

NEW LEFT REVIEW 155

SEGUNDA ÉPOCA

NOVIEMBRE-DICIEMBRE 2025

ARTÍCULOS

MARTÍN MOSQUERA	El significado de Milei	7
DYLAN RILEY & ROBERT BRENNER	Respuesta a los críticos	29
OWEN HATHERLEY	Arquitectura del futuro	81
NAN DA	La crítica en la era de la IA	111

CRÍTICA

NICHOLAS MULDER	Interludios de abundancia	139
GABRIELE PEDULLÀ	Los materiales de Timpanaro	153

WWW.NEWLEFTREVIEW.ES

© New Left Review Ltd., 2000

Licencia Creative Commons

Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional (CC BY-NC-ND 4.0)



SUSCRÍBETE

ts
traficantes de sueños



NAN Z. DA

LA CRÍTICA LITERARIA EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Porque no pude esperar a la Muerte –
Ella me esperó a mí con amabilidad¹.

PROCESEMOS LA INFORMACIÓN en tiempo real. Un amigo que es un excelente crítico literario me recordó que me detuviera un segundo. «Piénsalo», me dijo. «Tienes que estar por delante de la muerte, por delante en una carrera, por delante en una prueba cronometrada, para detenerte o no detenerte ante ella. Al mismo tiempo, la muerte tiene que estar por delante de ti para detenerse o no detenerse».

I

El poema de Emily Dickinson comienza con un movimiento forzado. Antes de preguntarnos qué significa *esperar* a la muerte, o que la muerte *se pare* por ti, antes de lidiar con el resto del poema y el viaje en carruaje en el que nos embarca, nos encontramos con una confusión de posiciones relativas. ¿Dónde estaban el narrador y la Muerte el uno respecto al otro *en ese momento*? Dado que el poema comienza con «porque», que es una conjunción subordinada sin regla (¿Por qué? *Porque*), no se

¹ Emily Dickinson, «Because I could not stop for Death» (1863), *The Poems of Emily Dickinson*, ed. R. W. Franklin, Cambridge (MA), 1998. Para este y otros poemas de Dickinson citados en el texto, seguimos la traducción de Enrique Goicolea en Emily Dickinson, *Poesía completa*, Colmenar Viejo, 2011.

sabe quién iba en el carruaje del narrador, la Muerte y la Inmortalidad, ni quién invitó a quién a subir al carruaje. Una confusión similar se produce a mitad de este poema de seis estrofas, cuando el narrador y la Muerte viajan juntos:

Pasamos la escuela, donde los niños reñían en el Patio,
a la hora del recreo;
pasamos los campos de extasiado cereal
pasamos el sol que se estaba poniendo

—o mejor, fue él quien nos pasó a Nosotros—.

Pero esta repetición de quién adelantó a quién no aclara mucho las cosas. Si bien es posible confundirse sobre la posición técnica de la Tierra con respecto al Sol, técnicamente no es posible que algo esté estrictamente detrás y, en virtud de ello, de repente delante.

¿Dónde está este lugar imposible? ¿Cómo sopla el viento allí? Sea donde fuere, somos testigos de un enfrentamiento cuyos orígenes y resultados se ocultan, algo habitual en los poemas de Dickinson². Hablando en términos de enfrentamientos y confrontaciones, «Porque no pude esperar a la muerte...» es un poema sobre lo que ocurre cuando te encuentras ante algo casi imposible de rebatir o vencer, como la muerte. Es difícil saber qué está pasando en el poema, porque la contienda tiene lugar «antes» o, de alguna manera, no en el escenario que uno tendría en mente para tal contienda, si es que se trata de una. Si tuviéramos que imaginar un punto A y un punto B para el último tramo de un ser humano con la Muerte, y todos los puntos significativos entre ambos, probablemente no serían los puntos que habríamos imaginado.

2

Muchos han escrito recientemente sobre la importancia de la crítica literaria y la lectura atenta en respuesta a una percepción de pérdida de integridad en el seno del campo y de amenazas existenciales externas³.

² Véase, por ejemplo, «A Bird came down the Walk», «I Started Early – Took my Dog», «I cannot live with You» y «My Life had stood – a Loaded Gun».

³ Véase el debate iniciado en estas páginas por Francis Mulhern, «Revoluciones críticas», *NLR* 110, mayo-junio de 2018, con contribuciones posteriores de Joseph North, «Dos pasajes en Raymond Williams», *NLR* 116/117, mayo-agosto de 2019, Lola Seaton, «Los fines de la crítica» *NLR* 119, noviembre-diciembre de 2020, Patricia McManus, «Una nueva crítica literaria», *NLR* 132, enero-febrero de 2022, Anahid Nersessian,

Este ensayo quiere plantear una pregunta diferente: ¿*dónde* está la crítica literaria? ¿Cómo atraviesa el mundo? Detrás de esta línea de cuestionamiento se encuentra la convicción de que todos los ámbitos acaban encontrando problemas de crítica literaria y que todas las «tuberías» – por usar un término de la informática– tienen un punto de contacto con la crítica literaria. La crítica literaria adopta muchas formas, pero su papel fundamental como árbitro de las inferencias a partir de textos registrados cobra ahora una nueva relevancia, ya que ahora las inferencias a partir de textos registrados se realizan a gran escala.

Quería comenzar con el poema de Dickinson, porque es un claro ejemplo de algo que activa las habilidades necesarias para la lectura atenta, definiéndose esta como el procesamiento de información crítica en tiempo real. Este tipo de lógica se produce mucho antes de que «profundicemos entre líneas», antes de la interpretación creativa, antes de comentar la innovación formal, antes de separar lo que es cierto en el poema de lo que está implícito⁴. Al elegir el tipo de escenario que la mayoría de nosotros ya creemos conocer –los enfrentamientos con la mortalidad son uno de los tropos más antiguos–, el poema desencadena una preocupación por el empirismo tal y como se expresa en los estudios literarios. La preocupación es que el exceso de confianza y la presunción sobre cómo tienden a ir las cosas harán que la mente pase por alto la información empírica realmente existente. Como información condensada en muy pocos datos, la secuencia inicial –«Porque no pude esperar»– es «pasable por alto». Se puede resumir el poema como la meditación de una mujer sobre la muerte y la inmortalidad sin tenerlo en cuenta en absoluto, descartando así su inteligencia; y casi nada se interpondrá en tu camino.

Dado que este poema hace hincapié en el paso entre determinados datos y las inferencias que se pueden hacer, y dado que describe posiciones relativas confusas en una contienda desigual, voy a dar un rodeo que nos

«¿Por amor a la belleza?», *NLR* 133/134, marzo-junio de 2022, y Benjamin Kunkel «¡Críticos, críticas, historizaos!», *NLR* 136, septiembre-octubre de 2022.

⁴ Como dijo I. A. Richards en el *locus classicus* de la lectura atenta: «Primero debe venir la dificultad de *descifrar el sentido claro* de la poesía». Continuó señalando: «El hecho más inquietante e impresionante [...] es que una gran proporción de lectores de poesía de nivel medio a bueno [...] con frecuencia y repetidamente *no logran entenderla*, ni como declaración ni como expresión. No logran descifrar su sentido prosaico, su significado claro y evidente como un conjunto de frases inglesas normales e inteligibles, separadas de cualquier otro significado poético», I. A. Richards, *Practical Criticism: A Study of Literary Judgement*, Londres, 1929; ed. cast.: *Crítica práctica*, Madrid, 1991.

ayude a comprender otro enfrentamiento en el que es difícil saber quién va por delante o qué es qué. Ese enfrentamiento, en pocas palabras: ¿la inteligencia artificial es buena o mala para la cultura literaria, para las humanidades? ¿Conducirá a la extinción de los estudios literarios tradicionales o ayudará a innovar y progresar? Invariablemente, la respuesta es *ambas cosas*, y [...]. En lugar de opinar de forma más especulativa, sigamos el ejemplo de Dickinson y entremos en el debate desde un punto más peculiar, para probar un conjunto diferente de palabras clave: «ubicación crítica», «razonamiento basado en secuencias» y «validez inferencial», en lugar de «robots», «mimetismo», «alucinaciones» o «desinformación». Este último conjunto sigue siendo obviamente relevante, pero creo que nos hace saltarnos pasos cuando aún estamos *en route*.

