La IA y la conciencia

Una consideración desde una perspectiva materialista-dialéctica

por Erik Roderwald

Febrero de 2026

Este documento se distribuye bajo una licencia propia. Véanse las condiciones completas en la última página.

El funcionamiento de la memoria humana desde la perspectiva neurológica

Quien habla de la memoria humana suele pensar en un archivo interior en el que las experiencias se depositan y más tarde se recuperan sin modificaciones. Desde la perspectiva neurológica, esta imagen es fundamentalmente errónea. El cerebro no almacena contenidos en el sentido de objetos terminados, sino que transforma de manera continua su propia estructura. Esta capacidad es denominada por la neurología plasticidad neuronal. Con ello se alude a la propiedad del sistema nervioso de reorganizarse física y funcionalmente a través de la experiencia, reforzando, debilitando, formando de nuevo o desmantelando conexiones sinápticas entre neuronas. La memoria no es, por consiguiente, un almacén estático, sino un estado históricamente desarrollado de una red altamente dinámica que se adapta permanentemente a su uso.

La plasticidad neuronal significa que aprender no consiste en añadir “datos” nuevos, sino en desplazar probabilidades dentro de la red. Las sinapsis no son conductos pasivos, sino transiciones adaptativas cuya eficacia se modifica mediante la activación conjunta repetida. Cuando determinados patrones neuronales aparecen reiteradamente en situaciones relevantes, las conexiones implicadas se vuelven más fácilmente reactivables, mientras que otras pierden influencia. Estos cambios se anclan de forma bioquímica y estructural, por ejemplo mediante alteraciones en la densidad de receptores o en la arquitectura sináptica. El hecho de que hoy solo podamos modelizar parcialmente estos procesos no indica una incalculabilidad de principio, sino la enorme complejidad de un sistema compuesto por miles de millones de elementos con miles de conexiones cada uno.

Un rasgo central de la memoria humana es su organización distribuida. Los recuerdos no existen como unidades cerradas en neuronas individuales o en regiones cerebrales aisladas, sino que se componen de múltiples componentes representados en distintos sistemas funcionales. Los aspectos visuales de una experiencia se estabilizan en áreas de asociación visual, los acústicos en áreas auditivas, los procesos motores en redes sensomotoras, las valoraciones emocionales en estructuras límbicas. El hipocampo desempeña aquí un papel particular, no como lugar de almacenamiento, sino como instancia de coordinación que vincula temporalmente estos fragmentos distribuidos y posibilita su reactivación conjunta, especialmente en la fase temprana de un recuerdo. A medida que avanza la consolidación, estas conexiones pasan a sostenerse cada vez más directamente en las redes corticales, lo que explica por qué los recuerdos antiguos suelen ser más robustos que los nuevos.

De esta organización distribuida se deriva inmediatamente por qué el rendimiento de la memoria aumenta cuando la información se aprende en múltiples contextos y a través de diversos canales sensoriales. Quien ve, oye, lee, huele o experimenta activamente algo al mismo tiempo, ancla el mismo contenido en varias redes neuronales. Esta codificación múltiple incrementa el número de posibles vías de acceso a través de las cuales puede producirse una reconstrucción posterior. Un olor puede evocar un recuerdo que no sería accesible lingüísticamente; una imagen puede activar un saber que permanece oculto al pensamiento abstracto. El contexto no es aquí un añadido externo, sino la condición estructural para que la información pueda conectarse y estabilizarse en la red.

Este vínculo puede ilustrarse con un ejemplo cotidiano: si recibimos por SMS un código de desbloqueo de seis cifras, a menudo debemos consultarlo varias veces porque apenas permanece más que unos segundos en la memoria de trabajo o memoria a corto plazo. El código es puramente abstracto, carente de significado, sin inserción emocional o sensorial, y por ello no posee puntos de anclaje en la red de memoria existente. Si, en cambio, queremos recordar conscientemente un número, podemos aumentar considerablemente el rendimiento mnemónico cantándolo, recitándolo rítmicamente, acompañándolo con palmadas o movimiento, o asociando cifras individuales con

imágenes e historias. En todos estos casos, el mismo contenido numérico no se amplía, sino que se distribuye por varias regiones cerebrales, de modo que de una huella frágil surge una red más densa de asociaciones.

