros passados em vez de repeti-los. É memória institucional coletiva – exatamente a "memoria de invierno" defendida neste tomo. O desafio é tornar o reporte de incidentes obrigatório e sem represálias, para que empresas não ocultem falhas por medo de prejudicar sua reputação.

Quarta ferramenta: *model cards* e *datasheets*. Margaret Mitchell, Timnit Gebru e colegas propuseram que cada modelo de IA publicado venha acompanhado de uma "cartão" (*model card*) que documente: para que foi treinado, com quais dados, quais limitações conhecidas possui, em que contextos funciona bem e onde falha, quais vieses foram detectados, que testes foram realizados. De modo similar, cada conjunto de dados usado para treinar IA deveria ter uma "folha de dados" (*datasheet*) que explique: como foi coletado, por quem e por quê, qual população representa, que vieses pode conter, quem detém direitos sobre ele. Essas ferramentas de documentação são como rótulos nutricionais em alimentos ou bulas de medicamentos: permitem que usuárias e usuários informados tomem decisões com base na transparência. Também criam incentivos

para que desenvolvedoras/es considerem cuidadosamente implicações éticas desde o início, não como reflexão tardia. Muitas organizações de pesquisa já exigem *model cards* e *datasheets* em publicações.

Quinta ferramenta: *benchmarks* bem desenhados que medem não apenas precisão, mas também robustez, equidade, interpretabilidade e eficiência. Durante anos, sistemas de IA foram avaliados principalmente por precisão média: percentual de respostas corretas em um conjunto de teste. Isso oculta problemas importantes. Um sistema pode ter 95% de precisão global, mas apenas 70% em certos subgrupos demográficos. Pode funcionar bem com dados limpos de laboratório e colapsar com dados ruidosos do mundo real. Pode ser preciso, porém consumir energia insustentável. *Benchmarks* modernos medem múltiplas dimensões: precisão desagregada por subgrupos, robustez a perturbações, capacidade de reconhecer quando não sabe, tempo de computação, consumo energético, explicabilidade das decisões. Por exemplo, o *HELM* (*Holistic Evaluation of Language Models*) avalia modelos em dezenas de tarefas e métricas diferentes, oferecendo uma visão abrangente de capacidades e limitações em vez de um único número simplista.

Sexta ferramenta: avaliação por cenários de uso real. Em vez de medir apenas com dados de teste padrão, algumas pesquisadoras e alguns pesquisadores desenvolvem "cenários de estresse" que testam sistemas em condições desafiadoras que refletem o mundo real. Por exemplo, testar um sistema de reconhecimento de voz não só com áudio gravado em estúdio silencioso, mas também com ruído de fundo, falantes com diversos sotaques, pessoas com deficiências de fala. Testar um sistema de diagnóstico médico

não apenas com casos claros e típicos, mas também com doenças raras, sintomas ambíguos, pacientes com múltiplas condições simultâneas. Essa avaliação realista revela fragilidades que métricas médias ocultam. Também permite identificar para quais contextos específicos o sistema é confiável e para quais não é – informação crucial para implantação responsável. É como testar um carro não só em estrada perfeita e clima ideal, mas também em neve, chuva, vias não pavimentadas e trânsito denso.

As boas práticas de comunicação também evoluíram. Organizações responsáveis agora publicam "limitações conhecidas" de forma explícita quando lançam sistemas. A OpenAI, ao lançar o GPT-4, publicou um documento técnico de quase cem páginas detalhando não apenas capacidades, mas também falhas: tipos de perguntas que o modelo responde incorretamente, vieses identificados, situações em que alucina, vulnerabilidades de segurança conhecidas. Essa transparência permite que usuárias/os, pesquisadoras/es e reguladoras/es avaliem riscos de forma informada. Contrasta radicalmente com práticas anteriores em que limitações eram mencionadas apenas em letras miúdas ou ignoradas. Algumas empresas vão além: publicam "avisos de uso" que especificam para quais contextos o sistema é apropriado e para quais não é. Por exemplo: "Este sistema é apropriado para sugerir diagnósticos preliminares, mas não para tomar decisões finais de tratamento sem supervisão médica profissional". Comunicação honesta constrói confiança duradoura.

