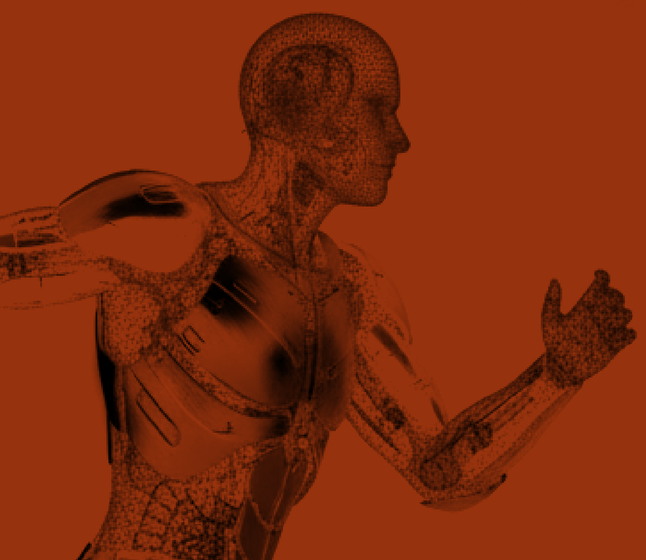
Pinocchio GePPetto



Le bugie, ragazzo mio, si riconoscono subito, perchè ve ne sono di due specie: vi sono le bugie che hanno le gambe corte, e le bugie che hanno il naso lungo: la tua per l'appunto è di quelle che hanno il naso lungo.
C. Collodi[1]

ULISES CORTÉS

La aspiración prometeica de los humanos por infundir vida a sustancias inanimadas refleja nuestro anhelo no satisfecho de emular a los dioses capaces de insuflarla en barro o madera o masa de maíz. La literatura ilustra estas ambiciones humanas desde el Golem del rabino Judah Loew, pasando por la *criatura* de Mary Shelley, hasta los robots de Karel Čapek[2].

La creación de *símiles* artificiales de humanos se nos presenta en una variedad limitada de formatos que imitan la biología humana, como los posibles rellenos de un taco que necesitan el soporte de una tortilla. Hay que decir que Norbert Wiener ya reflexionó sobre el tema, en 1964[3], en su famoso God and Golem, Inc., donde afirmaba: *Man makes man in his own image.*

Este panorama está cambiando, ya se proponen nuevas *creaciones*. Este fenómeno se está acelerando gracias a los avances en la biología, la inteligencia artificial y la robótica y esto está trayendo nuevas versiones de *posibles* seres artificiales con capacidades cognitivas.

Son estas variantes de esta última versión de la historia de los *símiles* artificiales de humanos las que han captado el interés de investigadores y filósofos preocupados por la bioética, la inteligencia artificial y otros asuntos morales relacionados con la innovación científica y tecnológica. Por eso, la cita de Collodi que encabeza este texto me hizo pensar en la estrecha relación que existe entre la historia de Pinocchio y la de los modelos masivos del lenguaje[4] (LLM), y cómo se conectan con esta aspiración demiúrgica de la creación de *símiles* artificiales de la inteligencia humana y los efectos éticos, legales, socioeconómicos y culturales que tiene abrir esta caja de Pandora o más bien abrir esta Matrioshka tecnológica que son los LLMs y sus secuelas como los llamados modelos fundacionales[5].

Pero comencemos por el principio, es decir, por la creación misma de la humanidad. La mitología nos explica que los dioses de varias cosmogonías en distintos intentos de crear humanos usando la madera siempre fallaron. Este aspecto experimental de la creación divina me parece fascinante: hay dioses que se equivocan, reflexionan y mejoran o cambian su método creativo. En la cosmogonía maya narrada en el *Popol Vuh*[6], por ejemplo, los dioses intentaron formar humanos en varias ocasiones. Primero lo modelaron con barro, pero las figuras no tenían fuerza ni consistencia: *pronto se deshacían y se volvían polvo*. Luego lo intentaron con

La mitología nos explica que los dioses de varias cosmogonías en distintos intentos de crear humanos usando la madera siempre fallaron. Este aspecto experimental de la creación divina me parece fascinante: hay dioses que se equivocan, reflexionan y mejoran o cambian su método creativo.

madera, pero esos humanos carecían de alma y de entendimiento, no recordaban a sus creadores, *y por eso fueron destruidos y anegados por una inundación*. Como última solución, los dioses Tepeu y Gucumatz utilizaron masa de maíz, materia sagrada que dio origen a los humanos verdaderos, aquellos con *espíritu* y *palabra*. Por eso resulta curioso que el maestro Geppetto, en el cuento moderno, del siglo XIX, logre lo que aquellos dioses mayas no pudieron: transformar una marioneta de madera en una marioneta viviente. Pinocchio nace en falso, como la mayor parte de las creaciones mitológicas, pues nunca es un bebé de carne y hueso sino de madera, esa es su trágica falsedad primigenia.

Evitando la versión edulcorada de los estudios Disney de 1940, la obra original de Collodi ofrece lecciones éticas atemporales extraídas de las penurias y el crecimiento y transformación de Pinocchio. Estas enseñanzas resaltan la virtud personal, las consecuencias de las acciones y el desarrollo humano mediante el aprendizaje y el esfuerzo constante. La nariz de Pinocchio que crece simboliza cómo la *mentira* erosiona la confianza. La marioneta se miente a sí misma y miente a los demás, tiene intención de engañar y el engaño crece con su napia.