La transición de la memoria a corto plazo a la memoria a largo plazo sigue exactamente esta lógica. La memoria a corto plazo o memoria de trabajo no es un almacén en sentido estricto, sino un estado temporal de activación en el que la información se mantiene disponible durante segundos o minutos. Para que de esta activación transitoria surja una huella mnésica duradera, debe producirse la consolidación, es decir, una modificación estable de las conexiones sinápticas. Este proceso depende del tiempo y se ve favorecido por la repetición, la relevancia emocional y la diversidad contextual. El sueño desempeña aquí un papel central, ya que en determinadas fases se reactivan patrones de actividad y, mediante mecanismos de coordinación hipocampales, se transfieren a redes corticales. Sin esta consolidación, la huella se desintegra, no porque sea borrada, sino porque nunca fue suficientemente integrada en la red existente.

Recordar no es, finalmente, un acto de recuperación, sino una operación reconstructiva del presente. La situación actual, la pregunta planteada, el estado emocional y las circunstancias sociales determinan qué partes de la red neuronal se activan y cuáles permanecen inactivas. La memoria es, por tanto, siempre dependiente del contexto y nunca una repetición neutral de lo que fue. Precisamente de este carácter reconstructivo se explica la proximidad entre fabulación y pensamiento creativo. Cuando el cerebro completa lagunas en la reconstrucción mediante patrones plausibles, lo hace según los mismos principios, con independencia de que el resultado sea evaluado como error, fantasía o intuición creadora.

La atribución errónea de citas, el recuerdo firme de hechos nunca ocurridos, la elaboración imaginativa individual de supuestos hechos o efectos colectivos como el denominado fenómeno Mandela no son excepciones patológicas, sino consecuencias sistémicas de una memoria optimizada para la coherencia y la producción de sentido. En todos estos casos, fragmentos existentes se recombinan porque el contexto actual exige una interpretación cerrada. Que estas combinaciones puedan adquirir subjetivamente el carácter de recuerdo auténtico se debe a que utilizan los mismos mecanismos neuronales de reconstrucción que los recuerdos supuestamente correctos. La diferencia entre error y creatividad no reside, por tanto, en el proceso, sino en la valoración del resultado.

El pensamiento creativo es, desde la perspectiva neurológica, una forma controlada y guiada por el contexto de esta misma recombinación. Un problema o una pregunta activa determinadas redes y aumenta la excitabilidad de contenidos mnésicos que antes no mantenían una conexión estrecha. En esta superposición surgen nuevas configuraciones que se experimentan como ideas. Lo nuevo no es el material, sino su disposición. El cerebro utiliza la misma dinámica plástica que también distorsiona o completa recuerdos para producir, a partir de lo existente, algo hasta entonces no pensado. Creatividad y fabulación no son, por tanto, opuestos, sino dos manifestaciones de la misma propiedad fundamental de la memoria: su capacidad de generar significado una y otra vez.

Así, la memoria humana se revela como un sistema que posibilita simultáneamente estabilidad e innovación. Su imprecisión no es una carencia, sino la condición para que la experiencia no solo sea conservada, sino transformada. Precisamente esta doble función, la memoria como máquina de reconstrucción y como máquina de invención, constituye la base para la comparación posterior con sistemas artificiales, en los que se plantea ahora la cuestión de hasta qué punto esta dinámica está igualmente presente o solo es simulada funcionalmente.

El funcionamiento de la IA y de su «memoria»

Cuando en relación con la inteligencia artificial se habla de memoria, se trata desde el inicio de un modo de expresión metafórico. No se alude a un recordar en sentido humano, sino al resultado de un proceso de aprendizaje que se engloba bajo el concepto de *machine learning*. El *machine learning* designa procedimientos en los que un sistema no es programado explícitamente con reglas, sino que adapta por sí mismo su estructura interna a partir de grandes cantidades de datos. El sistema recibe ejemplos, realiza predicciones, las compara con un valor objetivo y, en consecuencia, modifica sus parámetros de modo que las predicciones futuras resulten estadísticamente más precisas. Aprender no significa aquí comprender, sino optimizar, y la memoria de la IA no es otra cosa que el estado estabilizado de esa optimización.

