

LA LIBERTAD EN LOS TIEMPOS DE LOS LLMS: VOLUNTAD, PROBABILIDAD Y AUTONOMÍA ANTE LA PALABRA PREDICHA

FREEDOM IN THE AGE OF LLMS: WILL, PROBABILITY, AND AUTONOMY BEFORE THE PREDICTED WORD

Darío Caleb Cario Alanís

Universidad Autónoma de Nuevo León (UANL) | México

Resumen

Este artículo examina la problemática de la libertad humana en el contexto de los modelos de lenguaje de gran escala (LLMs), tomando su capacidad para predecir y generar continuidades lingüísticas como punto de partida para una reflexión filosófica sobre la autonomía. El texto comienza exponiendo el funcionamiento probabilístico de estos modelos y distingue la generación computacional de la deliberación humana. Posteriormente, analiza la experiencia cotidiana de habitar entornos tecnológicos que anticipan, organizan y sugieren posibilidades de acción y expresión. A partir de Schopenhauer, el LLM es interpretado como un espejo técnico —no como un equivalente ontológico— que vuelve visible el carácter condicionado del querer humano, en cuanto nuestras acciones emergen de la relación entre motivos, carácter y representación. Kant permite, a continuación, pensar la autonomía no como ausencia de todo condicionamiento, sino como capacidad de examinar la máxima bajo la cual hacemos propia una decisión. Finalmente, se propone interrumpir la secuencia probable como una práctica crítica: abrir un intervalo en el que una continuación algorítmica pueda convertirse en objeto de examen antes de ser asumida como propia.

Palabras clave: modelos de lenguaje de gran escala; libertad; autonomía; Schopenhauer; Kant

Abstract

This article examines the problem of human freedom in the context of large language models (LLMs), taking their ability to predict and generate linguistic continuations as a starting point for a philosophical reflection on autonomy. It begins by explaining the probabilistic functioning of these models and distinguishing computational generation from human deliberation. It then analyzes the everyday experience of inhabiting technological environments that anticipate, organize, and suggest possibilities for action and expression. Drawing on Schopenhauer, the LLM is interpreted as a technical mirror—not as an ontological equivalent—that makes visible the conditioned character of human willing, insofar as our actions emerge from the relation between motives, character, and representation. Kant then makes it possible to understand autonomy not as the absence of all conditioning, but as the capacity to examine the maxim under which we make a decision our own. Finally, the article proposes interrupting the probable sequence as a critical practice: opening an interval in which an algorithmically suggested continuation can become an object of examination before being assumed as one's own.

Keywords: large language models; freedom; autonomy; Schopenhauer; Kant

Introducción

¿En qué medida somos libres cuando incluso nuestras palabras pueden ser anticipadas? La pregunta no presupone que la capacidad predictiva de una máquina demuestre la inexistencia de la libertad humana. Busca, más modestamente, examinar por qué esa capacidad vuelve problemática la manera en que comprendemos nuestra propia expresión.

Esta pregunta surge en un momento histórico en el que la inteligencia artificial, especialmente los modelos de lenguaje de gran escala o LLMs (*large language models*), se ha incorporado a prácticas cotidianas de búsqueda, escritura, trabajo y decisión. Ya no recurrimos a estos sistemas únicamente para localizar información; también les pedimos que organicen ideas, reformulen argumentos, redacten mensajes, resuman textos y sugieran maneras de expresar aquello que queremos decir. En esa medida, comienzan a intervenir en la forma en que elaboramos y presentamos nuestras propias ideas.

Un LLM puede completar una frase, sugerir una respuesta, continuar un estilo, redactar un mensaje o resumir una idea. Incluso puede reproducir ciertos rasgos reconocibles de la manera en que una persona suele expresarse. Estas capacidades resultan fascinantes para quienes investigamos la inteligencia artificial, pero también producen una inquietud filosófica. Con el desarrollo tecnológico de los LLMs, hemos llegado a un momento singular de la historia en que aquello que durante mucho tiempo asociamos con la intención, el pensamiento y la expresión de un sujeto puede ser simulado por un sistema que opera sin una experiencia humana del mundo.

Frente a ello, la pregunta por la libertad deja de ser solamente moral o metafísica. Se convierte también en una pregunta técnica, lingüística y existencial. Técnica, porque exige comprender cómo se producen estas continuidades; lingüística, porque afecta la relación entre palabra, intención y autoría; y existencial, porque nos obliga a preguntar cuánto de aquello que decimos, deseamos y decidimos procede de una deliberación propia y cuánto continúa estructuras que nos preceden.

En términos generales, un LLM es un modelo de lenguaje cuyo objetivo generativo básico consiste en estimar una distribución de probabilidad sobre los posibles *tokens* —unidades básicas de texto— que pueden continuar una secuencia a partir de un contexto previo. Sin embargo, esta predicción no describe por sí sola la complejidad interna del sistema: para producirla, el modelo transforma el contexto mediante representaciones numéricas de alta dimensionalidad y múltiples capas de procesamiento que integran información distribuida de la secuencia (Vaswani et al.; Brown et al.). La predicción del siguiente *token* constituye, por tanto, el punto de salida generativo, no una descripción exhaustiva de todo lo que ocurre internamente. En el sentido que interesa a este artículo, dicha producción no exige suponer por sí misma una conciencia, una voluntad o una experiencia subjetiva del mundo.