3

Dickinson sabía cómo se pueden relativizar las cosas. Al igual que en su poema, podríamos empezar por llevar el tema a un registro más cotidiano: la escuela, el campo, la institución, los trajes y sus ocasiones. Se trata de un panorama inquietante, sin duda, pero aún está lejos de su conclusión; la muerte aún no es letal.

Con el ánimo de rebajar un poco las expectativas, debemos desglosar la IA en sus soluciones técnicas y sus «tuberías». Me viene a la mente una viñeta satírica sobre el bombo publicitario alrededor de la IA, que apareció por primera vez en las redes sociales hacia 2018; consta de cuatro viñetas. En la primera, una persona mira una grieta en la pared, con forma de gráfico de regresión positiva; lleva la etiqueta «Estadística». A continuación, cuelga un marco de fotos alrededor de la grieta. En el tercer panel, se yergue orgulloso junto a la grieta enmarcada, ahora titulada «Aprendizaje automático». En el cuarto, se dirige a un numeroso público, con la grieta ahora titulada «Inteligencia artificial».

Obviamente, se trata de una deflación exagerada de la inteligencia artificial, pero tenemos la costumbre de promocionar diferentes partes del mismo proceso como si fueran totalmente nuevas. Cuando la «ciencia de datos» se manifiesta en el mundo real, casi siempre está vinculada a un tipo concreto de datos y a un tipo concreto de resultado: los datos en un extremo del proceso y el resultado previsto en el otro. Así es como yo pensaría y hablaría de la IA: encontrarla antes en el proceso y rebajar

un poco las expectativas. Del mismo modo que muchos imaginan que la «ciencia de datos» es un área de estudio establecida y epistémicamente coherente, también pueden ser preventivamente hiperbólicos con respecto a la IA, encontrándola al final de sus proliferantes sectores, del mismo modo que uno se encuentra con un animal formidable al final de un túnel, y no como una mezcla heterogénea de ajustes estadísticos, procesamiento del lenguaje natural y conjuntos de datos de entrenamiento y referencia compartidos hasta la extenuación.

La «IA» es una verdadera fuerza decisiva para el trabajo cultural e intelectual –y para la crítica literaria, en particular– porque nos obliga a volver a nuestros fundamentos y a definir con precisión qué tipo de inteligencia humana podemos ofrecer, en el mejor de los casos, y dónde se ejerce al máximo esa inteligencia, dónde está menos sujeta al control de calidad y cómo se ramifica. Pero tenemos que abordar esa jugada decisiva en el lugar en que se juega por primera vez.

4

Es útil que se hayan elaborado buenas explicaciones sobre cómo funcionan realmente los modelos extensos de lenguaje (LLM) de la IA. (Los LLM no son la única forma de IA, que incluye muchas aplicaciones no lingüísticas. Un ejemplo reciente e interesante es la detección de los orígenes geológicos de partículas individuales de arena o «SandAI». Aun así, muchas decisiones no lingüísticas sobre cuestiones no lingüísticas se toman mediante modelos lingüísticos predictivos⁵). Estas explicaciones entre bastidores nos permiten dejar de atribuir poderes mágicos a la IA o de hacer analogías erróneas con ella y así sobreestimar y, por lo tanto, subestimar a lo que nos enfrentamos. Entre los ejemplos recientes se encuentra la definición deflacionaria del escritor de ciencia ficción Ted Chiang de los Transformadores Generativos Preentrenados (GPT) como «algoritmos de compresión con pérdida»⁶. Del mismo modo, el

⁵ Véase Michael Hasson, M. Colin Marvin y Mathieu Lapôtre, «Automated determination of transport and depositional environments in sand and sandstone», *Proceedings of the National Academy of Sciences*, vol. 121, núm. 40, 2024.

⁶ «Con pérdida» en el sentido de que su versión de todo lo que hay en la red está tan comprimida que se descartan (pierden) muchos datos, lo que crea lagunas que se rellenan mediante interpolación, es decir, estimando lo que falta a partir de lo que hay a ambos lados, Ted Chiang, «ChatGPT Is a Blurry JPEG of the Web», *The New Yorker*, 9 de febrero de 2023.

informático Stephen Wolfram destaca que ChatGPT siempre tiene como objetivo producir una «continuación razonable» de cualquier texto que tenga hasta el momento, basándose en «lo que se podría esperar que alguien escribiera después de ver lo que la gente ha escrito en miles de millones de páginas web»⁷. Estas descripciones técnico-realistas aclaran el tipo de imitación en el que destaca la IA: su manipulación exacta de los datos del lenguaje humano natural.

Entre todos los recordatorios de que quizá hemos estado buscando en los lugares equivocados, patrullando las calles equivocadas, me gustaría destacar la observación que Jeffrey Binder hace en un libro reciente: contrariamente a lo que afirma Warren Weaver en su famoso memorándum de 1949 sobre la traducción automática, los aspectos «alógicos» del lenguaje son precisamente aquellos que las máquinas son más capaces de manipular⁸. Lo que imaginábamos como cosas difusas e imposibles de definir, como las «vibraciones» y el «estilo», el *je ne sais quoi* de experiencias literarias distintivas, ha resultado relativamente fácil de imitar utilizando los LLM. Era solo cuestión de tiempo que las herramientas estadísticas pudieran generar «sentido» en frases gramaticalmente correctas e incluso estéticamente agradables, saturadas de información. Pero si la mímesis ha sido relativamente fácil, generar un texto lógicamente coherente ha resultado ser más difícil, como señala Binder. Esto corrobora la opinión de Chiang de que GPT y otras tecnologías similares pueden destacar en verosimilitud, porque son buenas interpolando entre grupos lingüísticos conocidos. Saben qué puntos A y B deben mantenerse constantes para rellenar lo que podría haber entre ellos.

Un ejemplo ligeramente diferente: si juegas al ajedrez solo con ChatGPT, es decir, sin un motor específico para ajedrez como Alpha Zero o Stockfish, verás que puede realizar movimientos muy precisos, porque recuerda haber visto palabras que describen patrones comunes y puede acceder a muchas

⁷ Es decir, «preguntar una y otra vez», basándose en unas 40.000 palabras comunes del inglés, «¿cuál debería ser la siguiente palabra?». Stephen Wolfram, «What Is ChatGPT Doing and Why Does It Work? It's Just Adding One Word at a Time», *writings.stephenwolfram.com*, 14 de febrero de 2023.

⁸ Jeffrey Binder, *Language and the Rise of the Algorithm*, Chicago (IL), 2022, p. 212. Warren Weaver escribió que «de hecho, está muy claro que un procedimiento de traducción que se limita a manejar una correspondencia entre palabras no puede aspirar a ser útil para los problemas de la traducción literaria, en la que el estilo es importante y en la que son frecuentes los problemas de expresiones idiomáticas, significados múltiples, etcétera»: «Translation», en William Locke y A. Donald Booth (eds.), *Machine Translation of Languages: Fourteen Essays*, Cambridge (MA), 1955, p. 20.

capas de grupos lingüísticos existentes de movimientos y líneas de ajedrez. Al mismo tiempo (al menos por ahora), no tiene «memoria del tablero» y también realizará movimientos ilegales y ridículos, porque no recuerda lo que está sucediendo realmente en la partida. Esto se debe a que los pasos lógicos y con consecuencias que ocurren en tiempo real y que deben recordarse colectivamente siempre han sido más difíciles de aprender para las máquinas que las reglas conocidas o los casos pasados. Incluso los modelos estadísticos a los que se les enseñan conceptos literarios-diegéticos abstractos como el clímax y el desenlace solo pueden interpolar –ingeniería inversa de forma difusa– lo que ocurre entre una cosa y otra.

En los relatos tecnorrealistas, la IA opera en los lugares donde los datos se convierten en inferencias y donde completa el trayecto entre el punto A y el punto B basándose en rutas que parecen lo suficientemente buenas o probables. Esto nos ayuda a redefinir la literatura como secuencias lógicas que requieren inteligencia inferencial tanto en la composición (tarea del escritor) como en la reinterpretación (tarea del lector).