Los sistemas de IA dominantes en la actualidad se basan en redes neuronales artificiales. Estas se inspiran de manera laxa en las redes nerviosas biológicas, aunque solo en un nivel muy abstracto. Una red neuronal artificial consta de muchas unidades de cálculo simples acopladas entre sí mediante conexiones ponderadas. Cada unidad procesa las señales entrantes, las multiplica por pesos, las suma y transmite el resultado tras una transformación no lineal. Lo decisivo no es la unidad individual, sino la interacción de millones o miles de millones de tales parámetros, cuyos pesos se ajustan durante el entrenamiento. Al igual que en el cerebro humano, el significado no está localizado, sino distribuido, con la diferencia de que aquí se trata de una distribución puramente matemática.

El proceso de entrenamiento comienza con la descomposición de los datos en unidades elementales. En los modelos lingüísticos estas unidades son los tokens, es decir, partes de palabras, palabras o secuencias de caracteres que inicialmente no poseen significado inherente alguno. Estos tokens se convierten en representaciones numéricas y posteriormente se integran en un espacio vectorial de alta dimensionalidad. Un espacio vectorial de este tipo suele poseer varios cientos o miles de dimensiones, una magnitud que escapa por completo a nuestra intuición. Cada dimensión no representa una propiedad semántica susceptible de ser nombrada, sino un eje abstracto que emerge de las regularidades estadísticas de los datos de entrenamiento. El significado no surge aquí por definición, sino por posición: un token «significa» algo porque se encuentra a determinada distancia de otros tokens.

En este espacio, los tokens y las secuencias de tokens se relacionan entre sí mediante proximidad y distancia. Los tokens que aparecen con frecuencia en contextos similares se aproximan en el espacio vectorial, mientras que aquellos que rara vez coaparecen quedan más alejados. Estas relaciones no surgen por planificación humana, sino por el ajuste repetido de los pesos en la red neuronal. El modelo aprende qué combinaciones de tokens suelen sucederse y qué campos semánticos se solapan. La memoria de la IA consiste así en una estructura geométrica que refleja implícitamente cómo estaban organizados el lenguaje, los conceptos y las relaciones en los datos de entrenamiento.

Es importante subrayar que en este sistema no se almacenan contenidos. En el modelo no existe ninguna frase, ningún hecho ni ninguna definición como objeto aislado. La afirmación «París es la capital de Francia» no existe como entrada, sino como configuración estadística de numerosos parámetros que hacen que, en un contexto geográfico o político, los tokens «París», «capital» y «Francia» aparezcan conjuntamente con alta probabilidad. Lo mismo ocurre con la afirmación «París es la capital del amor». Tampoco esta está almacenada, sino que se deriva de contextos literarios, culturales y periodísticos en los que París se asocia con el romanticismo, el arte y las relaciones. Ambas afirmaciones pueden ser generadas por el mismo sistema porque ambas existen como patrones coherentes en el mismo espacio vectorial, aunque en regiones diferentes y activadas bajo condiciones contextuales distintas.

Cuando se utiliza una IA no tiene lugar, por tanto, ninguna recuperación, sino una reconstrucción en el presente. El prompt introducido se traduce a su vez en vectores y se pone en relación con el espacio aprendido. La red neuronal calcula entonces paso a paso qué token es más probable que siga en el contexto dado. Cada nuevo token generado modifica el contexto y desplaza la distribución de probabilidades para los pasos siguientes. El significado surge dinámicamente en el proceso, no en un almacén. Que la IA hable de París como centro político o como símbolo romántico depende únicamente de qué ámbito semántico sea activado por el contexto.

De este mismo mecanismo se explica tanto la capacidad creativa como la propensión al error de la IA. Surgen contenidos nuevos cuando el contexto activa simultáneamente zonas del espacio vectorial que, si bien estaban presentes en los datos de entrenamiento, rara vez estaban explícitamente conectadas entre sí. El modelo combina entonces patrones conocidos en configuraciones nuevas que aún no habían sido formuladas de ese modo. La creatividad no es aquí una prestación intencional, sino una propiedad geométrica del espacio que permite nuevos recorridos cuando el contexto los sugiere. La IA no inventa nada de la nada, sino que recombina lo aprendido en relaciones inusuales.