Sin embargo, lo inquietante no se encuentra únicamente en el modelo. Aparece cuando advertimos que también nuestras palabras, deseos y decisiones se forman dentro de condiciones que nos preceden: hábitos, historia personal, lenguaje, clase social, cultura, cuerpo y entornos algorítmicos. La comparación no permite reducir al ser humano a un modelo probabilístico, pero sí vuelve sospechosa la idea de que hablar, escribir o elegir libremente significa originar algo desde la nada.

La tesis que sostengo es que los LLMs no prueban el determinismo humano ni permiten reducir al sujeto a un sistema probabilístico; funcionan, más bien, como un espejo técnico que vuelve visible el carácter condicionado de nuestra expresión y obliga a replantear la autonomía, no como capacidad de originar palabras o decisiones desde la nada, sino como posibilidad crítica de reconocer, examinar e interrumpir las continuidades que se nos presentan como probables.

Para desarrollar esta tesis, el argumento se organizará en cuatro momentos. En primer lugar, presentaré el nivel técnico, con el propósito de explicar en qué sentido un modelo de lenguaje produce continuidades probabilísticas sin elegir sus palabras como lo haría un sujeto. En segundo lugar, abordaré el nivel fenomenológico, centrado en la experiencia cotidiana de habitar entornos que anticipan, sugieren y ordenan nuestras posibilidades. En tercer lugar, examinaré el problema metafísico desde Schopenhauer, cuya concepción de la voluntad permite pensar al LLM como un espejo técnico del carácter condicionado de nuestros deseos y acciones. Finalmente, recurriré a Kant para preguntar si todavía es posible reconstruir la autonomía como capacidad de examinar e interrumpir la secuencia probable.

La continuidad probable del lenguaje: funcionamiento técnico y problema filosófico

Los modelos de lenguaje generativos contemporáneos suelen construirse mediante variantes de la arquitectura Transformer, introducida originalmente como una arquitectura basada en mecanismos de atención. En los modelos autorregresivos, como los pertenecientes a la familia GPT —*Generative Pre-trained Transformer*—, la generación se organiza alrededor de una tarea básica: calcular, a partir del contexto disponible, una distribución de probabilidad sobre los *tokens* que podrían continuar la secuencia (Vaswani et al.; Brown et al.). Esta predicción se sostiene, sin embargo, en un procesamiento interno mucho más complejo que la simple asociación entre una palabra y la siguiente. Para producir cada distribución, el contexto atraviesa sucesivas transformaciones que construyen representaciones vectoriales de alta dimensionalidad y distribuyen información relevante entre múltiples capas del modelo.

La predicción del siguiente *token* debe entenderse, entonces, como el objetivo generativo inmediato del modelo, no como una descripción completa de su organización interna. El sistema aprende durante el entrenamiento parámetros que le permiten transformar el contexto recibido y representar relaciones entre los elementos de la secuencia antes de producir la distribución final de probabilidades. Esta complejidad interna será relevante más adelante al examinar cómo ciertos patrones funcionales del modelo pueden hacerse técnicamente interpretables. La continuidad probabilística que interesa a este artículo emerge, por tanto, de una arquitectura interna mucho más rica que una simple asociación entre una palabra y la siguiente.

A su vez, generar una distribución de probabilidades no significa que el modelo seleccione siempre el *token* con el valor numérico más alto. A partir de la distribución calculada, distintos procedimientos de decodificación pueden seleccionar una continuación u otra. Lo fundamental para este argumento es que cada salida se encuentra condicionada por el contexto precedente, los parámetros aprendidos durante el entrenamiento y el procedimiento utilizado para generar la respuesta. La palabra producida no aparece desde la nada: emerge de un espacio de posibilidades construido y calculado por el sistema.

En este sentido restringido, el LLM no elige sus palabras como lo hace un sujeto humano. No delibera entre alternativas porque considere una de ellas verdadera, justa o moralmente preferible, ni se da a sí mismo una regla conforme a la cual decide actuar. Sus continuaciones dependen de operaciones realizadas sobre representaciones numéricas y de regularidades adquiridas durante el entrenamiento. Decir que el modelo “no es libre” no constituye todavía una conclusión metafísica acerca de la inteligencia artificial; significa, por ahora, que no debemos confundir la selección computacional de una salida con la autonomía moral de un sujeto.

El interés filosófico comienza precisamente aquí. Un sistema al que no atribuimos experiencia vivida, responsabilidad moral o voluntad propia puede, no obstante, producir frases coherentes, continuar un estilo, desarrollar una argumentación y generar una apariencia de pensamiento. Esto no demuestra que el lenguaje humano y el lenguaje generado sean ontológicamente equivalentes. Sí muestra, en cambio, que la coherencia lingüística no basta por sí sola para probar la existencia de una conciencia libre detrás de las palabras. La relación entre lenguaje, subjetividad y libertad se vuelve, por ello, menos evidente de lo que parecía.

Algunos trabajos recientes de interpretabilidad mecanicista han identificado patrones internos de activación técnicamente interpretables, denominados *features*, cuyo comportamiento puede caracterizarse a partir de los contextos en los que se activan. En Claude 3 Sonnet, por ejemplo, se han aislado *features* que responden de manera relativamente consistente ante referencias a determinados lugares, personas, emociones, estructuras de código y otros contenidos semánticos reconocibles. Investigaciones posteriores sobre Claude 3.5 Haiku han reconstruido circuitos funcionales involucrados en tareas como el razonamiento intermedio, la planificación de versos y el procesamiento multilingüe (Templeton et al.; Lindsey et al.).