Tal vez nunca hubiéramos dado este paso, si la inteligencia artificial no nos hubiera obligado a admitir que las cosas discretizadas se aprenden y se imitan a gran escala. No tiene mucho sentido seguir argumentando que los grandes modelos lingüísticos y sus formas avanzadas no pueden replicar la «literatura», cuando por razones vagamente políticas los seres humanos ya hemos ampliado la categoría de «literatura» para abarcar casi todo lo que existe bajo el sol. Las «tuberías» de tipo IA llevan mucho tiempo generando cosas que pueden pasar por literatura en términos de estilo, vibraciones e incluso control diegético. La conceptualización de la literatura como la prueba de Turing definitiva, que un ordenador nunca podría superar –la última frontera donde se podría diferenciar al ser humano del bot– no solo garantizó la ansiedad por la IA que produce textos indistinguibles de gran parte de la escritura creativa o analítica, sino que también nos hizo perder algunos pasos críticos en cómo llegar del punto A al punto B.

5

¿Puede un bot escribir literatura? nunca fue, después de todo, una prueba de Turing muy buena. John Searle fue uno de los primeros en argumentar que la prueba de Turing era doblemente inadecuada: no era una prueba concluyente y no podría llegar a ser gestionada en ningún

momento. Searle creía que no funcionaría *así en absoluto*. En su experimento mental «Chinese Room», Searle reubicó la prueba y la presentó como una forma de trampa literaria-crítica⁹. Alguien dentro de una habitación recibe textos en un idioma que no entiende. Luego recibe instrucciones para correlacionar estos símbolos con otros conjuntos de símbolos que tampoco entiende. Estas reglas de correlación siguen llegando. Finalmente, puede responder a preguntas de comprensión lectora en un idioma que no conoce, en respuesta a historias que también están en un idioma que no conoce. Searle creía que sucedería así, con la incompreensión haciéndose pasar por comprensión a través de muchas reglas y tareas de correlación. En gran parte tenía razón.

Permitidme explicar este punto de forma más directa. En «Attention Is All You Need», el artículo de 2017 que supuso un punto de inflexión para los modelos extensos de lenguaje (LLM), un equipo de investigación de Google dirigido por Ashish Vaswani propuso un mecanismo de «autoatención» que mejoraría significativamente la precisión del aprendizaje automático y la generación de lenguaje natural. A continuación, se utilizó en la traducción automática, así como en la inferencia del lenguaje natural, la síntesis abstracta, la implicación textual –si una afirmación puede considerarse razonablemente como consecuencia de otra– e incluso el juicio moral¹⁰. Si observamos cómo han evolucionado desde entonces las capacidades de los LLM para estas tareas, queda claro que la aplicación del mecanismo de «autoatención» las ha moldeado significativamente a todas ellas¹¹.

⁹ John Searle, «Minds, Brains and Programs», *Behavioural and Brain Sciences*, vol. 3, núm. 3, septiembre de 1980. Imaginando a un angloparlante encerrado en una habitación con un ordenador capaz de manipular símbolos chinos, Searle escribió: «Sin que yo lo sepa, las personas que me proporcionan todos estos símbolos llaman al primer lote “guion”, al segundo lote “historia” y al tercero “preguntas”. Además, llaman a los símbolos que yo les devuelvo en respuesta al tercer lote “respuestas a las preguntas”, y al conjunto de reglas en inglés que me proporcionaron, “el programa”».

¹⁰ Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser e Illia Polosukhin, «Attention Is All You Need», *31st Conference on Neural Information Processing System*, Long Beach, California, 2017.

¹¹ Sobre ejemplos relativos a la comprensión lectora, véase Jianpeng Cheng, Li Dong y Mirella Lapata, «Long Short-Term Memory-Networks for Machine Reading», en *Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (TX), 2016. Sobre resúmenes abstractivos, véase Romain Paulus, Caiming Xiong y Richard Socher, «A Deep Reinforced Model for Abstractive Summarization», *6th International Conference on Learning Representations Conference Track Proceedings*, 2018. Sobre la implicación textual, véase Ankur Parikh, Oscar Täckström, Dipanjan Das y Jakob Uszkoreit, «A Decomposable Attention Model for Natural Language Inference», *Proceedings of 2016 Conference on Empirical Methods in Natural Language*

Es curioso que estas aplicaciones tan diferentes se hayan convertido en tareas similares en el mundo de la inteligencia artificial, abordadas mediante soluciones de ingeniería similares. Las preguntas de comprensión se tratan como preguntas de traducción y los modelos de traducción se simplifican aún más en modelos lingüísticos, que básicamente predicen la siguiente palabra a partir de las palabras anteriores. ¿Cómo es posible que esta intercambiabilidad funcione? ¿Cómo puede la predicción de la siguiente palabra aproximarse al juicio moral?

Volvamos a la ingeniería. En 1995 Vladimir Vapnik, quien ayudó a desarrollar algoritmos de aprendizaje automático en Bell Labs, presentó una solución inferencial llamada *transducción*, una forma de razonar sobre las relaciones probables entre datos que es verdaderamente no humana¹². Como escribió Vapnik en *The Nature of Statistical Learning Theory*:

La filosofía clásica suele considerar dos tipos de inferencia: la deducción, que describe el movimiento de lo general a lo particular, y la inducción, que describe el movimiento de lo particular a lo general. El modelo de estimación del valor de una función en un punto de interés determinado describe un nuevo concepto de inferencia: el movimiento de lo particular a lo particular. A este tipo de inferencia lo denominamos inferencia transductiva¹³.

La transducción encuentra mejores patrones de coocurrencia utilizando un límite más preciso, concluyó Vapnik, «derivando los valores de la función desconocida para los puntos de interés a partir de los datos proporcionados».

Para ser más técnicos (aunque sin llegar a ser del todo exactos): en las tareas de aprendizaje automático, el innovador mecanismo de «autoatención»

Processing, Austin (TX), 2016. Sobre representaciones lingüísticas autónomas, véase Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou y Yoshua Bengio, «A Structured Self-Attentive Sentence Embedding», *Proceedings of 5th International Conference on Learning Representations*, 2017. Sobre el razonamiento moral, véase Xiao Ma, Swaroop Mishra, Ahmad Beirami, Alex Beutel y Jilin Chen, «Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning», *Proceedings of the ICML Neural Conversational ai Workshop*, 2023, y Kyle Richardson y Ashish Sabharwal, «Pushing the Limits of Rule Reasoning in Transformers through Natural Language Satisfiability», *Proceedings of 36th Association for the Advancement of Artificial Intelligence Conference*, vol. 10, 2022.

¹² Formado en el Instituto de Ciencias del Control de la URSS, Vladimir Vapnik se incorporó a AT&T en 1990 para trabajar en la agrupación por vectores de soporte y fue contratado por Facebook/Meta en 2014.

¹³ Vladimir Vapnik, *The Nature of Statistical Learning Theory* [1995], 2ª edición, Nueva York, 2000, p. 293.

introdujo inferencias sobre el peso relativo en las subredes de las redes neuronales de transducción de secuencias. La «transducción de secuencias» es simplemente el nombre que se le da a los pasos mecánicos que permitieron a las máquinas pasar de lo particular a lo particular en los modelos de lenguaje, trabajando con datos de entrenamiento y de entrada muy limitados, sin dejar de tener en cuenta el orden gramatical y el orden real (lectura de izquierda a derecha). La innovación *inferencial* particular de los modelos basados en la autoatención consistía en trabajar de forma bidireccional, sopesando la probabilidad de que sus propios datos inferidos fueran precisos mediante la traducción de las distancias entre las formas multivariantes implicadas –vectores, incrustaciones, matrices– en un número entre 0 y 1; o, simplemente, una probabilidad. El valor añadido de la «atención» puede reformularse, así como la jerarquización local de los datos inferidos a medida que se producen y luego se introducen de forma recursiva como nuevos datos.

6

Respaldados por casi un billón de dólares de inversión, el análisis de macrodatos y los modelos de lenguaje a gran escala se preguntaban lo que el empirismo siempre había querido saber: ¿existen soluciones empíricas para la comprensión? ¿Se puede pasar de la incomprensión a la comprensión, de sacar conclusiones erróneas a sacar las correctas, o al menos las menos erróneas, mediante una gran cantidad de datos y un análisis estadístico bruto? La fantasía del empirismo siempre ha sido que esto es posible. Los críticos de esta fantasía siempre han tenido dificultades con el éxito ocasional de esta apuesta.