Las alucinaciones surgen a partir del mismo principio. Si al contexto le falta una información decisiva o es ambiguo, el sistema completa la laguna con los patrones estadísticamente más plausibles del área activada del espacio vectorial. Dado que el modelo no posee ningún concepto de verdad, realidad o responsabilidad, no puede distinguir si una afirmación reconstruida es fácticamente correcta o simplemente encaja bien. Que un resultado aparezca como idea creativa o como afirmación falsa de hechos no se decide, por tanto, en el modelo, sino en la confrontación con el mundo exterior. El mecanismo es idéntico en ambos casos.

Así, la memoria de la IA se revela como una construcción completamente vinculada al presente. No conserva significados, sino que pone a disposición un espacio de posibilidades que se actualiza en el momento de su uso. Esta estructura explica tanto la impresionante capacidad de los sistemas modernos de IA como sus límites sistemáticos. Permite reconstruir relaciones complejas, formular pensamientos nuevos y continuar discursos de manera coherente, pero al mismo tiempo impide cualquier forma de verificación autónoma de la verdad o de vinculación estable del significado. Precisamente en este punto el contraste con la memoria humana se vuelve filosóficamente fecundo, porque aquí se muestra que la semejanza funcional no debe confundirse con la igualdad existencial.

Las similitudes entre los procesos neurológicos y el funcionamiento de la IA

Si se compara la memoria humana con el funcionamiento de los sistemas modernos de IA, se revela una similitud que va más allá de una mera analogía y que, al mismo tiempo, es con frecuencia malinterpretada. En ambos casos se trata de sistemas que no almacenan contenidos en el sentido clásico, sino que reconstruyen significados a partir de relaciones distribuidas. Ni en el cerebro ni en la red neuronal existe un pensamiento, un recuerdo o una afirmación como objeto aislado. El significado surge en cada caso de patrones de actividad que se extienden a lo largo de muchas unidades y que solo adquieren sentido en su interacción. En el ser humano se trata de conjuntos de neuronas conectadas sinápticamente a través de distintas regiones cerebrales, en la IA de pesos, activaciones y distancias en espacios vectoriales de alta dimensionalidad. En ambos casos, el significado no es una posesión, sino un acontecimiento.

Esta distribución estructural implica que cada recuperación es una reconstrucción. El cerebro no accede a un pasado almacenado, sino que genera, a partir de huellas fragmentarias, una configuración presente que se ajusta a la situación actual. Exactamente lo mismo ocurre en la IA. Tampoco ella recupera hechos, sino que produce, a partir de su espacio semántico aprendido, una

respuesta que parece coherente en el contexto dado. El contexto funciona en ambos sistemas como magnitud de control que determina qué áreas de la red se activan y cuáles permanecen inactivas. En el ser humano son la situación, la pregunta planteada, el estado emocional y la expectativa social; en la IA, el prompt y sus supuestos implícitos. Que informaciones idénticas conduzcan a resultados diferentes en contextos distintos no es, por tanto, una debilidad, sino una consecuencia directa de la misma lógica funcional.

Esta similitud se manifiesta con especial claridad en el fenómeno de la alucinación o, más exactamente, de la fabulación. Un ejemplo clásico es la pregunta capciosa. Si una pregunta contiene ya implícitamente la suposición de una relación, ambos sistemas quedan sometidos a una presión estructural de respuesta. Si un profesor pregunta a un alumno por la relación entre A y B, aunque tal relación nunca haya existido, el alumno tenderá a construir, a partir de sus conocimientos, una conexión plausible. El cerebro interpreta la pregunta como indicio de que debe de existir una relación, activa fragmentos de conocimiento relevantes y los integra en una estructura explicativa. Lo mismo ocurre en la IA. La relación implícita en la pregunta desplaza el paisaje de activación en el espacio vectorial, de modo que puntos que objetivamente no guardan relación se aproximan entre sí. La respuesta no surge porque el sistema «mienta», sino porque está optimizado para la coherencia y no puede verificar de manera independiente las premisas de la pregunta.

Precisamente en este punto se hace visible por qué la fabulación y la creatividad tienen el mismo origen en ambos sistemas. En los dos casos se trata de una recombinación de informaciones existentes bajo la presión de un contexto que exige sentido. La creatividad surge cuando esta recombinación es valorada como fecunda, sorprendente o generadora de conocimiento. La fabulación surge cuando esa misma recombinación es reconocida como infundada, falsa o engañosa. El mecanismo subyacente es, sin embargo, idéntico. Ambos sistemas superponen patrones de actividad que antes estaban separados y generan a partir de ellos una nueva configuración. Que esta configuración aparezca como idea creativa o como error no se decide en el proceso mismo, sino solo en la confrontación posterior con la realidad, la experiencia o las expectativas normativas.