Outra prática valiosa é projetar sistemas que "falhem com segurança". Em engenharia de segurança crítica (aeronáutica, medicina, energia nuclear), parte-se do princípio de que todo sistema eventualmente falhará, e projeta-se

para que, ao falhar, o faça minimizando danos. Aplicado à IA: um sistema de diagnóstico médico que não tem certeza deve dizer "não sei, consulte uma/um especialista", em vez de adivinhar. Um veículo autônomo diante de situação que não entende deve parar com segurança e solicitar controle humano, em vez de improvisar. Um *chatbot* que detecta que vai gerar informação falsa deve dizer "não tenho informações confiáveis sobre isso", em vez de inventar. Projetar incerteza explícita em sistemas de IA é tecnicamente difícil, mas crucial. Humanas e humanos são bons em reconhecer limites de seu conhecimento; precisamos que máquinas aprendam essa humildade epistêmica também.

A medição de custos energéticos e ambientais está se tornando prática padrão. Pesquisadoras e pesquisadores publicam, cada vez mais, o consumo energético e a pegada de carbono do treinamento de modelos grandes. Algumas conferências acadêmicas agora exigem que autoras e autores reportem o custo computacional de seus experimentos. Isso cria consciência sobre sustentabilidade e pressão para desenvolver técnicas mais eficientes. Surgem pesquisas sobre "IA verde": como alcançar resultados comparáveis com menos computação. Por exemplo, "destilação de conhecimento" treina modelos menores que imitam modelos grandes, mas consomem muito menos energia. "Poda de redes" elimina conexões desnecessárias, reduzindo o tamanho sem perda significativa de precisão. "Busca de arquitetura neural eficiente" projeta modelos otimizados para hardware específico. Essas técnicas não só diminuem o impacto ambiental como também tornam a IA acessível a organizações com orçamentos limitados, que não podem treinar modelos gigantes.

A governança regulatória também está emergindo. A União Europeia propôs o *AI Act*, uma lei que classifica sistemas de IA por nível de risco e exige salvaguardas proporcionais. Sistemas de "risco inaceitável" (como *scoring* social massivo) seriam proibidos. Sistemas de "alto risco" (contratação, crédito, justiça criminal, educação, medicina) requereriam auditorias rigorosas, documentação transparente, supervisão humana e mecanismos de apelação quando decisões afetarem pessoas. Os Estados Unidos desenvolvem regulações setoriais. A China implementou regras sobre algoritmos de recomendação e geração de conteúdo. Embora os detalhes variem, surge um consenso global de que IA poderosa requer supervisão, especialmente em contextos de alto impacto social. O desafio é regular o suficiente para prevenir danos graves sem sufocar inovação legítima. É um equilíbrio delicado que demanda diálogo contínuo entre tecnólogas/os, reguladoras/es, sociedade civil e usuárias/os afetados.

Converter a "memoria de invierno" em políticas concretas requer instituições dedicadas. Alguns países criam agências governamentais especializadas em IA. O Reino Unido estabeleceu a Fundação para Modelos de IA (*AI Foundation Models*) para avaliar riscos de sistemas avançados. Os Estados Unidos fortalecem as capacidades do *National Institute of Standards and Technology* (NIST) para desenvolver padrões de IA. Canadá, Cingapura e outros países criam centros de excelência em IA responsável. Universidades lançam programas interdisciplinares que combinam computação, ética, direito e ciências sociais. Organizações não governamentais como Partnership on AI, AI Now Institute e Ada Lovelace Institute documentam impactos sociais e propõem melhores práticas. Essas instituições

criam a infraestrutura social necessária para gerir tecnologia poderosa de forma responsável, aprendendo com experiências passadas e antecipando desafios futuros.