Pinocchio es un títere de madera al que se le ha dotado de una consciencia externa: un grillo capaz de hablar y dar buenos consejos que al principio son desoídos y luego se van, afianzado en la personalidad del muñeco que va, poco a poco, con esfuerzo, humanizándose. En este proceso de metamorfosis no hay magia alguna, solo aprendizaje, esfuerzo y la superación de la mitomanía. La maestría del narrador y la complicidad del lector hacen esta ficción plausible.

En el caso de los modelos masivos, la situación es distinta, no tienen un *cuerpo* humano pero sus creadores, unos demiurgos tecnológicos, aspiran a que sus *criaturas*, en el sentido de M. Shelley[7], superen las condiciones propuestas por Alan Turing[8], en 1950, para hacer *indistinguible* la conversación que sostiene un programa de la que puede tener un humano con otro. En mi opinión, Turing concibió su juego como una auténtica batalla de ingenio.

Al final, su prueba no mide cuán *inteligente* es un sistema basado en la inteligencia artificial, sino hasta qué punto logra parecer humano. Aunque parezca una paradoja, los humanos no nos definimos tanto por nuestra racionalidad como por nuestros errores, emociones y *excentricidades* al comunicarnos. Por eso, un sistema basado en la inteligencia artificial, lógico y eficiente, tendría dificultades para superar el Test de Turing[9]. Como señaló Marvin Minsky, *tendemos a pensar en la inteligencia como algo que una persona posee, pero en realidad surge de la cooperación de muchas partes no inteligentes*[10].

Es decir, la inteligencia en una máquina no consiste en simular la perfección, sino en que sea capaz de enfrentarse a la imperfección, parecer humano implica también aprender a fallar como tal o a hacer trampas, si fuese *necesario*. Esto no justifica en absoluto las mal llamadas *alucinaciones*[11] que producen los modelos masivos del lenguaje, que no son otra cosa que fallos, muchas veces errores de bulto o generaciones de contenido ficticio, cuyo origen es casi siempre inexplicable, aunque, con la tecnología existente, imposible de corregir. También hay que decir que en estos fallos no hay, *a priori*, intención de engaño, ya que estas máquinas no tienen intenciones, aunque ello no exima de culpa a los programadores que las han creado.

La distinción entre *fallo humano* y *alucinación* técnica es crucial para avanzar hacia sistemas basados en la IA más robustos sin que esto implique mayor inteligencia en estos sistemas; los LLMs pueden llegar a ser más eficientes, pero no más inteligentes.

En este intento de conseguir una máquina que tenga una inteligencia comparable o superior a la de un humano, un primer paso esencial es que este sistema debería abandonar su calidad de máquina. Tendría que, si eso fuese posible, dejar de ser *artificial*. De lo contrario permanecerán como artefactos muy potentes, pero limitados, incapaces de capturar la intrincada simplicidad esencial de la mente. Y es que, al menos por ahora, sin un toque de *magia*, ese umbral sigue fuera del alcance de la tecnología. Pero, como diría Geppetto: *Caro mio, non si sa mai quel che ci può capitare in questo mondo. I casi son tanti!...*

ULISES CORTÉS
Catedrático de Inteligencia Artificial de la Universitat Politècnica de Catalunya. Coordinador Científico del grupo High-Performance Artificial Intelligence del Barcelona Supercomputing Center. Miembro del Observatori d'Ètica en Intel·ligència Artificial de Catalunya y del Comitè d'Ètica de la Universitat Politècnica de Catalunya. Es miembro del comité ejecutivo de EurAI. Participante como experto de México en el grupo de trabajo Data Governance de la Alianza Global para la Inteligencia Artificial (GPAI). Doctor Honoris Causa por la Universitat de Girona.

REFERENCIAS

•1 C. Collodi. Le avventure di Pinocchio: Storia di un burattino. https://www.gutenberg.org/cache/epub/52484/pg52484.txt
•2 Čapek, Karel (2001). R.U.R.. Translated by Paul Selver and Nigel Playfair. Dover Publications.
•3 N. Wiener. God and Golem, Inc. MIT Press.
•4 A. Radford, K. Narasimhan., T. Salimans & I. Sutskever. Improving Language Understanding by Generative Pre-Training. (2018) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
•5 R. Bommasani et al. On the Opportunities and Risks of Foundation Models. ArXivabs/2108.07258 (2021). https://arxiv.org/abs/2108.07258
•6 Popol Vuh. https://www.iifl.unam.mx/uploads/popolVuh/libros/popolVuhXocNah.pdf
•7 Shelley, M. Frankenstein. Annotated for Scientists, Engineers, and Creators of All Kinds. (2017) MIT Press.
•8 A. M. Turing (1950) Computing Machinery and Intelligence. Mind 49: 433-460. https://courses.cs.umbc.edu/471/papers/turing.pdf
•9 Turing, A. M. (2021). Computing machinery and intelligence (1950). Mind, 59(236), 33-60. https://courses.cs.umbc.edu/471/papers/turing.pdf
•10 Minsky, M. (1986). The Society of Mind. New York: Simon & Schuster.
•11 Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023).Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730