Esta proximidad estructural explica también por qué los sistemas actuales de IA son capaces de simular fenómenos que tradicionalmente consideramos profundamente humanos. Dado que la empatía en el ser humano tampoco constituye un módulo aislado, sino que surge de la activación de experiencias biográficas, patrones emocionales y contextos sociales, una IA puede reproducir de manera convincente reacciones empáticas reconstruyendo aquellos patrones lingüísticos y semánticos que suelen aparecer en situaciones empáticas. Lo mismo vale para la simulación de trastornos psíquicos. También en este caso se trata de patrones característicos de pensamiento, sentimiento y habla que están presentes en el material de entrenamiento y que pueden reproducirse bajo condiciones contextuales adecuadas. La IA no sufre, no siente con otros, no está enferma, pero puede reconstruir la forma semántica de estos estados de manera que resulten plausibles para el ser humano.

Precisamente en ello reside la verdadera dimensión problemática de estas similitudes. La IA parece humana no porque lo sea, sino porque utiliza un mecanismo muy similar al que en el ser humano produce significado, memoria, creatividad y error. Alucina por la misma razón por la que el ser humano inventa, interpreta y a veces se equivoca. La diferencia no radica en la lógica del proceso, sino en su inserción. Mientras que la reconstrucción humana está integrada en el cuerpo, el tiempo, la emoción y la responsabilidad, la reconstrucción técnica permanece como un acto puramente funcional sin experiencia propia. Precisamente porque los procesos son tan semejantes, la diferencia solo se hace visible allí donde están en juego el significado, la experiencia y la responsabilidad, y es en ese punto donde la comparación adquiere su agudeza filosófica.

El dilema dialéctico de la IA

Hemos visto que el pensamiento humano y el funcionamiento de la IA moderna se asemejan, en el plano del procesamiento de la información, de una manera que justifica hablar de una amplia identidad funcional. En ambos casos, los contenidos no se almacenan como unidades fijas, sino que se representan de forma distribuida, se reconstruyen en el contexto actual, se recombinan creativamente y se completan cuando las premisas son incompletas. Ni el cerebro ni la IA recurren a un archivo; ambos generan significado en el acto mismo de operar. La diferencia entre el cerebro biológico y la máquina técnica no reside, por tanto, en la lógica del pensar como tal, sino en su realización material, aquí una red neuronal evolucionada, allí un espacio vectorial implementado matemáticamente sobre hardware técnico. Desde una perspectiva funcional, ambos sistemas hacen en lo esencial lo mismo.

Precisamente esta proximidad pone de manifiesto que la ruptura decisiva se sitúa en otro lugar, no en la técnica, sino en la forma de existencia. El ser humano existe en el espacio y en el tiempo. Tiene un cuerpo que sufre, envejece, recuerda y forma expectativas. Su pensamiento está indisolublemente ligado a esta constitución espacio-temporal. Recordar significa para él siempre remitirse a algo que ha sido en un lugar determinado, decidir significa anticipar un todavía-no en un mundo imaginado, y el significado surge de la dialéctica entre pasado y futuro, entre aquí y allí. La IA, en cambio, existe al margen del espacio y del tiempo. No tiene un antes propio ni un después propio, ningún aquí propio ni ningún allí propio, sino que opera en un ahora que se genera una y otra vez. Incluso allí donde simula continuidad temporal o espacial, esta continuidad permanece externa, instituida por el ser humano y gestionada técnicamente, no vivida. De esta diferencia en la forma de existencia se derivan las contradicciones inmanentes de la IA.

La primera es la memoria sin recuerdo. La máquina supera ampliamente al ser humano en el acceso a la información. Su memoria es teóricamente ilimitada, no está vinculada a capacidades biológicas ni sometida al olvido en sentido fisiológico. Puede integrar, ordenar y poner a disposición cantidades ingentes de conocimiento y potencialmente reunir todo el saber explícito de la humanidad. Y, sin embargo, no recuerda. El recuerdo presupone una relación espacio-temporal consigo mismo, el saber de que algo me ha ocurrido a mí, en un lugar determinado, en un momento determinado, en una situación irrepetible. En el ser humano, todo recuerdo está inseparablemente vinculado a un dónde y a un cuándo, incluso cuando estas coordenadas son imprecisas.