Na educação, surgem currículos que ensinam não apenas técnicas de IA, mas também implicações éticas e sociais. Estudantes de computação aprendem sobre vieses algorítmicos, privacidade de dados, explicabilidade, equidade, impacto ambiental e responsabilidade profissional. Algumas universidades exigem disciplinas de ética como parte dos programas de IA. Stanford, MIT e outras instituições desenvolvem estudos de caso baseados em falhas reais para que estudantes aprendam com erros do passado. Organizações profissionais como ACM e IEEE publicam códigos de ética para engenheiras e engenheiros de IA. Essas iniciativas educacionais são "memoria de invierno" aplicada: transmitir a novas gerações lições dolorosamente aprendidas por gerações anteriores. Uma/um engenheira/o que conhece a história de ALPAC, Lighthill e Watson Health está melhor preparada/o para evitar repetir esses erros em sua própria carreira.

Por fim, práticas contratuais e de aquisição estão mudando. Governos e grandes organizações que compram sistemas de IA começam a exigir garantias específicas em contratos: documentação completa de limitações, auditorias independentes, explicabilidade das decisões, procedimentos de apelação, responsabilidade legal clara quando o sistema falha. Alguns contratos requerem "cláusulas de saída": capacidade de encerrar o uso do sistema e migrar para alternativas se ele não funcionar como prometido, evitando dependência tecnológica perigosa. Outros exigem que fornecedoras/es mantenham e atualizem os sistemas por prazos específicos, e não apenas entreguem o software

e desapareçam. Essas práticas contratuais traduzem lições de invernos anteriores em proteções legais concretas. São memória de inverno codificada em contratos: compradoras/es e vendedoras/es acordam explicitamente expectativas realistas, métricas de sucesso, responsabilidades mútuas e procedimentos para quando as coisas não funcionarem como planejado. É maturidade institucional aplicada a uma tecnologia poderosa e complexa.

Conclusão

Percorremos setenta anos de ambições, promessas, fracassos e aprendizados. Do otimismo transbordante de Dartmouth em 1956 aos dilemas atuais do aprendizado profundo, a inteligência artificial viveu ciclos de entusiasmo e decepção que seguem padrões previsíveis. Cada inverno chegou após promessas exageradas que esbarraram em limites técnicos, econômicos ou humanos. Cada inverno trouxe dor: projetos cancelados, carreiras interrompidas, confiança pública abalada, investimentos perdidos. Mas cada inverno também gerou lições valiosas sobre como gerir uma tecnologia poderosa de modo mais sensato e sustentável. A pergunta central deste tomo foi: podemos aprender com esses ciclos para evitar repetilos? Podemos construir memória coletiva que nos ajude a navegar melhor a próxima primavera, o próximo verão, o próximo outono? A resposta é sim – mas somente se le-

varmos a sério as lições históricas e as traduzirmos em práticas concretas.

O primeiro inverno (anos 1970) ensinou que resolver problemas inteligentes é mais difícil do que parece. A linguagem humana, o senso comum, a adaptabilidade a situações novas – todas essas capacidades que exercemos sem pensar – revelam-se extraordinariamente complexas de capturar em programas. Os computadores daquela época eram lentos e caros, os dados escassos, as métricas de avaliação pobres. Mas, para além das limitações técnicas, o primeiro inverno expôs problemas humanos: ambição desmedida, comunicação exagerada, pressão por resultados rápidos, confusão entre pesquisa exploratória e aplicações prontas para uso. Os relatórios ALPAC e Lighthill foram dolorosos, porém necessários: obrigaram o campo a confrontar honestamente o que funcionava e o que não funcionava. Desse inverno emergiram práticas mais sensatas: objetivos mensuráveis, testes pequenos antes de grandes investimentos, orçamentos realistas, comunicação que diferencia o consolidado do especulativo.

