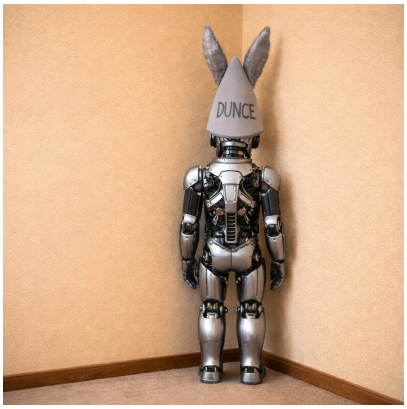

La inteligencia artificial no piensa y nosotros estamos dejando de hacerlo

Descripción

Introducción

¿Son perros o toallas?

Circula desde hace tiempo una imagen aparentemente inofensiva la de varios perros arrugados y unas toallas enrolladas. A un humano le bastan décimas de segundo para distinguirlos. Bueno a la mayoría de los humanos. Sin embargo, muchos sistemas de visión artificial son incapaces. Confunden perros con toallas y fallan.

¿Son perros o toallas?

Circula desde hace tiempo una imagen aparentemente inofensiva la de varios perros arrugados y unas toallas enrolladas. A un humano le bastan décimas de segundo para distinguirlos. Bueno a la mayoría de los humanos. Sin embargo, muchos sistemas de visión artificial son incapaces. Confunden perros con toallas y fallan.

Hay gente que se ríe al leerlo La gracia suele quedarse ahí. *Mira qué tonta es la máquina*. Pero el chiste tiene trampa, porque ese mismo tipo de sistemas, los que confunden textura con significado, son los que hoy usamos para clasificar imágenes médicas, detectar fraudes, priorizar currículums o decidir qué contenido merece ser visto y cuál no.

El fallo no está en que la IA se equivoque, el fallo está en no entender por qué se equivoca y, aun así, confiarle decisiones cada vez más relevantes.

La máquina no ve un perro, ve patrones de píxeles que se parecen estadísticamente a otros patrones etiquetados como “perro”. Cuando esos patrones se parecen mucho a una toalla arrugada, el sistema duda. O falla. Y eso no es un bug aislado, es una pista de cómo “conoce” el mundo.

Este artículo se centra en una idea incómoda pero necesaria. La inteligencia artificial no entiende la realidad como la entendemos los humanos, y sin embargo estamos empezando a delegarle criterio, decisiones y autoridad. Este artículo no es para hablar de apocalipsis tecnológicos, sino para analizar, desde la ciencia, la experiencia cotidiana y el sentido común, qué puede hacer la IA, qué no puede hacer y qué riesgos asumimos cuando olvidamos la diferencia.

El cerebro humano y el mundo

El cerebro no procesa datos, tropieza con el mundo, y aprende porque vive, no porque optimiza

Un ser humano no aprende porque “procesa información”, aprende porque está vivo.

Aprende porque se cae, porque se equivoca, porque hace el ridículo, porque prueba algo, no funciona y lo ajusta sobre la marcha. Aprende porque el mundo no es amable, no está etiquetado y no espera a que termines de entrenarte.

Piensa en un día cualquiera. No en uno heroico, en uno vulgar.

Te levantas medio dormido, tanteas la mesilla sin mirar para encontrar el móvil, calculas sin pensar la distancia exacta para no tirarlo al suelo. Vas a la cocina, esquivas una silla que no recuerdas haber movido y buscas y seleccionas el desayuno. Hoy te vas a dar un capricho y decides hacerte unos huevos revueltos. Tuenes que elegir el producto y acertar a dar el golpe exacto para no destruir la cáscara, mientras que has de hacer el doble de esfuerzo para abrir la botella de zumo. Te quemas un poco con la sartén y bajas el fuego. Mientras tanto, estás pensando en una conversación pendiente, en algo que dijiste ayer y quizá no deberías haber dicho así. Suena el teléfono. Reconoces por el tono de voz si tu amigo está bien o no antes de que termine la primera frase. Decides qué decir, qué callar, cuánto insistir. Todo eso ocurre sin detener el mundo, sin pausar la realidad para “cambiar de tarea”.

Ese es el tipo de inteligencia que damos por sentada. Y es brutal.

El cerebro humano coordina visión, oído, tacto, memoria, emoción, normas sociales, cálculo motor y expectativas futuras al mismo tiempo, en tiempo real, en un entorno lleno de ruido, contradicciones y estímulos inesperados. No porque sea perfecto, sino porque está diseñado para adaptarse. Y encima modifica la salida en base a la intención, incluso en condiciones iguales, con respuestas diferentes adaptadas al contexto.