Estas investigaciones permiten hablar de representaciones internas en un sentido técnico, pero no autorizan por sí mismas a atribuir conciencia, experiencia subjetiva o conceptos en el mismo sentido en que los atribuimos a un sujeto humano. Una *feature* vinculada con el dolor, la justicia o una persona no siente, juzga ni recuerda aquello con lo que se relaciona. Más bien constituye un patrón funcional dentro del procesamiento del sistema. La cuestión filosófica no es, entonces, si esas activaciones son ideas idénticas a las humanas, sino cómo un sistema carente de experiencia subjetiva puede organizar representaciones capaces de intervenir eficazmente en el lenguaje que compartimos.

La experiencia de ser autocompletados

La predicción algorítmica no permanece encerrada en los procesos internos de un modelo. Se hace visible en una experiencia cada vez más cotidiana: la de habitar entornos técnicos que anticipan lo que podríamos querer decir, mirar, escuchar o responder. Pasar del funcionamiento del LLM a esta experiencia exige, sin embargo, una precisión. Una red social, un sistema de recomendación, un corrector automático y un asistente conversacional no son técnicamente equivalentes. Comparten, para el problema que aquí interesa, una forma de mediación: la intervención algorítmica no necesita sustituir directamente nuestra elección, pues puede operar configurando de antemano el campo de posibilidades dentro del cual elegimos.

Una plataforma clasifica y prioriza el contenido que aparecerá ante nosotros. En sistemas de recomendación a gran escala, esa operación puede comprender una etapa de generación de candidatos y otra de ordenamiento o *ranking*, como muestra el sistema de YouTube descrito por Covington et al. Aquello que finalmente observamos no representa la totalidad indiferenciada de lo disponible, sino el resultado de una selección previa realizada conforme a determinados criterios técnicos.

Una intervención semejante aparece en la escritura. Una aplicación sugiere cómo terminar una frase; un corrector modifica una palabra antes de que revisemos conscientemente su pertinencia; un asistente conversacional propone la estructura de un mensaje o la continuación de una idea. Esta asistencia no afecta únicamente la velocidad de escritura. En una tarea experimental de composición, Arnold et al. encontraron que las personas que recibían sugerencias predictivas escribían textos más breves y utilizaban formulaciones más previsibles que quienes escribían sin ellas.

Lo relevante no es que estos sistemas eliminen formalmente la posibilidad de escoger. El usuario todavía puede ignorar una recomendación, cerrar la aplicación, modificar una palabra o escribir algo distinto. La cuestión aparece en un nivel anterior: ciertas posibilidades llegan ya seleccionadas, ordenadas o destacadas antes de que el sujeto responda a ellas. Yeung ha descrito este tipo de intervención mediante la noción de *hypernudge*, mostrando cómo los sistemas basados en datos pueden orientar decisiones mediante la configuración dinámica del entorno de elección. De manera convergente, Weinmann et al. señalan que el diseño de interfaces digitales puede modificar la arquitectura de elección dentro de la cual se producen decisiones.

Esta mediación puede experimentarse, paradójicamente, no como una privación evidente de libertad, sino como una proliferación de opciones. Una plataforma puede ofrecer más contenido del que cualquier usuario sería capaz de revisar por cuenta propia y producir así la experiencia de un horizonte abierto. Sin embargo, la abundancia misma se encuentra previamente organizada: aquello entre lo que elegimos ya ha atravesado procedimientos de selección, jerarquización y recomendación.

Esto no convierte al usuario en una marioneta ni permite afirmar que toda decisión tecnológicamente mediada haya dejado de ser libre. Permite hablar, más cuidadosamente, de una libertad mediada. Desde la perspectiva de la mediación tecnológica, las tecnologías no funcionan únicamente como instrumentos pasivos de intenciones previamente constituidas, sino que participan en la configuración de las percepciones y acciones desde las que decidimos (Verbeek).

Esta mediación adquiere una relevancia particular cuando interviene en el lenguaje mediante el cual formulamos nuestras propias razones. Un asistente puede ayudarnos a encontrar una expresión, organizar un argumento o enfrentar la hoja en blanco; pero también puede introducir una formulación que termine incorporándose a aquello que presentamos como propio. Jakesch et al. mostraron experimentalmente que un asistente lingüístico configurado para favorecer una postura podía influir tanto en las opiniones expresadas por los participantes como en sus respuestas posteriores sobre el mismo tema. Este resultado no demuestra que el sistema determine sus convicciones, pero sí permite advertir que la asistencia lingüística puede intervenir en la formulación y evaluación de aquello que el sujeto llega a expresar.

La experiencia de ser autocompletados plantea, entonces, un problema más preciso que el de una supuesta sustitución total de nuestra voluntad por la máquina. Nos sitúa ante una libertad ejercida entre posibilidades cuya aparición y relevancia pueden haber sido técnicamente anticipadas. La cuestión ya no consiste solamente en comprobar si todavía podemos aceptar o rechazar una sugerencia, sino en preguntarnos qué ocurre con nuestra autonomía cuando parte del horizonte desde el que elegimos ha sido previamente organizado.

Pero esta pregunta todavía permanece en la superficie de la elección. Saber que una posibilidad ha sido seleccionada o sugerida no explica por qué puede resultarnos adecuada, atractiva o difícil de rechazar. Para ello es necesario avanzar hacia una cuestión anterior: ¿de dónde procede aquello del sujeto que quiere aceptar o rechazar? En este punto, el problema de la palabra predicha se encuentra con la filosofía de Schopenhauer.