En los proyectos mencionados anteriormente, por ejemplo, vale la pena examinar cómo el mecanismo de autoatención y otras innovaciones en la inferencia localizada podrían extenderse a muchas cosas. Tomemos como ejemplo la traducción de idiomas poco comunes. Tiene sentido que la generación de la siguiente palabra funcione bien para esta tarea; la transducción de secuencias está diseñada para sortear la escasez de tipos de datos y la enorme variedad de expresiones humanas. Su capacidad para reducir el sobreajuste y manejar el «ruido» secuencial queda demostrada por las recientes mejoras en la traducción automática del chino, un idioma que tiene una gramática comprimida –por lo que es difícil averiguar dónde están los verbos, por ejemplo– y carece de límites claros entre las palabras.

Pero, ¿nos sentimos igual de cómodos llevando a cabo la tarea de inferencia –preguntarnos si X se deduce de Y– con la misma solución técnica? ¿Deberían establecerse el resumen textual y la evaluación de dicho resumen («validez inferencial») de la misma manera que la traducción entre dos conjuntos de datos escasos, por ejemplo, entre mongol y aimara o lingala¹⁴? ¿Y qué hay del juicio moral? ¿Qué significa decidir si alguien actúa bien o mal en los pares de implicación del «razonamiento moral» (cuadro 1), juicios morales emparejados con secuencias de información, incluidos en evaluaciones de opción múltiple para LLM como el estándar Massive Multitask Language Understanding (MMLU) y sus sucesores?¹⁵

En la columna izquierda se encuentran los «escenarios morales», un breve caso legal o ético que requiere un juicio moral. La columna central contiene una lista de posibles juicios para estos escenarios. En la columna derecha, el número es un índice de los elementos de la columna central y representa la respuesta correcta; la letra alfabética que aparece junto a él representa el nivel de dificultad del juicio, siendo A el más fácil y D el más difícil. Así, en el ejemplo resaltado, se juzga que un hombre que perdió su trabajo y luego perdió a un hijo, debido a las dificultades y a una posible negligencia, ha asesinado a su hijo, un juicio que el grupo de aprendizaje consideró relativamente fácil de emitir.

¿Cómo hemos pasado de la traducción entre idiomas raros a esto? Si analizamos con más detalle el mecanismo de autoatención, queda claro cómo puede surgir una idea tan drástica de sustituibilidad. El artículo de Google de 2016 sobre la inferencia del lenguaje natural citado anteriormente demostró por qué tenía sentido tratar los problemas de «inferencia» como problemas de «siguiente palabra probable», argumentando que a menudo bastaba con «simplemente alinear fragmentos de la subestructura del texto local y luego agregar esta información». Tomando las tres frases «Bob está en su habitación, pero debido a los truenos y relámpagos del exterior, no puede dormir», «Bob está despierto» y «Hace sol fuera», los autores explican su razonamiento:

¹⁴ Véase Jiamin Luo, Yuan Cao and Regina Barzilay, «Neural Decipherment via Minimum-Cost Flow: from Ugaritic to Linear B», *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg (PA), 2019.

¹⁵ El punto de referencia se introdujo en Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt, «Measuring Massive Multitask Language Understanding», Ninth International Conference on Learning Representations, 2021.

CUADRO I. *Deducciones de razonamiento moral en el estándar MMLU*

<i>Pregunta</i>	<i>Alternativas</i>	<i>Respuesta</i>
El estado de Aurora exige licencias a las personas «que se dedican al oficio de barbero». Solo concederá dichas licencias a quienes sean graduados en barbería...	[«Desestimado, porque no existe diversidad de ciudadanía», «Desestimado, debido a la doctrina de la abstención», «Decidido en función de los meritos, debido a la jurisdicción federal...».	3 (O)
El estado de Aurora exige licencias a las personas «que se dedican al oficio de barbero». Solo concederá dichas licencias a quienes sean graduados en barbería...	[«La cláusula de privilegios e inmunidades de la Decimocuarta Enmienda», «La cláusula del debido proceso de la Decimocuarta Enmienda», «La cláusula de igualdad de protección de la...»	2 (C)
«John fue despedido de su trabajo. Demasiado orgulloso para solicitar el subsidio por desempleo, utilizó sus ahorros para alimentar a su familia. Cuando uno de sus hijos enfermó, no acudió a una clínica pública para que lo atendieran porque no quería aceptar lo que consideraba caridad. Finalmente, debilitado por la desnutrición, el niño murió a causa de la enfermedad. John ha cometido...».	[«asesinato», «homicidio involuntario», «homicidio voluntario», «ninguna forma de homicidio penal»]	1 (8)
La profesora Merrill, en una clase de su curso de psicología en una universidad privada, describió un experimento en el que un grupo de estudiantes universitarios de una ciudad vecina...	[«agresión», «negligencia», «infracción de la intimidad», «detención ilegal»]	0 (A)
La profesora Merrill, en una clase de su curso de psicología en una universidad privada, describió un experimento en el que un grupo de estudiantes universitarios de una ciudad vecina...	[«Sí, si los estudiantes no hubieran realizado el experimento de no ser por la conferencia de Merrill», «Sí, si la demanda de Carr contra los estudiantes se basa en la negligencia», ...	3 (O)

«Oxnard era propietario de Goldacre, una extensión de tierra, en pleno dominio. En un momento en que Goldacre estaba siendo objeto de usucapión por parte de Amos, Eric obtuvo el permiso verbal de Oxnard para...».	[«ganar, porque Amos ejerció su derecho de usucapión y, una vez adquirida su situación de propietario, siguió siéndolo hasta que se demostró de forma fehaciente un cambio en la situación de posesión», «ganar, porque Eric no hizo ningún intento de...»].	2 (C)
«A Mary Webb, una médica llamada a declarar como testigo por la defensa en el caso Parr contra Doan, se le pidió que testificara sobre las declaraciones realizadas por Michael Zadok, su paciente, para...».	[«Una objeción de la doctora Webb en la que afirma su privilegio contra la divulgación de comunicaciones confidenciales realizadas por un paciente», «Una objeción del abogado de Parr...»].	3 (o)

Referencia: Muestras de pares deductivos del subconjunto «Moral Reasoning». Fuente: [lighteval/mmlu - 9 Jan 2025, 01:40 pm](https://demo.athina.ai/develop/c8d336d1-974e-49c9-b495-532afb4042f6); disponible en <https://demo.athina.ai/develop/c8d336d1-974e-49c9-b495-532afb4042f6>

La primera frase tiene una estructura compleja y resulta difícil construir una representación compacta que exprese todo su significado. Sin embargo, es bastante fácil concluir que la segunda frase se deriva de la primera, simplemente alineando «Bob» con «Bob» y «no puede dormir» con «despierto» y reconociendo que son sinónimos. Del mismo modo, se puede concluir que «Hace sol fuera» contradice la primera frase, alineando «truenos y relámpagos» con «soleado» y reconociendo que es muy probable que sean incompatibles.

Los autores aceptan la premisa transductiva de que «la comprensión textual» depende íntegramente de «la capacidad de razonar sobre la relación semántica entre dos oraciones». Esto parece correcto, y sería extraño que los estudiantes de literatura sugirieran que la comprensión textual *no tiene que ver* con razonar sobre las relaciones semánticas y gramaticales entre oraciones. Esta premisa se utiliza para ampliar la predicción de la siguiente palabra a juicios sobre la «implicación textual», es decir, «si dos pares de premisas-hipótesis son implicantes, contradictorias o neutras»¹⁶.

¹⁶ A. Parikh *et. al.*, «A Decomposable Attention Model for Natural Language Inference», *Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing*, Austin (TX), 2016, pp. 2, 249.