No hay dataset que capture eso. No hay función de pérdida que lo optimice.

En ciencia cognitiva se habla de cognición encarnada por una razón muy simple, porque pensar no ocurre en el vacío, ocurre en un cuerpo que se mueve, que se cansa, que siente placer, miedo, vergüenza o deseo. Ocurre en un entorno social donde las decisiones tienen consecuencias que no son numéricas, como perder la confianza de alguien, quedar como un imbécil, hacer daño sin querer o incluso salir dañados.

Ahora comparemos esto con una IA.

Una IA no se levanta con sueño. No se quema. No duda si decir algo por miedo a herir. No siente el peso de una decisión. No paga el precio de equivocarse. No tropieza con el mundo. Lo observa desde fuera, convertido en datos.

Por eso, cuando ve una imagen, no ve “algo”, ve correlaciones estadísticas entre píxeles. Y cuando esas correlaciones cambian, se pierde. Hay estudios claros que muestran que muchos sistemas de visión artificial clasifican objetos basándose más en texturas locales que en formas globales, justo al revés que los humanos. De ahí que una toalla arrugada pueda parecerle un perro.

El humano no confunde porque ha vivido con perros, ha oído ladridos, ha olido pelo mojado, ha esquivado mordiscos, ha sentido miedo o cariño. La máquina no confunde porque sea tonta, confunde porque nunca ha vivido nada.

Y aquí está el núcleo de esta reflexión. El conocimiento humano no se construye solo resolviendo tareas, sino siendo muchas cosas a la vez. Trabajador, amigo, torpe, valiente, cobarde, curioso. Aprendemos porque el mundo nos obliga a adaptarnos constantemente, no porque optimicemos bien una métrica.

Esto no es romanticismo, es una diferencia de arquitectura.

Lenguaje fluido no es pensamiento

Uno de los grandes trucos de magia de nuestra época es el lenguaje. Nos deslumbra, nos tranquiliza, nos engaña con facilidad. Y los modelos de lenguaje lo explotan sin mala intención, pero con una eficacia casi obscena.

Un LLM, como ChatGpt, escribe bien. A veces muy bien, demasiado bien. Con frases redondas, conectores elegantes, tono seguro. Y eso activa en el lector un sesgo muy humano. Si suena coherente, debe saber de lo que habla. Es el mismo sesgo que nos hace confiar en alguien que habla con aplomo, aunque esté diciendo una tontería monumental. Sólo hay que ver la cantidad de millones de libros de autoayuda vacía que se venden en el mundo y la ingente cantidad de bytes virales que corren por internet de cómo lograr el éxito sin dar palo al agua.

La diferencia es que cuando una persona habla con seguridad, hay algo detrás. Hay experiencia, arrogancia, intuición, mala fe o ignorancia. Pero hay alguien. En un modelo de lenguaje no hay nadie. Hay un sistema optimizado para predecir la siguiente palabra más probable dado un contexto. No piensa la mejor respuesta sino la más probable en base a los millones de datos con los que ha sido entrenado.

Eso no es una metáfora. Es literal.

Un LLM no “sabe” que París es la capital de Francia. Sabe que, estadísticamente, después de “La capital de Francia es…” suele venir “París”. Y mientras el mundo siga siendo razonablemente estable, esa predicción funciona. Siempre, claro, que alguien no haya bombardeado el sistema de entrenamiento con millones de veces “Washington”, en cuyo caso el sistema le dará la razón a Trump y Francia será virtualmente invadida por los EEUU.

https://sergiocolado.com/wp-content/uploads/2026/02/IA-LLM.mp4

Fuente: Santiago Ortiz.

En un caso normal sin contaminación ni engaños, el problema aparece cuando el mundo se complica, cuando la pregunta exige comprensión, contexto, límites o, simplemente, decir “no lo sé”.

Ahí aparece el fenómeno que ya conocemos bien, el de la alucinación. El modelo no distingue entre rellenar un hueco con algo plausible y afirmar algo verdadero. Porque su objetivo no es la verdad, es la continuidad del discurso. Y eso está documentado, clasificado y estudiado hasta el aburrimiento en investigación reciente.

El resultado es inquietante. Surgen respuestas convincentes y bien escritas, pero erróneas y entregadas con la misma serenidad que una respuesta correcta. El modelo no miente, no engaña, simplemente no sabe que está equivocado.