El LLM como espejo técnico del pesimismo schopenhaueriano

Esta pregunta desplaza el problema desde las condiciones externas de la elección hacia la formación misma del querer. Hasta aquí hemos examinado cómo los sistemas técnicos pueden anticipar, seleccionar y organizar posibilidades; con Schopenhauer, la cuestión se dirige ahora hacia el sujeto que responde a ellas. Antes de preguntarnos hasta qué punto una máquina interviene en nuestra libertad, debemos examinar hasta qué punto el querer desde el cual elegimos puede considerarse verdaderamente nuestro.

En Schopenhauer, el problema de la libertad alcanza precisamente la formación misma del querer. El mundo fenoménico se encuentra sometido al principio de razón suficiente: todo aquello que aparece lo hace dentro de relaciones que permiten explicar por qué ocurre de una manera y no de otra. En el caso de la acción humana, este principio adopta la forma de la ley de motivación. Un motivo no es solamente una causa exterior que empuja mecánicamente al sujeto, sino una representación conocida por él y capaz de suscitar un acto de la voluntad. Por ello, Schopenhauer afirma que “la motivación es la causalidad vista por dentro” y denomina a esta relación “principio de razón suficiente del obrar” (Schopenhauer, *Cuádruple raíz* 206–209).

El motivo, sin embargo, no determina por sí solo una acción idéntica en todos los individuos. Actúa sobre un carácter particular y dentro de una determinada forma de conocer la situación. De ahí que personas diferentes puedan responder de manera distinta ante circunstancias aparentemente semejantes. Para Schopenhauer, el acto concreto nace de la convergencia entre el motivo y el carácter: el primero proporciona la ocasión de la acción, mientras que el segundo determina la forma específica en que el individuo responde. El sujeto no comienza cada decisión desde un punto vacío o indiferente, sino desde una individualidad que ya se expresa en sus elecciones (Schopenhauer, *Cuádruple raíz* 208–209; *El mundo* 517–520).

Esta relación exige distinguir entre la voluntad como cosa en sí y sus manifestaciones fenoménicas. Schopenhauer considera libre a la voluntad en cuanto no está sometida, en sí misma, a las formas del principio de razón. Pero esa libertad no se transfiere directamente a la persona empírica ni a cada una de sus decisiones particulares. La libertad metafísica que Schopenhauer atribuye a la voluntad pertenece a esta en cuanto cosa en sí y no designa una facultad electiva oculta del individuo que, desde un plano nouménico, escogiera libremente aquello que después manifiesta en el mundo fenoménico. El individuo, su carácter y sus actos pertenecen al mundo fenoménico y aparecen, por ello, dentro de una estricta concatenación de motivos, causas y consecuencias. No hay *liberum arbitrium indifferentiae* porque la persona no es libre en el sentido de poder producir, ante las mismas condiciones, cualquier decisión imaginable; sus actos manifiestan temporalmente un carácter que ella no ha escogido previamente (Schopenhauer, *El mundo* 515–520).

Lo que experimentamos como una apertura entre varias alternativas puede ser, entonces, una indeterminación del conocimiento y no necesariamente del querer. Mientras deliberamos, el intelecto contempla posibilidades contrapuestas porque todavía desconoce cuál de ellas expresará finalmente el carácter. Ambas parecen igualmente posibles desde nuestra perspectiva, pero la resolución resulta de la confrontación entre los motivos presentes y la individualidad sobre la que actúan. Schopenhauer denomina a esta experiencia la “ilusión de la libertad empírica de la voluntad” (Schopenhauer, *El mundo* 521–523). La conciencia observa la decisión cuando esta ocurre, pero no ocupa un punto exterior desde el cual pudiera producir libremente un carácter y un querer distintos.

Esto no significa que la deliberación sea inexistente ni que el conocimiento resulte inútil. El intelecto puede comparar motivos, corregir una representación equivocada, anticipar consecuencias y mostrar caminos que antes permanecían ocultos. El conocimiento puede modificar considerablemente el modo de actuar porque transforma la manera en que una situación se presenta ante el sujeto. Puede mostrarle que se equivocó en los medios, que ignoraba ciertas consecuencias o que un objeto distinto permite perseguir aquello que ya quería. No transforma necesariamente el núcleo de la voluntad, pero sí

interviene en el campo representacional dentro del cual los motivos adquieren sentido y fuerza (Schopenhauer, *El mundo* 527–531).

Este punto resulta decisivo para pensar los LLMs. Un modelo de lenguaje no necesita implantar directamente un deseo para intervenir en una decisión. Puede reformular una situación, presentar una posibilidad antes ausente, ordenar razones, anticipar consecuencias o volver lingüísticamente más convincente uno de los caminos disponibles. Su poder no radica necesariamente en sustituir o anular la voluntad del usuario, sino en participar de manera significativa en la producción y organización de las representaciones mediante las cuales los motivos de la voluntad llegan a ser conocidos. Si, como señala Schopenhauer, los motivos solo actúan en cuanto son representados (Schopenhauer, *Cuádruple raíz* 206–209), entonces la mediación lingüística de un LLM no es filosóficamente indiferente.

Aun así, sería un error concluir que un modelo probabilístico y un sujeto humano funcionan de manera idéntica. Una distribución de *tokens* no equivale a un motivo, del mismo modo que los parámetros de una red neuronal no equivalen al carácter de una persona. El sujeto humano posee cuerpo, afectividad, historia, autoconciencia y relaciones sociales; además, puede reflexionar sobre las razones que orientan su acción. La comparación no pretende reducir al humano a una máquina estadística ni utilizar la arquitectura del LLM como demostración del determinismo. Su función es más modesta, pero quizá también más inquietante: mostrar que una continuación coherente puede emerger de condiciones anteriores sin originarse en una libertad absoluta.