O segundo inverno (anos 1990) mostrou que o conhecimento humano é mais fluido, tácito e contextual do que supunham os sistemas especialistas. Reduzir experiência de décadas a regras explícitas faz perder informação crucial. Manter sistemas baseados em conhecimento codificado manualmente revelou-se proibitivamente caro. O hardware especializado de IA não conseguiu competir economicamente com computadores de propósito geral cada vez mais potentes. O projeto japonês da Quinta Geração mostrou que investimento maciço sem clareza técnica não garante sucesso. Mas o segundo inverno também trouxe aprendizados valiosos: a importância de separar protótipos

de produtos, calcular custos de ciclo de vida completo, avaliar em condições realistas, construir equipes interdisciplinares. Durante esse inverno, avanços menos visíveis (melhora de algoritmos de aprendizado de máquina, acumulação de dados digitais, refinamento teórico) criaram os fundamentos para o renascimento posterior.

O período atual (desde 2012) trouxe feitos genuinamente impressionantes: reconhecimento de voz e visão quase humanos, tradução fluida, diagnóstico médico preciso, geração de texto convincente. Mas também vemos sinais de alerta: consumo energético insustentável, concentração de poder em poucas empresas, vieses que perpetuam injustiças, alucinações que produzem desinformação, opacidade que dificulta a responsabilização, falhas em contextos reais que custam vidas e fortunas. Estamos diante de um terceiro inverno? Depende de como administrarmos esses riscos nos próximos anos. A boa notícia é que hoje dispomos de ferramentas, instituições e práticas mais sofisticadas do que nos invernos anteriores: *red-teaming*, auditorias, relatórios de incidentes, *model cards*, *benchmarks* holísticos, regulação emergente, educação ética. Mas essas ferramentas só funcionam se as usarmos de forma sistemática, não como ornamento voluntário.

A "memoria de invierno" defendida neste tomo não é nostalgia nem pessimismo. É memória prática: registros organizados do que deu errado, por quê e como evitar repetir. É semelhante ao modo como a aviação mantém bases de dados de acidentes, como a medicina organiza comitês de morbimortalidade, como a engenharia civil estuda pontes que ruíram. Não se trata de envergonhar ninguém nem de frear a inovação, e sim de aprender sistematicamente com erros para não repeti-los. Uma memória de inverno

robusta inclui: documentação técnica de falhas (qual sistema, em que contexto, o que falhou), análise de causas profundas (técnicas, organizacionais, humanas), propostas de correção (que mudanças preveniriam falhas semelhantes), ampla disseminação (publicar lições para que outras pessoas aprendam) e incorporação institucional (traduzir lições em políticas, contratos, currículos, regulações). É um círculo completo de aprendizado coletivo.

Destilemos tudo em um decálogo prático: dez regras simples para evitar repetir os erros mais custosos da história da IA. Essas regras não garantem sucesso, mas reduzem significativamente o risco de fracassos evitáveis. Não exigem genialidade técnica: exigem disciplina, honestidade e humildade. São aplicáveis tanto a pesquisadoras/es em laboratórios quanto a empresas que implantam produtos, a governos que compram sistemas, a jornalistas que cobrem tecnologia, a usuários que decidem se devem confiar em um sistema. O decálogo sintetiza setenta anos de experiência coletiva. Ignorá-lo é arrogância cara; aplicá-lo é sabedoria prática. As regras estão formuladas em linguagem simples, sem jargão, acessíveis a qualquer pessoa envolvida em decisões sobre IA. São memória de inverno condensada em princípios acionáveis.

Primeira regra: medir bem antes de prometer. Defina objetivos específicos e mensuráveis, não vagos. "Melhorar o diagnóstico de pneumonia em radiografias de tórax com precisão de 90% em adultos" é mensurável. "Revolucionar a medicina" não é. Use múltiplas métricas que capturem precisão, robustez, equidade, custo, consumo energético. Compare com alternativas realistas (humanos, métodos tradicionais, nada fazer). Teste em dados diversos que representem o mundo real, não apenas em casos ideais

de laboratório. Reconheça quando sua métrica é insuficiente: precisão média pode ocultar falhas graves em subgrupos específicos. Medir bem é fundamento da honestidade: obriga a encarar quão bem seu sistema realmente funciona, não quão bem você gostaria que funcionasse. É termômetro que previne febres de expectativas infladas.