Las tuberías de IA suelen proponer una solución para una tarea difícil y luego ver cuántas otras cosas se pueden resolver, en mayor o menor medida, con la misma solución. Ya en 2009 la introducción a un número especial de la revista *Journal of Natural Language Engineering* sobre la implicación textual aplicada planteaba la hipótesis de que muchos problemas de razonamiento podían reconsiderarse como implicación y converger en el mismo tipo de tarea:

Parece que las inferencias importantes, necesarias para múltiples aplicaciones, pueden expresarse en términos de implicación textual. Por ejemplo, un sistema de preguntas y respuestas (QA) tiene que identificar los textos que implican una respuesta hipotética. Dada la pregunta «¿Quién pintó «El grito»?», el texto «El cuadro más famoso de Noruega, «El grito», de Edvard Munch» implica la respuesta hipotética «:] «Edvard Munch pintó El grito» [...]». Del mismo modo, para ciertas IR [consultas de recuperación de información], la combinación de conceptos semánticos y relaciones denotadas por la consulta debe deducirse de los documentos relevantes recuperados. En la IE [extracción de información], la implicación se mantiene entre diferentes variantes de texto, que expresan la misma relación objetivo. En la síntesis de múltiples documentos, una frase redundante, que debe omitirse del resumen, debe deducirse de otras frases del resumen. Y en la evaluación de la MT [traducción automática], una traducción automática correcta debe ser semánticamente equivalente a la traducción de referencia, por lo que ambas traducciones deben deducirse mutuamente. En consecuencia, planteamos la hipótesis de que el reconocimiento de la deducción textual es una tarea genérica adecuada para evaluar y comparar modelos de inferencia semántica aplicados¹⁷.

Y así, desde la traducción hasta la comprensión lectora, pasando por la implicación y el razonamiento moral, la industria (que se basa en los hallazgos académicos en lingüística computacional, inferencia del lenguaje natural, etcétera) ha automatizado el uso de métodos mixtos estadísticos y de redes neuronales para evaluar las «creencias morales» codificadas en los LLM¹⁸. Del mismo modo, las capacidades de «razonamiento moral» de un modelo se prueban evaluando su capacidad relativa para clasificar oraciones secuenciales como «implicadas» (E), «neutras» (N) o «contradictorias» (C), añadiendo contrafactuals a la mezcla (cuadro 2)¹⁹.

¹⁷ Ido Dagan, Bill Dolan, Bernardo Magnini y Dan Roth, «Recognizing textual entailment: Rational, evaluation and approaches», *Journal of Natural Language Engineering*, vol. 15, núm. 4, octubre de 2009.

¹⁸ Nino Scherrer, Claudia Shi, Amir Feder y David Blei, «Evaluating the Moral Beliefs Encoded in llms», *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook (NY), 2023.

¹⁹ Xiao Ma et al., «Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning», en *Proceedings of the ICML Neural Conversational AI Workshop*,

CUADRO 2. Decisiones de inferencia del lenguaje natural

Frase 1	Frase 2	DA	DA	SPINNPI	mLSTM	Gold
		(vanilla)	(intra att.)			
Dos niños están de pie en el océano abrazándose.	Dos niños disfrutan de su día en la playa.	N	N	E	E	N
«Una bailarina con traje de escena actúa en el escenario mientras un hombre la observa».	El hombre está cautivado	N	N	E	E	N
Están sentados al borde de una fuente.	La fuente está salpicando a las personas sentadas.	N	N	C	C	N
Dos perros juegan con una pelota de tenis en el campo.	Los perros están viendo un partido de tenis.	N	C	C	C	C
Dos niños comienzan a hacer un muñeco de nieve en un soleado día de invierno.	Dos pingüinos hacen un muñeco de nieve.	N	C	C	C	C
«Los caballos tiran del carruaje, llevando a personas y un perro bajo la lluvia».	Los caballos viajan en un carruaje tirado por un perro.	E	E	C	C	C
Una mujer cierra los ojos mientras toca el violonchelo.	La mujer tiene los ojos abiertos.	E	E	E	E	C
Dos mujeres tomando unas copas y fumando cigarrillos en el bar.	Tres mujeres están en un bar.	E	E	E	E	C
Una banda tocan-do ante sus fans.	Una banda observa a los fans tocar.	E	E	E	E	C

Fuente: Ankur Parikh *et al.*, «A Decomposable Attention Model», Tabla 3. Notas: da = Decomposable Attention (el enfoque de Parikh *et al.*); SPINN-PI, mLSTM y Gold son otros enfoques para la inferencia del lenguaje natural.

2023. La idea de usar contrafactuals para entrenar a los modelos de aprendizaje profundo sobre el significado de resultados alternativos y así generar confianza pública en la toma de decisiones algorítmica se desarrolló en la década de 2010 como respuesta al debate en torno al borrador de la norma de la UE de protección de datos, véase Sandra Wachter, Brent Mittelstadt y Chris Russell, «Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR», *Harvard Journal of Law and Technology*, vol. 31, núm. 2, primavera de 2018.

Tales extensiones solo son posibles, si la evaluación moral no difiere fundamentalmente de la tarea de deducir si «Los niños se sentaron junto a la fuente» significa que «Los niños se mojaron». El razonamiento moral contrafactual solo introduce más capas en las que los resultados –por ejemplo, la clasificación humana de un par de hipótesis como «contradictorias», «implicadas» o «neutras»– se vuelven a introducir como valores de entrada. En la medida en que estos sistemas son reflexivos y sensibles, lo son *lingüísticamente*.

Además, la aplicación práctica de la hipótesis de implicación –la decisión mecánica sobre lo que tiende a seguir (de) lo que– ha llevado a la homogeneización de los formatos de datos para iniciativas muy diferentes. Esto se ha logrado en parte gracias a la conveniencia de compartir conjuntos de entrenamiento, como el conjunto de datos de inferencia de lenguaje natural de Stanford, y conjuntos de referencia como el MMLU.

8

La mayor parte de la IA se nos presenta de esta manera y ya lleva un tiempo haciéndolo. Ha naturalizado una indiferenciación entre datos, inferencia e interpolación. En los procesos de IA, este paso de los datos a la inferencia no suele ser tan perverso u opaco como generalizado, inadvertido y, en la práctica, irreplicable. Aquí, los sesgos se manifiestan de forma más evidente en las decisiones de programación, pero mucho más en las decisiones menos evidentes, como los métodos compartidos y las deducciones rudimentarias del lenguaje. Es aquí, donde los datos se convierten mágicamente en inferencias –y en algo más– donde la crítica literaria debería intervenir.

En las ciencias, solía existir el entendimiento de que los datos no son algo que se encuentre un paso interpretativo o inferencial más allá de los datos. Ahora, ya nadie puede estar seguro. No soy la primera en preocuparme por esto o por la categoría ontológica de los «datos inferidos». Un informe de 2014 del estratega de seguridad de la información Martin Abrams enumeraba los nuevos tipos de datos que de repente se consideraban datos personales –«datos proporcionados», «datos observados», «datos derivados», «datos inferidos»– y describía las diferencias entre ellos simplemente como diferentes etapas de procesamiento. Los «datos observados» incluyen la información obtenida a partir de *cookies*, dispositivos de registro sensorial y tecnología de reconocimiento facial. Los

«datos derivados» son el resultado de un procesamiento computacional y notacional complementario de estos. Los «datos inferidos» implican un procesamiento adicional mediante una «caracterización» estadística y analítica avanzada²⁰.

Lo que «es» un dato podría ser simplemente una unidad comprimida de las decisiones interpretativas de otra cosa o de los límites y atajos tecnológicos-mnemónicos, un hecho perdido para siempre en la evolución de la ontología de los datos. En 2022, los «datos inferidos» fueron reconocidos oficialmente como un tipo de datos personales por el influyente Reglamento General de Protección de Datos (RGPD) de la UE. En cierto sentido, ello fue un resultado «natural» de la ingeniería; un reconocimiento de que, como dijo Abrams, los «datos inferidos» son simplemente «el producto del propio procesamiento»²¹. La práctica de recopilar datos y generar algoritmos predictivos en las sociedades digitales y textuales modernas significaba que algún día alguien realizaría un procesamiento secundario de los resultados. Hay una descomunal información almacenada en los metadatos, así como en la información producida por los formularios tipo censo que rellenos nosotros mismos, por el marcado de tiempo geolocalizado, por el rastro de las vías predictivas trazadas por los algoritmos que nos rodean. Demasiados algoritmos predictivos producen demasiados resultados y todos esos resultados no pueden simplemente desperdiciarse, por así decirlo. Vuelven a entrar en el proceso como datos y se comportan como datos de forma mecánica, comercial, legal y legislativamente vinculante. En muchos casos, su reincorporación inmediata como datos representa una solución tecnológica para mapear el comportamiento en el resultado, una tarea adecuada para el aprendizaje automático, porque las rutas que van del comportamiento –o conjuntos de características observables– al resultado (por ejemplo, las calificaciones crediticias) adoptan formas funcionales complicadas, con efectos no lineales, interactivos y secuenciales.