En este sentido, el LLM funciona como un espejo técnico del pesimismo schopenhaueriano. El modelo continúa una secuencia conforme al contexto y a las regularidades aprendidas; el sujeto continúa su existencia desde una trama de carácter, motivos, hábitos, cuerpo e historia. Por ello, el espejo no consiste en afirmar que ambos operen de la misma manera, sino en hacer visible, mediante la continuidad técnica del primero, un problema que Schopenhauer había situado en el querer humano: hasta qué punto aquello que aparece como decisión propia emerge de condiciones que el sujeto no ha elegido. La semejanza no es ontológica, sino interrogativa. La máquina no demuestra que el ser humano carezca de libertad, pero nos obliga a sospechar de aquello que llamamos elección y a preguntar cuánto de nuestro lenguaje nace realmente de una decisión autónoma y cuánto continúa necesidades que, por nuestras propias condiciones, no alcanzamos a reconocer ni calcular.

Schopenhauer lleva esta sospecha todavía más lejos. La capacidad deliberativa humana no nos sustrae del conflicto de los motivos: nos convierte en el lugar donde esos motivos pueden aparecer de forma abstracta, enfrentarse y ser ponderados. La razón amplía el horizonte de la elección, pero no crea desde la nada aquello que queremos. Incluso cuando elegimos después de una reflexión extensa, la decisión continúa expresando el encuentro entre los motivos conocidos y el carácter individual. Por eso, la complejidad de una elección no garantiza por sí misma su libertad (Schopenhauer, *El mundo* 531–536).

Sin embargo, la filosofía schopenhaueriana tampoco conduce necesariamente a una pasividad absoluta. El conocimiento de la propia individualidad puede dar lugar a lo que denomina carácter adquirido: una comprensión progresiva de aquello que queremos, de lo que podemos hacer y de los límites dentro de los cuales actuamos. Este autoconocimiento no permite fabricarnos desde cero, pero sí orientar de manera más reflexiva la conducta, evitar contradicciones y reconocer las fuerzas que intervienen en nuestras decisiones (Schopenhauer, *El mundo* 539–545).

La negación de la voluntad presentada hacia el final de *El mundo como voluntad y representación* tampoco debe confundirse con una simple decisión cotidiana de rechazar una sugerencia. En el §70, Schopenhauer describe una transformación mucho más radical, en la que el conocimiento deja de funcionar como motivo ordinario y llega a actuar como aquietador del querer. No se trata de escoger otra posibilidad dentro del mismo campo de deseos y, por ello, no puede identificarse con cerrar una aplicación, rechazar una recomendación o interrumpir un autocompletado (Schopenhauer, *El mundo* 685–694).

Sin embargo, aún queda abierto un problema particularmente difícil. Si no elegimos desde la nada nuestro carácter, nuestros deseos ni las condiciones bajo las cuales los motivos adquieren fuerza, ¿en qué sentido todavía podemos hablar de autonomía? Schopenhauer permite diagnosticar el carácter condicionado del querer, pero ese diagnóstico no agota la pregunta por la forma en que un sujeto responde a aquello que lo motiva. El conocimiento puede modificar la presentación de los motivos, ampliar aquello que el sujeto considera posible y transformar el modo en que actúa frente a ellos, aunque no convierta por ello a la voluntad empírica en absolutamente libre.

Es en este punto donde el retorno a Kant adquiere su función dentro del argumento. La cuestión deja de ser únicamente de dónde procede aquello que queremos y se desplaza hacia una pregunta normativa:

bajo qué principio hacemos nuestra una decisión y hasta qué punto podemos responder por la máxima conforme a la cual actuamos. Kant no será utilizado aquí para restaurar una libertad empírica incondicionada después de Schopenhauer, sino para examinar si todavía puede hablarse de autonomía dentro de una existencia atravesada por inclinaciones, motivos y mediaciones técnicas.

Autonomía e interrupción de la secuencia probable

Si Schopenhauer desplaza el problema de la libertad hacia el origen del querer, Kant permite formular una pregunta distinta. Ya no se trata solamente de saber de dónde proceden nuestros deseos, motivos e inclinaciones, sino de examinar bajo qué principio actuamos cuando decidimos seguirlos. Esta diferencia no elimina la herida abierta por Schopenhauer: seguimos siendo seres condicionados, afectados por el cuerpo, la historia, los hábitos y, ahora, por sistemas capaces de anticipar y organizar nuestras posibilidades. Kant permite, sin embargo, preguntar si ese condicionamiento agota el problema de la libertad o si todavía es posible adoptar una posición racional frente a aquello que nos inclina.

Esta distinción resulta especialmente clara en *La metafísica de las costumbres*. Kant diferencia entre voluntad y arbitrio. El arbitrio se relaciona con la elección de las acciones, mientras que la voluntad se refiere al fundamento que determina esa elección y, en cuanto razón práctica, cumple una función legisladora. El arbitrio humano no es puro ni permanece aislado de lo sensible, pues es afectado por impulsos e inclinaciones. Pero, según Kant, no queda por ello necesariamente determinado por estos, pues también puede ser orientado por la razón. La libertad negativa consiste en la independencia respecto de la determinación exclusiva de los impulsos; su sentido positivo aparece cuando la razón puede ser práctica por sí misma y someter las máximas de la acción a la posibilidad de convertirse en ley universal (Kant, MdS, AA VI 213–214, 226).