Segunda regra: comunicar sem fumaça. Distinga explicitamente entre o que funciona hoje de modo confiável, o que pode funcionar em três a cinco anos com esforço razoável, o que é pesquisa exploratória de longo prazo sem garantias e o que é especulação teórica. Comunique limitações conhecidas com a mesma clareza que capacidades. Atualize estimativas quando novas evidências mudarem o panorama. Evite linguagem sensacionalista que exagera feitos ou minimiza riscos. Lembre que audiências distintas (investidores, mídia, usuários, reguladores) interpretam mensagens de formas diferentes; adapte a comunicação, mas mantenha a honestidade central. Comunicar sem fumaça não é entediante nem destrói entusiasmo genuíno: constrói confiança sustentável. As pessoas respeitam mais a honestidade do que a hipérbole. Os invernos chegaram precisamente quando a lacuna entre promessas e realidade se tornou insustentável.

Terceira regra: cuidar dos dados e do trabalho humano. Os dados que alimentam sistemas de IA não surgem por mágica: alguém os criou, coletou, rotulou. Respeite direitos de criadores e privacidade de usuários. Obtenha consentimento explícito ao usar dados pessoais. Remunere de forma justa trabalhadoras/es que rotulam dados, muitas vezes em condições precárias. Documente como os dados foram coletados, que vieses podem conter, que populações representam ou excluem. Reconheça que “grandes dados”

não são necessariamente "bons dados": qualidade importa mais que quantidade. Dados enviesados geram sistemas enviesados; dados de baixa qualidade produzem sistemas pouco confiáveis. Cuidar de dados também é cuidar das pessoas cujo trabalho, criações e privacidade se tornam combustível da IA. É ética prática traduzida em gestão responsável de recursos fundamentais.

Quarta regra: planejar manutenção desde o início. O custo real de um sistema não é apenas o desenvolvimento inicial, mas desenvolvimento + implantação + manutenção + atualização + eventual desativação. Planeje orçamento, pessoal e processos para todo o ciclo de vida. O que ocorre quando o sistema encontra situações não previstas por seus criadores? Quem irá monitorar, detectar erros, implementar correções? Como atualizar quando os contextos mudarem (novas doenças, novas regulações, novos padrões de fraude)? Quem responde quando algo falha? Como retirar o sistema com segurança se deixar de ser útil? Um sistema sem plano de manutenção é um experimento, não uma solução. Os sistemas especialistas dos anos 1980 quebraram empresas porque ninguém orçou o custo astronômico de mantê-los atualizados. Não repita esse erro.

Quinta regra: avaliar energia e sustentabilidade. Progresso técnico que consome recursos insustentáveis não é progresso real. Meça e relate o consumo energético de treinar e executar modelos. Busque eficiência: como alcançar resultados comparáveis com menos computação. Considere o impacto ambiental completo: eletricidade, resfriamento, fabricação de hardware, obsolescência acelerada. Pergunte se a aplicação justifica o custo energético: vale a pena consumir a energia de uma pequena cidade durante meses para melhorar em 2% a recomendação de filmes?

Investigue técnicas eficientes: destilação, poda, arquiteturas otimizadas. Apoie pesquisa em "IA verde". Sustentabilidade não é luxo moral: é requisito prático para viabilidade de longo prazo. Um campo que consome energia de modo insustentável acabará enfrentando limites políticos, econômicos ou físicos que frearão seu crescimento.