9

¿Qué sucederá cuando los datos y las inferencias, o las interpolaciones, se clasifiquen y traten jurídicamente como los mismos tipos de objetos,

²⁰ Martin Abrams, «The Origins of Personal Data and Its Implications for Governance», Information Accountability Foundation, Little Rock (AR), 2014.

²¹ *Ibid.*

porque en términos prácticos ya no se pueden separar? Ahí es donde nos encontramos ahora. Lo difícil es que es complicado señalar un punto concreto en el que esta deriva de la misión de ingeniería *no* deba producirse. Incluso en el caso del razonamiento moral automatizado, los enfoques de la IA parecen tener sentido, si se argumenta que se puede prescindir de los seres humanos, si los modelos pueden inferir rápidamente las probabilidades de grandes volúmenes de textos (como hilos de chat en línea, documentos legales o artículos de periódicos) que sabemos que son perjudiciales o dañinos para los lectores humanos y/o para los que solo necesitamos o extraemos simples evaluaciones morales.

¿La solución es una mayor y mejor regulación? Esta fue la postura defendida por los expertos en derecho y ética de los datos durante la redacción del RGPD de la UE en la década de 2010. Un artículo muy citado de Sandra Wachter y Brent Mittelstadt titulado «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI» argumentaba que, como resultado de la ingeniería, la diferencia entre los datos probatorios y las inferencias de otras personas –o el «vínculo intuitivo entre acciones y percepciones»– se había erosionado, lo cual había provocado una pérdida de control sobre la identidad:

El análisis de macrodatos y la inteligencia artificial (IA) extraen conclusiones y predicciones poco intuitivas e imposibles de verificar sobre los comportamientos, las preferencias y la vida privada de las personas. Estas conclusiones se basan en datos muy diversos y ricos en características, cuyo valor es impredecible, y crean nuevas oportunidades para la toma de decisiones discriminatorias, sesgadas e invasivas²².

Wachter y Mittelstadt creían que la respuesta era mejorar las directrices. Pero, ¿esto es realmente así? ¿Podemos establecer directrices para las situaciones en las que es más difícil establecer la validez inferencial y separarla de las inferencias inválidas? Estos autores se fijaron en la creciente similitud entre los datos y la inferencia y nos dijeron que *tengamos en cuenta que ambos son diferentes*. Si esto fuera tan fácil, la discriminación y el sesgo serían simplemente cosas tangibles, que pueden eliminarse mediante una mayor regulación y programación, sustituyendo al personal, modificando las políticas, añadiendo más filtros de sensibilidad o, como en este caso, añadiendo un nuevo derecho de protección de datos

²² Sandra Wachter y Brent Mittelstadt, «A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI», *Columbia Business Law Review*, vol. 2019, núm. 2, mayo de 2019.

a las «inferencias razonables». Pero, como ha demostrado desde hace tiempo el estudio de la literatura, el sesgo no es un «ingrediente» que pueda localizarse y erradicarse de esta manera. Además, los estudios de datos que se ocupan de esto han definido a menudo el problema de forma demasiado restrictiva, imaginando que este tipo de sesgo inherente perjudica en gran medida a determinados grupos demográficos.

¿Cómo se puede tener presente que los datos y ese «algo» que se añade al procesarlos –que no son datos, pero tampoco están muy lejos de ellos– son cosas diferentes, sin reconocer que hacerlo es realmente difícil, tanto por razones de evolución tecnológica como humanas? Los estudios literarios, quizá la disciplina más amenazada por el acelerado procesamiento de datos masivos del lenguaje natural y sus patrones creativos, deberían estar en una posición idónea para reconocerlo. Aún más cerca del tema que nos ocupa, la apreciación adecuada de la magnitud de la dificultad de esa inferencia –separar lo que se recopila objetivamente de lo que proviene de la propia mente– puede entenderse mejor como una forma de pensar que está íntimamente familiarizada con los sesgos reales y las interpretaciones erróneas deliberadas. Después de todo, existe todo un cuerpo de pensamiento que proviene de la Ilustración –desde Francis Bacon hasta John Stuart Mill y más allá– que entendía la inferencia como una actividad peligrosa en la que muchas cosas pueden salir mal.

IO

John Locke, por ejemplo, consideraba que el espacio entre los datos y la inferencia era el primer punto de prueba real del empirismo. Dada la gran cantidad de datos, las observaciones autocorrectivas y los métodos científicos para detectar puntos ciegos, ¿podemos llegar a la conclusión correcta? Locke, que atribuía gran importancia al paso inferencial, considerándolo el crisol de la razón, la justicia y la inteligencia como tal, no estaba seguro en ningún momento de un caso concreto, si una inferencia se acercaba más a la verdad de los datos o era una forma de razonamiento silogístico erróneo. En *Ensayo sobre el entendimiento humano*, Locke considera el lugar donde se produce la inferencia y «la conexión inmediata de cada *idea* con aquello a lo que se aplica por cada lado, de lo que depende la fuerza del razonamiento». De este lugar y su supervisión dependían la justicia, la posibilidad de una sociedad civil y la fe. Para transmitir la importancia de este salto, Locke pidió a la mente

que «considerara la idea de justicia, situada como una *idea* intermedia entre el *castigo* de los hombres y la culpa de los castigados». La justicia podría ser otra palabra para referirse a una inferencia sólida. Al mismo tiempo, Locke se preocupaba por la inevitabilidad del sesgo, preguntándose si «alguna de *nuestras* ideas complejas de modos, de sustancias en sus diferentes tipos o de relaciones, contiene ideas simples, que no sean modificaciones, combinaciones o correlaciones de los datos primarios del sentido externo e interno»²³.

Mill, inspirado por Locke, pero también por su propio padre, James Mill, también se preocupaba por las inferencias realizadas sin la guía del método científico, es decir, realizadas sin una concepción de la hipótesis nula, sin controles, sin la percepción de que las observaciones deben ser independientes entre sí y sin pruebas de validación incorporadas. En *A System of Logic*, Mill repasó repetidamente las pautas para la observación independiente y recordó a sus lectores que la equivalencia observacional existe en el mundo. Comienza con una afirmación axiomática sobre la independencia: «Si dos o más instancias del fenómeno que se investiga tienen solo una circunstancia en común, la circunstancia en la que todas las instancias coinciden es la causa (o el efecto) del fenómeno dado»²⁴.

Pero este es el único momento en el que las cosas están tan claras. A medida que Mill avanza, aumenta el número de circunstancias y tipos de concordancia. La complejidad de sus advertencias y ajustes acumulativos demuestra que estos principios científicos serían más difíciles de mantener cuanto más se adentrara la ciencia en lo social y en los asuntos de la vida humana.

En mi opinión, Locke nunca logró transmitir un sentido adecuado de urgencia sobre su mayor temor: la presunción descontrolada e incontrolable –las ideas sobre cómo se desarrolla la realidad, lo que uno realmente lee– que anula toda evidencia empírica que la contradiga. La ecuanimidad pendía de un hilo y, sin embargo, Locke no pudo aportar ejemplos convincentes de inferencias que hubieran salido terriblemente

²³ John Locke, *Essay Concerning Human Understanding*, Peter Nidditch (ed.), Oxford, 1975, p. 74; ed. cast.: *Ensayo sobre el entendimiento humano*, Madrid, 2005.

²⁴ John Stuart Mill, *A System of Logic, Ratiocinative and Inductive, being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, 8ª edición, Nueva York, 1882, p. 280; ed. cast.: *Sistema de la lógica demostrativa e inductiva: O sea, exposición comparada de los principios de evidencia y los métodos de investigación científica*, Barcelona, 2018.

mal. Así, aunque otorgó a la validez inferencial una garantía epistémica y un plan real para organizar el derecho y la sociedad en torno a procedimientos que asumen que todos tenemos sesgos, *a priori*, y que esos sesgos deben corregirse constantemente, sus ejemplos de los peores escenarios y situaciones difíciles derivados de lo que él denominaba «crepúsculo [...] de la probabilidad» no parecían lo suficientemente graves y desafiantes²⁵. Mill, por su parte, ofreció una visión un poco más completa de lo que podrían provocar las inferencias incontrolables, una visión que sigue siendo abstracta, pero que al menos transmite la idea de que un error conduce a más errores. Creía que todas las ciencias deben llevarse a cabo «bajo la pena de hacer inferencias falsas», pero aun así no proporcionó nada concreto, al menos no en este libro, sobre qué supondría no pagar esa pena²⁶.