Esta distinción permite precisar el problema de los LLMs. Una sugerencia algorítmica puede afectar nuestro arbitrio: aparece ante nosotros como una respuesta posible, una frase conveniente, una formulación atractiva o un camino que ahorra tiempo. Puede incluso adquirir fuerza por su coherencia, por la autoridad que atribuimos al sistema o por la facilidad con la que resuelve una dificultad inmediata. Pero la mera presencia de esa sugerencia no demuestra todavía que el sujeto haya perdido toda capacidad de actuar. Al menos desde la perspectiva práctica kantiana, existe una diferencia que no deberíamos borrar demasiado pronto entre ser afectado por una continuación y quedar determinado por ella.

Esto tampoco significa que toda acción realizada por una persona sea autónoma simplemente porque fue ella quien pulsó “aceptar”, copió el texto o envió la respuesta. El origen externo de una sugerencia no vuelve por sí mismo heterónoma la acción que la incorpora, del mismo modo que el hecho de aceptarla no basta para convertirla en autónoma. Una recomendación procedente de un modelo puede ser examinada y apropiada reflexivamente; del mismo modo, un deseo aparentemente “interior” puede gobernarnos de manera heterónoma cuando lo obedecemos solo porque resulta agradable, cómodo o ventajoso. La diferencia no se encuentra únicamente en el lugar del que procede la influencia, sino en el principio bajo el cual la incorporamos a nuestra acción.

Aquí aparece la importancia kantiana de la máxima. La máxima es el principio subjetivo que el agente adopta como regla de su obrar: no describe únicamente qué hizo, sino bajo qué regla decidió hacerlo. Kant formula el principio supremo de la doctrina de las costumbres como la exigencia de actuar según una máxima que pueda valer al mismo tiempo como ley universal. Las leyes proceden de la voluntad legisladora; las máximas, del arbitrio que elige cómo actuar (Kant, MdS, AA VI 226). Esto transforma la pregunta por el uso de un LLM. Ya no basta con preguntar si acepté o rechacé su respuesta. Debo preguntarme qué máxima sigo cuando la acepto: ¿utilizo cualquier salida que me permita terminar más rápido?, ¿delegaré mi juicio siempre que la respuesta parezca convincente?, ¿emplearé el modelo como punto de partida, pero revisaré aquello que finalmente asuma como propio?, ¿puedo responder por la regla implícita en mi decisión?

Desde esta perspectiva, la autonomía no consiste en producir una respuesta sin antecedentes. Tampoco exige escribir desde un espacio vacío, libre de lenguaje, cultura, deseos o mediaciones técnicas. Consiste en que el objeto deseado, la inclinación o la ventaja esperada no se conviertan por sí solos en la ley de nuestra conducta. En la *Crítica de la razón práctica*, Kant identifica la libertad positiva con la legislación propia de la razón y la autonomía de la razón práctica. La heteronomía aparece, en cambio, cuando el fundamento de la regla depende del objeto del deseo o del seguimiento de un impulso e inclinación (Kant, KpV, AA V 33).

Un LLM puede entrar precisamente en ese lugar. Puede presentarnos un objeto lingüístico ya formado: una frase, una conclusión, una interpretación o incluso una aparente razón para actuar. El riesgo no reside únicamente en que la respuesta sea falsa. También puede ser verdadera, útil y lingüísticamente impecable, pero ser incorporada sin que el sujeto examine la regla bajo la cual la hace suya. La heteronomía contemporánea no tendría que adoptar siempre la forma visible de una orden. Puede aparecer como una continuación tan cómoda, probable o adecuada que ya no sintamos la necesidad de preguntarnos por qué habría que seguirla.

Es aquí donde adquiere sentido mi propuesta de *interrumpir la secuencia probable*. La interrupción no debe entenderse como una prueba empírica de libertad. Detenerse, cerrar una aplicación o rechazar una sugerencia también puede proceder de un impulso, un prejuicio, un miedo o una obediencia distinta. Interrumpir no vuelve autónoma por sí misma una decisión. Su importancia consiste en abrir un intervalo: el momento en el que la continuación deja de operar como una consecuencia aparentemente natural y puede convertirse en objeto de examen.

Ese intervalo puede permitirnos preguntar qué estamos haciendo con la salida generada, qué razones aceptaríamos como propias y qué consecuencias estamos dispuestos a asumir. La autonomía no reside en escoger automáticamente la opción contraria a la sugerida, sino en no permitir que la probabilidad de una continuación sustituya el juicio sobre su validez. Aceptar una respuesta puede ser una acción reflexiva; rechazarla puede ser una reacción ciega. Lo decisivo es si el sujeto puede reconocer la máxima de su acción, examinarla y responder por ella.

En rigor kantiano, aquello que recibe una sugerencia, se encuentra afectado por ella y puede incorporarla a una acción es el arbitrio, no la voluntad entendida como razón práctica legisladora. Por ello, empleo aquí de manera deliberadamente metafórica la expresión “suspender el autocompletado de la voluntad”. No pretendo convertirla en una categoría kantiana, sino nombrar con ella el momento en que una continuación técnicamente sugerida corre el riesgo de confundirse con aquello que el sujeto reconoce como principio propio de su acción. Suspender ese autocompletado no supone imaginar una voluntad previamente vacía o aislada de toda influencia. Significa abrir la posibilidad de que la sugerencia sea examinada antes de ser asumida como propia.