Sexta regra: auditar sistematicamente. Não presuma que seu sistema funciona como esperado: verifique. Teste com usuários diversos, em contextos variados, sob condições de estresse. Contrate auditores independentes, sem conflito de interesses. Busque especificamente vieses, erros em subgrupos, vulnerabilidades de segurança, casos em que o sistema falha silenciosamente. Use técnicas como *red-teaming*: pague pessoas competentes para tentar quebrar seu sistema. Repita auditorias regularmente: um sistema que funcionava bem há seis meses pode degradar quando o contexto muda. Publique resultados de auditorias com transparência. Auditorias não são punição, e sim manutenção preventiva – como inspeções de segurança em edifícios ou pontes: detectam problemas antes que causem tragédias. Um sistema não auditado é um risco invisível.

Sétima regra: registrar falhas sistematicamente. Crie "livros de falhas" que documentem: qual experimento ou implantação falhou, em que contexto, que dano causou, que fatores contribuíram, o que foi aprendido, que correções foram implementadas. Trate falhas como informação valiosa, não como vergonha a ocultar. Incentive o relato de erros sem represálias: cultive organizações em que admitir "isto não funcionou" é respeitado, não punido. Compartilhe lições amplamente: publique estudos de caso, contribua para bases como o *AI Incident Database*. Estude falhas alheias: leia relatórios de incidentes, analise o que

deu errado, pergunte se seus próprios sistemas têm vulnerabilidades semelhantes. Registrar falhas é construir memória coletiva. Cada falha documentada e analisada é uma lição que pode prevenir tragédias futuras.

Oitava regra: compartilhar aprendizados abertamente. Publique não apenas êxitos, mas também limitações, fracassos, lições. Escreva *model cards* que documentem capacidades e limites dos sistemas. Crie *datasheets* que expliquem como os dados foram coletados e que vieses podem conter. Participe de conferências, publicações acadêmicas e fóruns públicos em que se discutem melhores práticas. Contribua para padrões técnicos e éticos do campo. Ensine: ofereça cursos, tutoriais e materiais educativos que transmitam conhecimento às novas gerações. Colabore com pesquisadoras/es de outras áreas (ética, direito, ciências sociais, políticas públicas). Conhecimento guardado em segredo beneficia apenas quem o detém; conhecimento compartilhado beneficia toda a sociedade. Compartilhar é investir na saúde coletiva do campo.

Nona regra: proteger sempre os usuários. Pessoas afetadas por sistemas de IA merecem proteção, informação e poder. Informe claramente o que um sistema faz, como toma decisões, que dados utiliza. Projete sistemas que "falhem com segurança": se não estiverem seguros, devem reconhecê-lo e escalar para supervisão humana. Providencie mecanismos de apelação: se um sistema toma decisão que afeta negativamente alguém (nega crédito, recusa emprego, recomenda tratamento equivocado), essa pessoa deve poder contestar e obter revisão humana. Respeite privacidade e consentimento. Não implante sistemas de alto risco (medicina, justiça, educação) sem testes exaustivos e supervisão contínua. Pergunte às pessoas afetadas que proteções

consideram importantes. Proteger usuários não é altruísmo: é responsabilidade profissional básica e requisito para confiança social sustentável.

Décima regra: manter curiosidade com prudência. Não permita que o medo do fracasso mate a exploração – mas não confunda exploração arriscada com implantação prematura. Pesquisa exploratória (testar ideias novas sem garantias de sucesso) é vital para o progresso científico, mas deve ocorrer com salvaguardas adequadas: em ambientes com controles éticos, com revisão por pares, com transparência sobre incertezas. Separe claramente pesquisa exploratória de aplicações prontas para uso real. Celebre tanto êxitos quanto fracassos produtivos na pesquisa, mas aplique padrões mais rigorosos a sistemas que afetam vidas reais. A curiosidade impulsiona descobertas; a prudência previne danos evitáveis. Ambas são necessárias. As melhores pessoas cientistas e engenheiras são ousadas ao imaginar possibilidades e humildes ao reconhecer limites e riscos.