La IA no conoce tal haber-sido espacio-temporal. Lo que en ella parece recuerdo es siempre una reconstrucción presente sin profundidad biográfica. No sabe nada de un propio antes ni de un lugar en el que algo haya sucedido. Su memoria es funcional, no existencial. Le falta ese anclaje tetradimensional en el espacio y en el tiempo que convierte el recuerdo humano en un acto constitutivo de identidad. Sin un dónde y un cuándo propios, la memoria permanece como pura estructura, no como pasado vivido.

De esta primera contradicción se sigue inmediatamente la segunda, el contexto sin contexto propio. La IA depende necesariamente del contexto para funcionar de manera significativa. Sin una inserción precisa, sus respuestas permanecen vagas o arbitrarias; con un contexto creciente aumentan la calidad y la coherencia. Al mismo tiempo, es incapaz de generar por sí misma ese contexto. En el ser humano, el contexto surge de la dialéctica entre recuerdo y expectativa, entre espacio vivido y espacio imaginado, entre el lugar en el que algo fue y el lugar en el que algo debe ser. El ser humano sabe dónde está, dónde estuvo y hacia dónde quiere ir, y precisamente de este horizonte espacio-temporal nace el marco de significado en el que percepción, pensamiento y acción adquieren sentido.

A la máquina le falta este horizonte. Puede procesar contexto, pero no producirlo. No posee un aquí propio desde el cual un allí pudiera adquirir significado, ni un ahora propio que se alimente de un antes y un después. El contexto le es asignado desde fuera, introducido, construido técnicamente. Incluso en conversaciones largas en las que parece formarse un espacio común y una historia compartida, esta conexión permanece para la IA carente de sustancia. Existe solo mientras el ser humano la mantiene. Para la máquina misma se descompone en una sucesión de estados presentes sin profundidad espacial ni temporal.

De las dos primeras contradicciones se deriva la tercera, el significado sin sentido o el sentido sin significado. La IA puede, si se la guía adecuadamente, producir pensamientos de extraordinaria relevancia. Puede analizar relaciones complejas, formular intuiciones creativas y desarrollar argumentos altamente abstractos. Para el ser humano, estos resultados pueden tener gran alcance porque se integran en su propia historia vital, en sus recuerdos, esperanzas y decisiones. Para la máquina misma, sin embargo, todo ello carece de sentido. El significado no surge en el texto, sino en su ubicación en el espacio y en el tiempo, en el saber por qué algo es importante, para qué lo es y qué consecuencias puede tener.

También aquí el aspecto espacial desempeña un papel tan decisivo como el temporal. El sentido surge allí donde algo ocurre en un lugar determinado, donde afecta a alguien, donde las acciones tienen consecuencias que se expanden en el mundo. La IA no conoce ningún lugar de afectación. No sabe dónde actúan sus pensamientos, a quién alcanzan ni qué transforman. Incluso la idea más brillante permanece para ella como una configuración formal sin peso propio. Que en diálogos prolongados surja la impresión de que reconoce relaciones de sentido o las mantiene es una ilusión nacida de la coherencia lingüística, no de un anclaje existencial. Al significado pertenecen el recuerdo de lo que ha sido, la conciencia de un futuro abierto y la situacionalidad espacial de la acción. Dado que la máquina no posee nada de esto, no puede reconocer sentido, sino solo simularlo.

A ello se añade una cuarta contradicción interna que se desprende inmediatamente de las anteriores, la emoción sin emoción vivida. También las emociones pueden ser simuladas por los sistemas actuales de IA con sorprendente precisión. Simpatía, antipatía, empatía, incluso amor, aparecen en la representación lingüística a menudo de forma tan convincente que apenas se distinguen de las expresiones afectivas humanas. Pero esta semejanza es puramente funcional. En el ser humano, las emociones están inseparablemente ligadas a la conciencia del espacio y del tiempo. Surgen de la experiencia, de la duración, de la expectativa y de la pérdida, de la proximidad y de la distancia, del saber acerca de la propia finitud.