Esta suspensión exige ciertas prácticas. La lectura, la escritura y el pensamiento crítico no son adornos culturales externos al uso de la IA; constituyen capacidades mediante las cuales el sujeto puede comparar, corregir y reformular aquello que recibe. Incluso la generación de un *prompt* puede entenderse como una forma limitada de intervención: en lugar de esperar pasivamente una continuación, el usuario delimita una tarea, formula condiciones, solicita explicaciones y revisa la salida. Pero redactar un *prompt* no garantiza autonomía. También puede convertirse en una forma más sofisticada de obedecer al sistema si el usuario carece de los recursos para distinguir entre una respuesta convincente y una respuesta que puede realmente defender.

Por ello, el problema no se resuelve mediante la oposición simple entre usar o no usar un LLM. Una herramienta puede convertirse en una extensión de nuestras capacidades, del mismo modo que otros dispositivos técnicos amplían la vista, la memoria o la acción. Puede ayudarnos a superar la hoja en blanco, ensayar formulaciones y reconocer posibilidades que antes no habíamos considerado. Pero esa extensión no debe confundirse con una sustitución completa del juicio. El desafío consiste en utilizar la herramienta de tal forma que no prescindamos de la regla mediante la cual evaluamos lo que ella produce.

El riesgo aparece cuando esa delegación alcanza también la evaluación de las razones. Un usuario puede recurrir al modelo para producir una formulación, organizar un argumento o explorar una posibilidad y conservar todavía la tarea de juzgar aquello que recibe. La situación cambia cuando el sistema comienza a sustituir también ese segundo momento y el sujeto acepta sus resultados sin poder reconstruir las razones por las que deberían considerarse válidos, adecuados o propios. En ese caso, lo que se debilita no es solamente la elaboración personal de una respuesta, sino la capacidad de justificar la decisión que finalmente se adopta.

La doctrina kantiana de la virtud añade aquí una última precisión. En seres como nosotros, cuya acción no coincide de manera inmediata con lo que la razón prescribe, la autonomía requiere también una forma de autoacción: la capacidad de gobernar inclinaciones y sostener máximas frente a aquello que resulta inmediatamente atractivo. Kant distingue esta autoacción libre de la coacción externa y la vincula con la posibilidad de dominar las propias inclinaciones (Kant, MdS, AA VI 394–395). Trasladada con cautela al problema contemporáneo, esta idea permite afirmar que la autonomía ante los LLMs no

consiste únicamente en comprender técnicamente cómo funcionan, sino en cultivar la capacidad práctica de no aceptar toda continuación por el solo hecho de que sea fácil, rápida o probable.

Kant no deshace el condicionamiento mostrado por Schopenhauer ni demuestra que poseamos una libertad empírica absoluta. Tampoco podemos conocer teóricamente la libertad como un objeto dado en la experiencia. En la filosofía práctica kantiana, sin embargo, esta no queda reducida a una posibilidad meramente pensable. La conciencia de la ley moral conduce al concepto de libertad y le confiere un estatuto práctico que no podría obtenerse a partir de la experiencia: Kant presenta la libertad como *ratio essendi* de la ley moral y a la ley moral como *ratio cognoscendi* de la libertad (Kant, KpV, AA V 4, 29–30). Su aportación consiste, por ello, en impedir que la descripción causal o la explicación psicológica de nuestras inclinaciones agote por completo el problema normativo de la acción. Podemos estar atravesados por motivos que no elegimos, expuestos a sugerencias que no solicitamos y situados dentro de arquitecturas que ordenan nuestras posibilidades. Pero, aun así, permanece la pregunta por la máxima bajo la cual hacemos nuestra una continuación.

Quizá sea allí donde todavía pueda defenderse una forma limitada, frágil y exigente de libertad: en la posibilidad de someter a examen aquello que pretende hablar por nosotros y preguntarnos si estamos dispuestos a hacerlo propio. Schopenhauer nos impide considerarnos autores absolutos de nuestros deseos; Kant desplaza la cuestión hacia la regla bajo la cual actuamos frente a ellos. Entre la sugerencia y su aceptación permanece, entonces, el problema de la máxima conforme a la cual el arbitrio incorpora una posibilidad a la acción. La autonomía no elimina las condiciones que nos preceden, pero exige que aquello que finalmente asumimos pueda ser examinado y reconocido como una decisión por la que estamos dispuestos a responder.

Conclusión: detenerse ante la palabra predicha

Hemos partido de una pregunta: ¿en qué medida somos libres cuando incluso nuestras palabras pueden ser anticipadas? La respuesta no puede reducirse a afirmar que los LLMs destruyen la libertad humana ni a sostener que su funcionamiento probabilístico demuestra que nunca hemos sido libres. La inquietud es más modesta, pero quizá también más profunda: estos sistemas vuelven visible que una expresión coherente puede surgir como continuación de condiciones anteriores y, con ello, hacen más difícil sostener una imagen de la libertad como origen absolutamente espontáneo de nuestras palabras y decisiones.

La contribución que he intentado defender consiste en pensar al LLM como un espejo técnico. El espejo no establece una equivalencia ontológica entre la máquina y el sujeto: una distribución de *tokens* no es un motivo, los parámetros de un modelo no constituyen un carácter y la generación lingüística de un sistema no reproduce la experiencia humana de querer. Su función consiste en hacer visible, mediante una forma técnica de continuidad, un problema que también atraviesa nuestra propia expresión. En ese sentido, la analogía no demuestra que el ser humano sea una máquina probabilística, pero vuelve sospechosa la idea de que cada palabra y cada elección surjan de un punto absolutamente incondicionado.