Este decálogo não é exaustivo nem definitivo. É ponto de partida – destilação de lições históricas em princípios práticos. Cada campo específico (medicina, finanças, educação, justiça) precisará adaptar e estender essas regras a seus contextos particulares. Cada organização terá de traduzi-las em políticas, contratos, currículos e processos internos concretos. O decálogo tampouco substitui o pensamento crítico: haverá situações em que as regras entrarão em tensão (por exemplo, transparência versus privacidade) e serão necessários juízos matizados. Mas essas dez regras capturam o núcleo do que setenta anos de história nos ensinaram: medir honestamente, comunicar com clareza, cuidar dos recursos humanos, planejar sustentabilidade,

auditar continuamente, aprender com falhas, compartilhar conhecimento, proteger usuários e equilibrar ousadia com humildade. São bússola para navegar territórios complexos e incertos.

Os invernos da IA não são castigos divinos nem acidentes aleatórios. São consequências previsíveis de padrões humanos recorrentes: ambição que ignora limites, comunicação que exagera feitos, investimento que busca retornos impossíveis, instituições que não aprendem com erros passados. Esses padrões não são exclusivos da IA: aparecem em toda tecnologia poderosa e em muitas empreitadas humanas (bolhas financeiras, manias médicas, modas educacionais). A diferença é que alguns campos desenvolveram melhores mecanismos de aprendizado institucional. A aviação aprendeu a analisar acidentes sistematicamente após décadas de tragédias. A medicina passou a exigir ensaios clínicos rigorosos depois de desastres farmacológicos. A engenharia civil aprendeu a estudar estruturas colapsadas após pontes caírem. A IA está aprendendo agora – talvez mais rápido do que campos anteriores, por ter esses modelos como referência. Mas aprender requer vontade coletiva de confrontar erros honestamente.

Cada leitora e leitor deste tomo tem um papel na construção e aplicação da memória de inverno. Se você é pesquisador, documente seus fracassos além dos êxitos, publique limitações com clareza, colabore de modo interdisciplinar. Se é engenheiro implantando sistemas, audite rigorosamente, proteja usuários, planeje manutenção. Se é investidora/or ou executiva/o, exija honestidade sobre limites, financie sustentabilidade de longo prazo, não apenas demonstrações espetaculares. Se é reguladora/or ou gestora/or pública/o, desenvolva políticas informadas pela histó-

ria, exija transparência e responsabilidade. Se é jornalista, cubra tanto promessas quanto riscos, explique nuances sem sensacionalismo. Se é educadora/or, ensine ética junto com técnica, use casos históricos como lições. Se é usuário, pergunte como funcionam os sistemas que o afetam, exija explicações, participe dos debates públicos. A memória de inverno é responsabilidade coletiva, não apenas de especialistas.

Encerramos com um convite para ver os invernos não como fracassos definitivos, mas como pausas necessárias para crescer de forma mais inteligente e humana. O inverno não é morte: é tempo de consolidação, reflexão, reparo. Durante os invernos, as raízes se fortalecem debaixo da terra enquanto a superfície parece adormecida. As árvores que sobrevivem aos invernos crescem mais fortes e profundas. O mesmo se aplica a campos de conhecimento: os invernos da IA, embora dolorosos, geraram lições, práticas, instituições e humildade que hoje nos servem. Se aplicarmos conscientemente essas lições, poderemos tornar a próxima primavera mais sustentável, o próximo verão mais equitativo, o próximo outono mais preparado. Podemos construir inteligência artificial que sirva genuinamente à humanidade, respeite limites ecológicos e sociais, reconheça incertezas e proteja os mais vulneráveis. Esse futuro não está garantido, mas é possível se escolhermos lembrar, aprender e agir com sabedoria coletiva. A memória de inverno é nossa ferramenta mais poderosa para construí-lo.