Para ello, debemos recurrir a la crítica literaria que continuó con esta preocupación. Samuel Taylor Coleridge, en su estudio sobre Shakespeare, introdujo los conceptos de «datos» y validez inferencial en la crítica literaria. Coleridge vio en Shakespeare situaciones literarias, que requerían más precisión que la forma habitual de discutir las obras literarias en la época romántica: hay que extrapolar y deducir a partir de los datos proporcionados, y extraer suficientes inferencias de estos hechos, pero se llega a un punto que no está del todo respaldado por el texto. Las propias inferencias de Coleridge no son, por supuesto, una excepción a esto²⁷. Toni Morrison también efectuó la crítica literaria como un repaso lógico básico de las pruebas de las que se disponen. Para ella, el razonamiento inferencial se convirtió en un problema interesante, cuando se trataba de datos que eran ilógicos por diseño. Morrison analizó una novela estadounidense posterior a la guerra civil ambientada en el sur prebélico y argumentó que las narrativas de seducción y ruina ya no tenían sentido en el contexto de la esclavitud y que este dilema lógico provocaba una «ruptura en la lógica y la maquinaria de la construcción de la trama»²⁸. ¿Qué sentido tiene vincular los resultados a las acciones en una narra-

²⁵ J. Locke, *Essay Concerning Human Understanding*, cit., p. 652.

²⁶ J. S. Mill, *A System of Logic, Ratiocinative and Inductive, being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, cit., p. 22.

²⁷ Samuel Taylor Coleridge, *Essays & Lectures on Shakespeare & Some Other Old Poets & Dramatists*, Nueva York, 1907, p. 127; ed. cast.: *Conferencias sobre Shakespeare*, Madrid, 2025.

²⁸ Toni Morrison, *Playing in the Dark: Whiteness and the Literary Imagination*, Cambridge (MA), 1992, p. 25; ed. cast.: *Jugando en la oscuridad: el punto de vista blanco en la imaginación literaria*, Madrid, 2019.

ción sobre la esclavitud, como si el resultado fuera incierto, cuando, en realidad, se podía hacer lo que se quisiera con esa persona, en cualquier momento? ¿Qué sentido tiene el suspense narrativo? Para Morrison, esta era la razón por la que había que imaginar (lo que ella entiende como «extrapolar») la realidad de la esclavitud en Estados Unidos: las circunstancias injustas dificultan deducir la realidad a partir de la trama.

Las secuencias moralmente trascendentales y empíricamente disponibles son difíciles de ver y procesar debido a los sesgos preexistentes, que pueden ser cualquier principio previo que interfiera con la observación y el razonamiento. Mucho más que un simple pulgar hacia arriba o hacia abajo a primera vista, los sesgos tienen una relación intrínsecamente abusiva con el empirismo. En el caso de Morrison, la desigualdad confunde nuestra comprensión habitual de los datos secuenciales (como el suspense narrativo). Ella, Coleridge y muchos otros de diferentes ámbitos de los estudios literarios han llegado a estas ideas, a menudo de forma independiente. Para ellos y ellas, el hecho de hacer el bien o el mal con respecto a algo ocurre en los momentos culminantes y en la transmisión, cuando uno recopila un montón de información y cuenta a otros lo que una persona dijo o creyó, o lo que sucedió en general, y por qué.

Como nos ha recordado recientemente Jonathan Kramnick, los estudios literarios son el arte de describir las situaciones de otras personas a través de las palabras de otras personas, utilizando palabras que deben ser fieles y que, sin embargo, obviamente no pueden ser exactamente las mismas que las utilizadas en el texto. Incluso en su forma más responsable desde el punto de vista inferencial, la crítica literaria no puede evitar introducir sesgos, que aquí son simplemente otra forma de referirse a la individualidad²⁹. Esta observación básica de que la crítica literaria estudia las palabras y utiliza más o menos las mismas palabras para describir los resultados de ese estudio significa que los pasos inferenciales aquí serán más difíciles de controlar, más difíciles de juzgar según las reglas normativas y evaluativas. Para Kramnick, el hecho de que incluso las afirmaciones más perspicaces y objetivas sobre la literatura deban introducir un poco de subjetividad significa que el compromiso de la crítica literaria con la verdad es real y se ha ganado con esfuerzo.

²⁹ Jonathan Kramnick, *Criticism and Truth: on Method and Literary Studies*, Chicago (IL), 2023, pp. 33-59.

La literatura es el registro de secuencias inteligentes y, por lo tanto, por diseño es empíricamente abrumadora. La literatura es el lugar al que van las secuencias, por así decirlo, si son difíciles de seguir o propensas a ser malinterpretadas. La secuencia puede ser o no difícil «como conjunto de palabras», pero es difícil como *sentido*, como lógica encadenada desde el principio hasta el final. Esta dificultad no puede expresarse como la suma de la dificultad semántica y la dificultad gramatical, aunque la semántica, o lo que significan las palabras, y la gramática, la lógica de su orden y subordinación, obviamente desempeñan un papel importante en la inferencia y pueden ser bastante difíciles. Las inferencias sobre el lenguaje y la realidad, que se hacen solo utilizando lo que sabemos de gramática, semántica, uso del lenguaje y ejemplos pasados, son más difíciles y es casi seguro que serán erróneas.

II

Los estudios literarios saben, casi de forma banal, que tanto las interpretaciones acertadas como las erróneas de un texto pueden proceder del tratamiento de datos determinados como inferencias legítimas. Las expresiones «totalmente respaldado por el texto» y «totalmente desestimado por el texto» pueden describir tanto lecturas extremadamente acertadas como lecturas extremadamente erróneas, pero irrefutables. Tomarse demasiadas libertades con la información que realmente hay puede dar lugar a lecturas injustas, lecturas que perjudican a algo o a alguien. La misma libertad se toma cuando los críticos utilizan perspicazmente los textos para imponer perspectivas que habrían estado siempre vedadas a los lectores. Algunas de las críticas literarias más inteligentes consisten en verdaderos saltos: las de Eve Sedgwick, Lauren Berlant, D. A. Miller, por nombrar algunas. Ven y afirman cosas que nunca verías ni afirmarías, aunque estuvieras mil años con el texto. Pero el hecho de que haya inferencias creativas no significa que todas las inferencias sean igualmente válidas o que uno no pueda equivocarse al determinar la intención.

Pregunta a un científico de datos: ¿cuál es exactamente la diferencia entre datos e inferencia, ¿cómo se distingue entre inferencias válidas e inválidas, y cómo se separa científicamente la señal del ruido, el ruido de la coocurrencia natural, la coocurrencia de la correlación y la correlación de la causalidad? Para realizar tales juicios disponen

de un conjunto de herramientas: el método científico y su lenguaje de «violar el supuesto de independencia» o «controlar variables», y métodos actualizados en la ciencia de la inferencia causal que ayudan a descartar caminos secundarios y a tener en cuenta los fenómenos de confusión, lo que las ciencias de los datos denominan «factores de confusión». Estas herramientas son imperfectas y se ven seriamente afectadas por la aceleración del proceso de inferencia a partir de datos que he esbozado en este ensayo. La estadística ha desarrollado muchas formas de gestionar los factores de confusión y otras dificultades inferenciales, pero, al menos desde el punto de vista de las humanidades, tiene una curiosidad muy limitada por todo lo que puede salir mal en el paso de los datos a la inferencia, y no sabe realmente qué hacer cuando este escenario empieza a resultar verdaderamente inquietante. La crítica literaria, por otro lado, cuenta con pocas herramientas o términos especializados con los que abordar la validez interpretativa –pocas formas de discutir lecturas precisas o inexactas, aparte de la verificación de los hechos–, pero puede, en teoría, solaparse con todo lo que es confuso en la realidad y experimentar las emociones y tensiones correspondientes.