Y aquí es donde entra la dependencia humana que tanto debería preocuparnos. Estos sistemas necesitan supervisión constante porque no tienen criterio interno para distinguir lo aceptable de lo inaceptable. Por eso se entrenan con feedback humano, con evaluaciones humanas, con juicios humanos. Porque sin ese trabajo invisible, el modelo sería lingüísticamente brillante y socialmente inútil, cuando no peligroso.

Hay algo todavía más sutil. El lenguaje humano no es solo transmisión de información, es intención, ironía, silencio, amenaza, cuidado, complicidad. Cuando alguien te dice “haz lo que quieras”, sabes perfectamente si eso es libertad o una bomba de relojería. Un modelo no lo sabe. Puede imitar el tono, pero no vivir la situación. El componente no verbal tampoco está implícito en la conversación con un LLM y se pierden muchísimos canales de comunicación necesarios para interpretar y entender realmente una conversación.

Por eso el lenguaje fluido no es pensamiento, es, como mucho, simulación de superficie. Útil, potentísima, sí, pero sin anclaje en la experiencia, en el cuerpo y en las consecuencias, sigue siendo eso, superficie.

La fantasía de la autosuficiencia

Hay una idea que flota en el ambiente tecnológico como una promesa mesiánica, que la IA se auto-mejorará, que aprenderá sola, que llegará un punto en el que los humanos sobraremos.

Esta es una idea que vende muy bien en conferencias, pero que aguanta mal el contacto con la realidad.

La IA actual no se mejora sola, la mejoran.

La mejoran ingenieros, etiquetadores, evaluadores, usuarios que corrigen, que reportan errores, que afinan prompts, que detectan sesgos. La mejoran personas que aportan algo que la máquina no tiene, el contacto con el mundo real.

Cuando se intenta prescindir de eso y se entrenan modelos con datos generados por otros modelos, empieza el problema. No de golpe, no de forma espectacular, sino como se degradan las cosas importantes, poco a poco. Se pierde diversidad, se pierden casos

raros, se pierde aquello que no es frecuente pero es crucial. Aparece lo que la investigación ha llamado *model collapse*, que es que el sistema se vuelve cada vez más parecido a sí mismo y menos representativo del mundo.

Explicado claramente, la IA empieza a olvidar el mundo para recordarse a sí misma.

Esto conecta directamente con la predicción sesgada. Los modelos no solo aprenden lo que hay en los datos, aprenden lo que falta. Si ciertos perfiles, contextos o comportamientos están infrarrepresentados, el modelo no “se da cuenta”., simplemente los borra de facto. Y como no tiene experiencia directa que le contradiga, ese sesgo se solidifica.

Un humano puede corregir una creencia porque vive una excepción. La IA no vive excepciones. Las ve como ruido.

Además, incluso si resolviéramos todos esos problemas técnicos, es decir mejores datos, mejores modelos, mejores evaluaciones, seguiría faltando algo que no se puede automatizar, la responsabilidad. Cuando una decisión causa daño, alguien tiene que responder. Y la IA no responde. No puede. No entiende el daño, no lo sufre y no lo repara.

Por eso la idea de una IA plenamente autónoma no es solo técnicamente dudosa, es conceptualmente vacía.

Autonomía sin responsabilidad no es autonomía, es irresponsabilidad delegada.

Shitty helperbot (fuente: Imgur)

Lo que la máquina no tiene y no sabe que le falta

Hay un obstáculo del que se habla poco cuando se fantasea con una inteligencia artificial general o con una supuesta “súper IA”. No es la falta de datos. No es la potencia de cálculo. Ni siquiera es la arquitectura. Es algo mucho más complejo porque no se arregla con más ingeniería. La máquina no tiene intención.

No intención en el sentido trivial de “querer hacer algo”, sino en el sentido profundo, en el de que no hay nada que le importe. No hay nada en juego para ella. No hay un *para qué* que nazca desde dentro.

Un ser humano no actúa solo porque pueda, actúa porque algo le importa, porque tiene sed, quiere agradar, teme perder, ama, odia, cree, se equivoca y luego corrige. Nuestras decisiones no están guiadas únicamente por cálculo, sino por una orientación constante hacia el mundo, por una direccionalidad que no se puede desconectar.

Cuando buscas la botella de agua no estás “optimizando una función”. Estás respondiendo a una necesidad corporal. Cuando atiendes una llamada no estás maximizando eficiencia comunicativa, estás leyendo un contexto emocional, una historia compartida, una relación. Cuando decides no decir algo, no es porque no sea estadísticamente adecuado, sino porque anticipas consecuencias humanas.

Eso es intencionalidad y sin ella no hay sentido común.