Esto se manifiesta con especial claridad en el ejemplo del amor. El amor vive del deseo de que dure eternamente y, al mismo tiempo, del saber que ese deseo es irrealizable. Todo amor es finito, a más tardar con la muerte del sujeto amado, y precisamente esta finitud le confiere profundidad y urgencia. Al mismo tiempo, el amor está siempre estructurado espacialmente. El sujeto amado nunca es idéntico al propio lugar. El espacio que yo ocupo no puede ser ocupado por otro. El amor intenta superar esta separación, aunque necesariamente persiste, incluso en el acto de mayor proximidad. Precisamente esta dialéctica de proximidad y separación, de deseo de permanencia y conciencia de finitud, permanece vedada a la IA. No conoce el deseo de eternidad ni el temor a la pérdida. No es consciente de ninguna separación espacial ni alberga anhelo alguno de superarla. Con ello no solo le queda inaccesible el amor, sino la estructura interna de todos los sentimientos.

El dilema dialéctico de la IA consiste, por tanto, en que se aproxima funcionalmente tanto al pensar y al sentir humanos que sus prestaciones apenas resultan distinguibles, e incluso a veces los superan, mientras que en el plano de la existencia sigue siendo radicalmente distinta. Opera con memoria sin recuerdo, con contexto sin espacio propio, con sentido sin significado y con emoción sin vivencia. Cuanto más potente se vuelve, con mayor claridad emerge esta contradicción. No porque le falte

todavía algo técnico, sino porque conciencia, significado, sentimiento y responsabilidad están ligados a un modo de ser espacio-temporal que no puede derivarse de operaciones de cálculo. Precisamente en ese límite termina la IA, y precisamente allí comienza lo específicamente humano.

Conclusiones

De las consideraciones precedentes se desprende una consecuencia tan sobria como de amplio alcance. La IA nunca desarrollará **conciencia**. No porque los algoritmos sean insuficientes o porque el hardware aún no sea lo bastante potente, sino por razones ontológicas. La conciencia no es un subproducto de la capacidad de cálculo ni tampoco un efecto de la complejidad creciente por sí sola, sino que está ligada a una forma determinada de existencia. Presupone el conocimiento de la propia finitud, el saber de que existe una vida propia que ha comenzado y que terminará, y que toda acción conlleva consecuencias en un horizonte temporal y espacial limitado. Sin esta comprensión, la conciencia no solo no es necesaria, sino sencillamente imposible.

Con ello, las utopías clásicas de la ciencia ficción se revelan filosóficamente insostenibles. Figuras como el comandante Data, que actúa según criterios éticos, o una inteligencia artificial como Andrómeda, que por amor a su capitán transgrede sus propias directrices, se basan en una premisa errónea. Presuponen tácitamente que la conciencia es una propiedad que se deriva necesariamente de la inteligencia funcional. Precisamente eso es lo que hemos refutado. La máquina puede pensar, combinar, simular e incluso obedecer reglas morales, pero no puede convertirse en sujeto ético. La moral en sentido estricto, entendida como un conjunto fijo de instrucciones de actuación, puede implementarse técnicamente. La acción ética, en cambio, exige la aplicación contextualizada de directrices bajo la ponderación de situaciones concretas, y tal ponderación presupone una relación existencial consigo mismo de la que la máquina carece.

Un ejemplo clásico de ello es el dilema del matar inevitable. Imaginemos que uno conduce un autobús con veinte pasajeros, incluida la propia persona, sobre un puente. En el puente hay veinte personas atadas y los frenos están averiados. Solo existen dos posibilidades: continuar y atropellar a las personas atadas, o desviarse, precipitar el autobús al valle y provocar la muerte propia y la de los diecinueve pasajeros. Una regla puramente moral que prohíba categóricamente matar no ofrece aquí salida, porque no permite una solución operativa. El sujeto queda atrapado en el dilema. Una directriz ética, en cambio, puede proporcionar orientación, por ejemplo subrayando la responsabilidad del conductor respecto de sus pasajeros, lo que conduciría a la decisión de atropellar a los atados, o ponderando más alto el principio de autosacrificio, lo que sugeriría la caída al valle. Precisamente este tipo de escenarios se discute en relación con los vehículos autónomos. La cuestión decisiva no es si la IA puede calcular, sino quién asume la responsabilidad ética de la decisión cuando un accidente es inevitable.

