Revista Colombiana de Pensamiento Estético e Historia del Arte

Edición 18
Julio-diciembre 2023
E-ISSN 2389-9794

Facultad de Ciencias Humanas y Económicas
Sede Medellín

Universidad Nacional de Colombia

ARTÍCULO

Dossier "Procesos creativos y cognitivos en la digitalización cultural"

De fábulas y autómatas: reflexiones iniciales de la narrativa para la alineación de valores de la inteligencia artificial

Óscar-Darío Villota-Cuásquer

Revista Colombiana de Pensamiento Estético e Historia del Arte

Edición 18 (Julio - diciembre de 2023)
E-ISSN 2389-9794

De fábulas y autómatas: reflexiones iniciales de la narrativa para la alineación de valores de la inteligencia artificial*

DOI: https://doi.org/10.15446/rcpeha.n18.106078

Óscar-Darío Villota-Cuásquer**

Resumen: el artículo aborda la necesidad de alinear la inteligencia artificial (IA) con los valores humanos ante el desarrollo de la IA general, capaz de igualar o superar la inteligencia humana. Se plantea que la narrativa puede ser fundamental para infundir principios éticos en las IA, dado su rol en la interpretación del mundo y la representación de experiencias humanas. Mediante una revisión bibliográfica de fuentes académicas, el estudio explora la relación entre la IA y la narrativa, enfocándose en la identidad narrativa, el conocimiento tácito y la experiencialidad narrativa. Se analizan casos como AlphaGo y ChatGPT para ilustrar la diferenciación entre IA estrecha y general, y se examina cómo las filosofías utilitarista y deontológica pueden influir en la ética de la IA. Los hallazgos identifican tres perspectivas fundamentales: la identidad narrativa, que configura la identidad a través de historias; el conocimiento tácito, que resalta el aprendizaje implícito y la experiencia intuitiva; y la experiencialidad narrativa, que subraya la importancia de la corporeidad y la vivencia temporal en la comprensión de relatos. La narrativa emerge como un enfoque viable para la alineación ética de la IA, aunque enfrenta desafíos relacionados con la experiencia vital y la corporeidad, elementos cruciales para una auténtica comprensión narrativa.

Palabras clave: inteligencia artificial; inteligencia artificial general; inteligencia narrativa; narrativa; alineación de valores; ética; AlphaGo; ChatGPT.

* **Recibido:** 28 de noviembre de 2022 / **Aprobado:** 22 de junio de 2023 / **Modificado:** 21 de julio de 2023. Artículo de investigación derivado de un trabajo doctoral más amplio cuyo título tentativo es "Diseño orientado a la experiencia y mediación técnica. Fundamentos para el diseño de obras inmersivas e interactivas". El artículo no contó con financiación institucional.

** Candidato a doctor en Diseño y Creación por la Universidad de Caldas (Manizales, Caldas). Magíster en Diseño y Creación Interactiva y profesor del Departamento de Diseño en la misma institución https://orcid.org/0000-0002-7604-5577 oscar.villota@ucaldas.edu.co

Cómo citar / How to Cite Item: Óscar-Darío. "De fábulas y autómatas: reflexiones iniciales de la narrativa para la alineación de valores de la inteligencia artificial". *Revista Colombiana de Pensamiento Estético e Historia del Arte*, no. 18 (2023): 53-76. https://doi.org/10.15446/rcpeha.n18.106078

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Of Fables and Automata: Initial Reflections on the Narrative for Value Alignment of Artificial Intelligence

Abstract: the article addresses the need to align artificial intelligence (AI) with human values in the face of the development of general AI capable of matching or surpassing human intelligence. It argues that narrative may be central to instilling ethical principles in AI, given its role in interpreting the world and representing human experiences. Through a literature review of academic sources, the study explores the relationship between AI and narrative, focusing on narrative identity, tacit knowledge, and narrative experientiality. Cases such as AlphaGo and ChatGPT are analyzed to illustrate the differentiation between narrow and general AI, and it examines how utilitarian and deontological philosophies may influence AI ethics. The findings identify three fundamental perspectives: narrative identity, which shapes identity through stories; tacit knowledge, which highlights implicit learning and intuitive experience; and narrative experientiality, which stresses the importance of embodiment and temporal liveness in understanding narratives. Narrative emerges as a viable approach to the ethical alignment of AI, although it faces challenges related to life experience and embodiment, crucial elements for authentic narrative understanding.

Keywords: artificial intelligence; artificial general intelligence; narrative intelligence; narrative; value alignment; ethics; AlphaGo; ChatGPT.