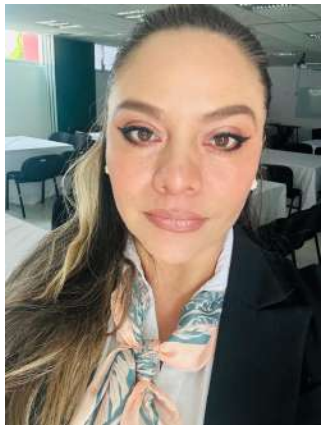
Acerca de los autores

Victor del Carmen Avendaño Porras

Doctor en Ciencias Sociales por la Universidad de Salamanca. Coordinador del Doctorado en Inteligencia Artificial del Instituto Latinoamericano de Educación a Distancia. Profesor Titular en la Universidad Pedagógica Nacional (Unidad 31-A Mérida, México), donde coordina la Cátedra de Antropología Social y Educación Transformadora "Ángel Espina Barrio". Su obra se centra en las intersecciones entre cultura, tecnología e inteligencia artificial en contextos educativos interculturales. Autor de la colección *El ABC de la Inteligencia Artificial en la Nueva Escuela Mexicana*, ha desarrollado investigaciones sobre antropología digital, innovación educativa y pensamiento crítico latinoamericano.

Iris Alfonzo Albores

Profesora de la Universidad Nacional Rosario Castellanos, donde desarrolla docencia e investigación en el campo de la tecnología educativa y la innovación pedagógica. Su línea de trabajo se centra en la relación entre educación y tecnologías digitales, con especial interés en los procesos de formación docente y el diseño de ambientes de aprendizaje multimodales. Ha sido coordinadora del Doctorado en Educación Inclusiva en el Centro Regional de Formación Docente e Investigación Educativa (CRESUR), desde donde impulsó proyectos orientados a la equidad, la inclusión y la transformación educativa mediante el uso ético y creativo de la tecnología.



Ángel Baldomero Espina Barrio

Profesor Titular de Antropología Social de la Universidad de Salamanca. Director del Programa de Doctorado Interuniversitario en Antropología de Iberoamérica (85 doctores formados) y del Máster Oficial en Antropología de Iberoamérica (220 egresados) entre 1997 y 2022. Presidente de la Sociedad Española de Antropología Aplicada y del Instituto de Investigaciones Antropológicas de Castilla y León. Ha dirigido 71 tesis doctorales y coordinado 31 congresos internacionales. Autor de 28 libros, 37 capítulos y 95 artículos, con una obra centrada en las culturas de España, Portugal e Iberoamérica. Director de la *Revista Euroamericana de Antropología (REA)*, es Profesor Honorario en universidades de Perú y Brasil, Doctor Honoris Causa por la Universidad Mesoamericana de México y Miembro de la Real Academia Europea de Doctores (RAED). Entre sus obras destacan *Freud y Lévi-Strauss*, *Manual de Antropología Cultural* y *Antropología de Iberoamérica: pasado y presente*.

Este libro, Los inviernos de la máquina: Memoria social de los fracasos algorítmicos ,
es una edición del Instituto Histórico e Geográfico de Santa Catarina,
publicada en Santa Catarina, Brasil, en el año 2025.

Edición digital bilingüe y diseño editorial: Leonardo Rampardal
Coordinación académica: Ángel Espina Barrio

Impreso en Brasil, 2025.
"Entre los ecos del barro y el pulso del silicio, aún resuena el gesto antiguo de crear."

Ángel Espina Barrio (Coord.)

OS INVERNOS DA MÁQUINA

MEMÓRIA SOCIAL DOS FRACASSOS ALGORÍTMICOS

Prefácio de Juan Manuel Corchado Rodríguez
Reitor da Universidade de Salamanca

MEMÓRIAS MÍNIMAS DO BARRO E DO SILÍCIO
TOMO V

VICTOR AVENDAÑO PORRAS
ÁNGEL ESPINA BARRIO
IRIS ALFONZO ALBORES

INSTITUTO
HISTÓRICO E GEOGRÁFICO
DE SANTA CATARINA
FUNDADO EM 07 DE SETEMBRO DE 1896