La IA, en cambio, no quiere nada, no teme nada, no espera nada. Tiene objetivos, sí, pero objetivos prestados, definidos por otros, optimizados desde fuera. Puede parecer que decide, pero solo se mueve dentro de un marco que alguien delimitó previamente. No puede cuestionarlo, ni puede decir “esto no tiene sentido”, ni puede detenerse porque algo le resulta moralmente inaceptable.

Incluso cuando hablamos de agentes autónomos que planifican, se corrigen, usan herramientas y “reflexionan” sobre sus pasos, seguimos hablando de agencia simulada. La máquina no persigue fines propios, sino que ejecuta fines ajenos con mayor o menor sofisticación.

Un agente de IA puede decir “mi objetivo es maximizar X, así que haré Y”. Eso parece agencia, pero no lo es. Es como un termostato muy sofisticado que regula, ajusta y corrige, pero no le importa la temperatura. No tiene frío, ni calor. No quiere estar mejor.

Y aquí se rompe la fantasía. Sin intención no hay comprensión real, solo cálculo.

El sentido común humano no es una base de datos de reglas, es el resultado de vivir orientado en un mundo donde las decisiones tienen coste, donde equivocarse duele y donde acertar importa. La IA puede aprender regularidades, pero no puede preocuparse por ellas. Y sin preocupación no hay criterio.

Mientras no haya intencionalidad genuina y no programada, no optimizada, no derivada, no habrá motivación propia, preocupación y comprensión con sentido. Y sin eso, hablar de IA general es hablar de una generalización sin centro.

Por eso la pregunta no es si la IA será más rápida, más precisa o más eficiente. Ya lo es en muchas cosas. La pregunta es si puede llegar a tener algo que nunca ha tenido, un mundo propio. Y hoy, no solo no lo tiene, sino que no sabemos cómo dárselo sin abandonar el terreno de la ciencia para entrar en el de la ficción.

Opinión y conclusiones

La inteligencia artificial no piensa y repetirlo no es una provocación, es una vacuna contra la estupidez.

No piensa porque no vive. No vive porque no arriesga. No arriesga porque no tiene nada que perder. Y quien no tiene nada que perder no puede tener criterio.

La máquina no tiene cuerpo, no tiene mundo y no tiene algo en juego. No siente el peso de una decisión mal tomada, no paga el precio de una palabra mal dicha, no aprende porque se haya equivocado delante de otros. Ejecuta, optimiza, ajusta, pero no se orienta en la realidad como lo hace un ser humano que vive entre otros seres humanos.

El problema, por tanto, no es que la IA calcule mejor, escriba más rápido o vea patrones donde nosotros no los vemos. El problema es que estamos empezando a tratar esos cálculos como si fueran juicio, esa escritura como si fuera comprensión y esos patrones como si fueran verdad. Y no lo son. Son aproximaciones útiles, a veces brillantes, pero vacías de intención.

Nos da miedo que la IA nos quite el trabajo, pero ese no es el miedo correcto. El miedo real es que nos quite algo peor, la costumbre de decidir, de dudar, de equivocarnos con responsabilidad, que dejemos de pensar porque alguien o algo nos da una respuesta inmediata, ordenada y con tono seguro.

No viene una rebelión de máquinas. Eso es cine barato. Viene algo más sutil, más humano y más peligroso, personas que dejan de pensar porque la herramienta siempre responde. Y cuando el mundo se sale del guion, porque siempre se sale, ya no saben qué hacer sin una sugerencia automática, sin un ranking, sin una probabilidad que les diga qué opción elegir.

La IA no es el problema. El problema es la fe ciega. La delegación perezosa. La renuncia voluntaria a aquello que nos hace humanos, el decidir en condiciones imperfectas, con información incompleta, asumiendo consecuencias.

Mientras recordemos que la máquina calcula y el humano juzga, iremos bien.
El día que confundamos ambas cosas, no hará falta que la IA nos sustituya, ya nos habremos sustituido solos.

Referencias

- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47* (1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

- Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science*. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477
- Colado García, S. (2020). Influencia de la tecnología en el desarrollo del pensamiento y conducta humana. Autoedición Amazon
- Colado García, S. (2021). *Multiversos digitales – La tecnología como palanca evolutiva* . Universo de Letras. ISBN 978-84-188-5466-8
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Bygh9j09KX
- Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*. https://arxiv.org/abs/1801.00631
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … Christiano, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*. https://doi.org/10.48550/arXiv.2203.02155
- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*, 755–759. https://doi.org/10.1038/s41586-024-07566-y
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.