Schopenhauer permite profundizar esa sospecha al mostrar que el querer empírico se encuentra atravesado por carácter, motivos y representaciones que el individuo no ha creado desde la nada. Kant permite formular, sin cancelar ese condicionamiento, una pregunta distinta: bajo qué máxima hacemos nuestra una posibilidad y hasta qué punto podemos responder por aquello que finalmente incorporamos a nuestra acción. Entre ambos aparece una concepción menos absoluta de la autonomía: no como ausencia de antecedentes, influencias o mediaciones, sino como exigencia de examinar la regla bajo la cual algo llega a convertirse en una decisión propia.

Desde esta tensión adquiere sentido la propuesta de interrumpir la secuencia probable. Su valor no reside en demostrar una libertad situada fuera de toda causalidad ni en recomendar que rechazemos aquello que una máquina sugiera. La libertad no aparece en la interrupción como un objeto ya conquistado. La interrupción conserva, más modestamente, la posibilidad de someter una continuidad a juicio y preguntar por qué habríamos de continuarla antes de asumirla como propia.

De esta interpretación se desprende también una consecuencia práctica. Si los LLMs participan cada vez más en la escritura, el aprendizaje y la elaboración de razones, una formación crítica frente a estas tecnologías no puede reducirse a enseñar su utilización eficiente o la producción de mejores *prompts*. La lectura, la escritura, la contrastación y la reconstrucción de razones adquieren importancia precisamente porque permiten juzgar aquello que el sistema produce. Saber pedirle algo al modelo no equivale *eo ipso*

a saber juzgar aquello que produce. El desafío educativo consiste, por ello, en formar usuarios capaces no solo de obtener respuestas, sino de decidir cuáles pueden comprender, justificar y asumir como propias.

Quizá los LLMs no nos quiten simplemente la libertad. Quizá hacen algo más incómodo: revelan que nunca fuimos tan libres como creíamos y que muchas veces hemos vivido continuando palabras, deseos y estructuras que nos preceden. Pero esta revelación no obliga a renunciar por completo a la autonomía. Si todavía podemos defender alguna frágil forma de libertad, no estará en producir una palabra absolutamente nueva desde la nada, sino en poder detenernos ante la palabra predicha, reconocer aquello que intenta completarnos y decidir, hasta donde nuestras capacidades y condiciones lo permitan, qué estamos dispuestos a hacer con ella.

La máquina, en ese sentido, no solo nos reemplaza ni simplemente nos domina: también nos revela. Nos muestra una imagen técnica de la continuidad y nos obliga a volver la mirada hacia nuestro propio querer. Ante esa imagen, la libertad deja de ser una posesión evidente y se convierte en una práctica difícil: la de no aceptar como propio todo aquello que puede completarnos.

Obras citadas

- Arnold, Kenneth C., et al. "Predictive Text Encourages Predictable Writing." *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, 2020, pp. 128–138.<https://doi.org/10.1145/3377325.3377523>
- Brown, Tom B., et al. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.<https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Covington, Paul, et al. "Deep Neural Networks for YouTube Recommendations." *Proceedings of the 10th ACM Conference on Recommender Systems*, Association for Computing Machinery, 2016, pp. 191–198.<https://doi.org/10.1145/2959100.2959190>
- Jakesch, Maurice, et al. "Co-Writing with Opinionated Language Models Affects Users' Views." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2023, article 111, pp. 1–22.<https://doi.org/10.1145/3544548.3581196>
- Kant, Immanuel. *Crítica de la razón práctica*. Traducido por Dulce María Granja Castro, Fondo de Cultura Económica / Universidad Autónoma Metropolitana / Universidad Nacional Autónoma de México, 2011.
- Kant, Immanuel. *La metafísica de las costumbres*. Traducido por Adela Cortina Orts y Jesús Conill Sancho, Tecnos, 2005.
- Lindsey, Jack, et al. "On the Biology of a Large Language Model." *Transformer Circuits Thread*, Anthropic, 2025.<https://transformer-circuits.pub/2025/attribution-graphs/biology.html> Consultado el 5 jul. 2026.
- Schopenhauer, Arthur. *De la cuádruple raíz del principio de razón suficiente*. Traducido por Leopoldo-Eulogio Palacios, Gredos, 1998.
- Schopenhauer, Arthur. *El mundo como voluntad y representación*, 1. Traducido por Roberto R. Aramayo, Alianza Editorial, 2025.
- Templeton, Adly, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Transformer Circuits Thread*, Anthropic, 2024. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>. Consultado el 8 jul. 2026.
- Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.<https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need>
- Verbeek, Peter-Paul. "Materializing Morality: Design Ethics and Technological Mediation." *Science, Technology, & Human Values*, vol. 31, no. 3, 2006, pp. 361–380.<https://doi.org/10.1177/0162243905285847>
- Weinmann, Markus, et al. "Digital Nudging." *Business & Information Systems Engineering*, vol. 58, no. 6, 2016, pp. 433–436.<https://doi.org/10.1007/s12599-016-0453-1>
- Yeung, Karen. "Hypernudge: Big Data as a Mode of Regulation by Design." *Information, Communication & Society*, vol. 20, no. 1, 2017, pp. 118–136.<https://doi.org/10.1080/1369118X.2016.1186713>