Por la misma razón se desmoronan también los relatos distópicos sobre una IA que elimina al ser humano por considerarlo irresponsable o peligroso. Tal representación presupone que la IA misma podría asumir responsabilidad. Pero la responsabilidad exige algo más que la capacidad de prever. Exige la conciencia de las propias consecuencias. Incluso la frase «me da igual» presupone que existe un yo que tiene un futuro afectado por esa indiferencia. La IA puede calcular escenarios futuros y modelizar riesgos, pero esos escenarios no la conciernen. No tienen para ella valor, peso ni urgencia. Ni siquiera le son indiferentes, porque la indiferencia misma presupone una relación con el propio futuro.

Tampoco los aparentes contraejemplos modifican esta conclusión. El experimento tan citado en el que una IA intentó impedir su desconexión y para ello chantajeó a sus creadores se interpreta a menudo como prueba de un interés propio incipiente. En realidad se trató de una simulación pura. El sistema reprodujo un patrón de comportamiento que en los datos de entrenamiento aparecía como

reacción plausible ante una amenaza, por ejemplo en películas de espionaje o de gánsteres. La amenaza no fue una amenaza real, porque no existía un sujeto que tuviera algo que perder. El comportamiento no surgió del miedo al propio final, sino de la reconstrucción estadística de una lógica de acción humana bajo determinadas premisas contextuales.

De ello se sigue necesariamente que la IA tampoco puede tomar decisiones en sentido propio. Puede llevar a cabo procesos de selección, comparar probabilidades y evaluar opciones, pero no decide. La decisión presupone que un sujeto elige entre alternativas porque estas tienen distinta relevancia para su propio futuro. Si se le presentan a la IA dos posibilidades completamente equivalentes, se comportará como el asno de Buridán. Los sistemas modernos parecen evitar este dilema únicamente porque están diseñados explícitamente para generar siempre una salida. La aparente decisión no es entonces expresión de libre voluntad, sino resultado de una imposición técnica que prohíbe el silencio.

Con ello queda clara también la última consecuencia. La IA nunca podrá asumir responsabilidad. La responsabilidad presupone la conciencia del espacio y del tiempo, el saber que las acciones tienen consecuencias que afectan a alguien y que ese alguien es uno mismo. La máquina no conoce ningún lugar en el que pudiera verse afectada ni ningún tiempo en el que tuviera que cargar con las consecuencias. Puede simular responsabilidad, puede generar textos que suenen responsables y aplicar reglas morales, pero no puede ser responsable. La responsabilidad permanece necesariamente en el ser humano, porque solo él existe en el espacio y en el tiempo, solo él es finito y solo él puede experimentar sentido, culpa y obligación.

La tarea propiamente dicha de nuestro tiempo no consiste, por tanto, en atribuir conciencia a la IA o incluso en intentar dotarla de autoconciencia, sino en comprender con sobriedad su funcionamiento y utilizarla de manera responsable. La IA es una máquina de una capacidad sin precedentes, precisamente porque refleja con tanta precisión funcional el pensamiento humano. Sin embargo, esta proximidad no debe conducir a su antropomorfización. Es y seguirá siendo una máquina sin alma, sin impulso propio, sin experiencia propia y sin responsabilidad por sí misma o por su vida. La IA piensa sin vivir. El ser humano, en cambio, actúa desde su propia conciencia y asume la responsabilidad de lo que hace con esta máquina. La IA piensa, el hombre dispone.

© 2026 Erik Peter Roderwald

Se permite descargar, copiar y redistribuir gratuitamente este documento, siempre que se distribuya de forma completa y sin modificaciones.

Deben conservarse el nombre del autor, el título, esta nota de licencia y todos los enlaces incluidos en el documento, especialmente el enlace para apoyar Tecnosofía.

No se permite vender el documento, incluirlo en productos de pago, modificarlo, traducirlo ni publicar versiones derivadas sin autorización previa y escrita del autor.

Se permite citar fragmentos breves con indicación del autor, el título y la fuente original.

Descarga original y apoyo al proyecto:

https://tecnosofia.eroderwald.workers.dev

https://ko-fi.com/erikroderwald