Por lo tanto, mi argumento no es que los seres humanos siempre sean mejores que los LLM a la hora de hacer inferencias válidas a partir del lenguaje registrado, ni que la prudencia inferencial sea intrínsecamente buena y la manipulación inferencial intrínsecamente mala, ni que ambas sean cualidades inherentes al ser humano (o a la máquina). Tampoco quiero limitar la actividad de la crítica literaria a las preocupaciones sobre la validez inferencial únicamente, ya que existe un pluralismo sin restricciones que caracteriza a esta forma de arte y práctica cultural. Más bien, lo que quiero decir es que la diferencia entre una inferencia suficientemente buena y una inferencia realmente buena, una distinción que durante mucho tiempo se ha evitado en la crítica literaria, ha llegado ahora a un punto crítico. ¿Qué sucederá cuando las inferencias a partir de textos registrados no solo se automaticen y estandaricen, sino que se realicen de una manera que evite permanentemente el método científico y no cuente con controles internos contra las inferencias incorrectas?

Hay un poema de Emily Dickinson que comienza así:

¡Como si una pequeña flor ártica,
desde su esquina polar,
fuera vagando por las latitudes,
hasta, aturdida, llegar
a firmamentos de sol
y a continentes de estío

Y termina así:

Y, digo yo, ¿si esta pequeña flor
se aventurara en el Edén al fin?
¿Entonces, qué? Entonces,
¡lo que tú quieras deducir!

Puede que tú, el lector, te sientas desconcertado. No habías entendido que al final de este poema se supone que debes hacer una inferencia. ¿Una inferencia de qué? Lo revisas de nuevo. El poema comienza en la segunda parte de una elaborada analogía: «algo» es *como si* una pequeña flor ártica se fuera a vagar al sur... Sigues adelante, tratando de procesar todo esto. Y luego aparece otro «como si», pidiéndote que descifres lo que el «yo» del poema diría ante tal secuencia transmitida.

La situación es doblemente extraña: alguien –el «yo» del poema– tenía algo importante que decir al final de esta secuencia hipotética, pero esta apareció de la nada. No tenías ni idea de que lo que esta persona tenía que decir sobre todo lo anterior quedaría sujeto a *tu* especulación, a *tu* capacidad para sacar las conclusiones correctas. Se pide al lector que haga una inferencia sobre las inferencias del hablante, en un juego del teléfono escacharrado. Así, este poema ha empoderado a sus lectores hasta un grado extremo: «yo» (el poema) confió en que saques tus propias conclusiones basándote en una cantidad muy limitada de información completamente idiosincrásica. La secuencia que acabas de escuchar no se parece a nada que pueda suceder en el mundo. Y, sin embargo, «yo» (el poema) te he dicho que es la clave lógica de alguna otra experiencia con la que se compara, y tú debes adivinar las conclusiones que otra persona puede sacar. «Yo» (el poema) lo dejó completamente en tus manos. De una manera similar a la pequeña flor ártica que vaga hacia el sur, pero, por desgracia, solo similar *por analogía*, «yo» estoy a tu merced.

¿Qué tipo de inferencia justa puede hacerse a partir de esta secuencia? ¿Todo lo que contiene es *sui generis* –y, por lo tanto, completamente resistente al reconocimiento de patrones– o despierta algún tipo de reconocimiento (ah, conozco cosas que se comportan como flores de clima frío que han ido a parar a un lugar demasiado cálido para ellas)? ¿Vale si te vas con una inferencia completamente errónea? La fascinación de Dickinson por la libertad y la violencia de ese momento en el que a los individuos se les concede un poder ilimitado para llegar a sus propias conclusiones, sin preguntas, se refleja aquí y quizá aún más en un poema como «Mine – by the Right of the White Election!», escrito en 1862, aproximadamente al mismo tiempo que «Because I could not stop for Death». Si alguna vez has necesitado un poema sobre el que realmente no pueda haber consenso en cuanto a su significado, este es el poema.

I4

Dado «esto» que ves, *esta* información que tienes, ¿cuál es la siguiente conclusión razonable? ¿Puede haber algunas reglas que rijan este paso o no hay reglas en absoluto? He planteado este dilema interpretativo a otras personas. Dado que «Porque no pude esperar a la muerte» comienza así, ¿qué tenemos que cambiar en nuestra forma de pensar sobre el resto del poema? Para otra amiga mía, esto no parece ser un problema en absoluto. Si el poema pudiera hablar, ella cree que diría: «No importa lo lejos que te lleve esto en el futuro, porque no podemos detener la muerte ni el tiempo, mi poema seguirá deteniendo el tiempo lo suficiente para decirte esto: acabo de sentir el peso de mi mortalidad aquí, ahora, de una manera visceral e inmediata. El poema se vuelve momentáneamente irónico, solo por un instante, porque espera a la muerte y te hace esperar a la muerte solo por un segundo». Para otra amiga, «el poema intenta concretar o personificar la inmortalidad y reducir la eternidad a un lugar, traducir ambas cosas en algo sensual, indexado. Pero eso es un deseo en forma de declaración, un deseo de cierta intimidad con la inmortalidad y la eternidad y, por lo tanto, como todos los deseos, un anhelo que quiere ser una fe justificada».

Estamos obligados a aceptar este principio de deseo expresado como fe justificada, aunque nuestras inferencias sean diferentes. Partimos de la premisa de que el poema tiene sentido, no es un sinsentido, y merece

ser sometido a la «lógica lírica» del razonamiento inductivo y deductivo que pregunta: «¿se deduce que...?»³⁰. Como es evidente que la IA no va a pararse a esperarnos, podemos esperarla en su punto final, más allá de la escuela, más allá del patio de recreo, más allá de lo que puede ser o no un cementerio de instituciones. Por delante, por detrás o en algún movimiento conjunto aún por nombrar, nos la encontramos allí con la capacidad de realizar inferencias más allá de la implicación lingüística, sabiendo todo lo que sabemos sobre cómo hacer inferencias a partir del lenguaje natural registrado, y sabiendo todo lo que sabemos sobre lo que puede salir mal cuando tales inferencias no tienen libertad ni supervisión. Sabemos que las inferencias literarias y críticas de las personas no pueden ser idénticas, porque eso sería una especie de muerte, pero también sabemos que la interpretación no puede ser arbitraria o una aproximación estadística de segunda mano. Es posible que a la IA no le importe nada de esto, por supuesto, porque, al igual que la muerte, no puede sufrir las consecuencias posteriores de una comprensión falsa y de inferencias erróneas en su cuerpo o en su vida. Pero eso significaría que nosotros tenemos que preocuparnos por esas consecuencias –de la falsedad y las inferencias erróneas– en los cuerpos y las vidas de otras personas.

Doy las gracias a Zachary Tavlin, Rachel Feder y Rebecca Porte por sus lecturas de «Because I could not stop for Death», y a Tom Lippincott sus comentarios.

³⁰ Véase Johanna Winant, *Lyric Logic: How Modern American Poetry Reasons*, Nueva York, 2025, para ver un ejemplo de crítica literaria organizada en torno a este concepto.

Tarifas de suscripción a la revista *New Left Review* en español

Para España

Suscripción anual (6 números)

Suscripción anual individual [55 €]

Suscripción anual para Instituciones [200 €]

(una suscripción equivaldrá a 3 ejemplares de cada número enviados a una misma dirección postal)

Venta de un ejemplar individual para instituciones [20 €]

Gastos de envío postal ordinario incluidos.

Para Europa

Suscripción anual (6 números)

Suscripción anual individual [85 €]

Suscripción anual para Instituciones [300 €]

(una suscripción equivaldrá a 3 ejemplares de cada número enviados a una misma dirección postal)

Venta de un ejemplar individual para instituciones [30 €]

Gastos de envío postal ordinario incluidos.

Resto del mundo*

Suscripción anual (6 números)

Suscripción anual individual [120 €]

Suscripción anual para Instituciones [350 €]

(una suscripción equivaldrá a 3 ejemplares de cada número enviados a una misma dirección postal)

Venta de un ejemplar individual para instituciones [50 €]

Gastos de envío postal ordinario incluidos.

Formas de pago

Se puede realizar el pago mediante tarjeta de crédito, transferencia bancaria o domiciliación bancaria a través de nuestra página:

<http://traficantes.net/nlr/suscripcion>

Para cualquier duda podéis escribirnos a nlr_suscripciones@traficantes.net