De fábulas e autômatos: reflexões iniciais sobre a narrativa para o alinhamento de valores da Inteligência Artificial

Resumo: o artigo aborda a necessidade de alinhar a inteligência artificial (IA) aos valores humanos diante do desenvolvimento de uma IA geral capaz de igualar ou superar a inteligência humana. Argumenta-se que a narrativa pode ser fundamental para incutir princípios éticos na IA, dada sua função de interpretar o mundo e representar as experiências humanas. Por meio de uma revisão da literatura de fontes acadêmicas, o estudo explora a relação entre IA e narrativa, concentrando-se na identidade narrativa, no conhecimento tácito e na experiência narrativa. Casos como o AlphaGo e o ChatGPT são analisados para ilustrar a diferenciação entre IA restrita e geral, e o estudo examina como as filosofias utilitarista e deontológica podem influenciar a ética da IA. As descobertas identificam três perspectivas principais: identidade narrativa, que molda a identidade

por meio de histórias; conhecimento tácito, que destaca o aprendizado implícito e a experiência intuitiva; e experiencialidade narrativa, que enfatiza a importância da incorporação e da experiência temporal na compreensão das narrativas. A narrativa surge como uma abordagem viável para o alinhamento ético da IA, embora enfrente desafios relacionados à experiência de vida e à incorporação, elementos cruciais para a compreensão autêntica da narrativa.

Palavras-chave: inteligência artificial; inteligência artificial geral; inteligência narrativa; narrativa; alinhamento de valores; ética; AlphaGo; ChatGPT.

Introducción

En la actualidad, las denominadas inteligencias artificiales débiles o estrechas —es decir, aquellas que se dedican a una tarea específica— se hallan en nuestra cotidianidad. Su implementación incluye asistentes virtuales, *software* de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento de voz y rostro, drones, vehículos autónomos y avances en el Internet de las cosas. No obstante, esta tecnología aún está lejos de lo que se espera conseguir con una inteligencia artificial general o fuerte, la cual, se plantea, podría equiparar o superar a la inteligencia humana. Con tal capacidad de resolución de problemas, se espera que esta tecnología implique un avance decisivo en muchos aspectos de la vida humana, y al mismo tiempo, vuelve imperativo que esta incidencia sea una deseable y alineada con los principios y deseos humanos; en otras palabras, que podamos confiar en ella. El presente texto versa sobre esta necesidad, y sobre cómo la narrativa, esa manera de dar sentido al mundo y representar experiencias, puede ser de ayuda para el desarrollo de inteligencias artificiales con un sentido ético. Finalmente se pregunta: ¿qué aspectos de la inteligencia narrativa serán cruciales al involucrar la narrativa en el desarrollo de la alineación de valores en la inteligencia artificial?

Para ello, el artículo empieza ilustrando la diferenciación entre la inteligencia artificial dedicada, la inteligencia artificial general y aspectos de la inteligencia humana, a través de una reflexión sobre el enfrentamiento del juego de mesa chino Go entre AlphaGo —un programa de IA desarrollado por DeepMind— y el jugador coreano Lee Sedol en 2016. Adicionalmente, se tienen en cuenta adelantos recientes en el campo con la denominada inteligencia artificial generativa, en particular, con ChatGPT, un modelo procesador de lenguaje natural creado por OpenAI en 2022. Posteriormente, se introduce el problema de la alineación

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

ética de la inteligencia artificial, analizando cómo diferentes corrientes filosóficas, como el utilitarismo y la ética deontológica, pueden influir en la dirección que toma este desarrollo tecnológico. La ética utilitarista, que busca maximizar el bienestar general, podría abogar por una IA que tome decisiones basadas en la maximización de resultados positivos para la mayor cantidad de personas, mientras que una perspectiva deontológica se centraría más en reglas y deberes establecidos que la IA debería seguir, independientemente de los resultados.

Por último, y como opción para establecer explícitamente principios éticos, se introduce la posibilidad de que agentes de IA infieran estos principios del análisis de situaciones y datos, desde donde se desprende un componente adicional y crucial: la narrativa. Los seres humanos no solo razonan y toman decisiones basados en principios éticos estrictos, sino que también interpretan y dan sentido a su realidad a través de historias. Esta inteligencia narrativa, intrínseca en la naturaleza humana, influye en valoraciones y decisiones éticas, y es esencial en la construcción de la identidad y entendimiento del mundo. Para concluir, el artículo profundiza en los aspectos cognitivos y experienciales que son inherentes a la inteligencia narrativa en los seres humanos, con el objetivo de estudiar su potencial impacto en el ámbito de la inteligencia artificial. Este análisis se abordará desde tres perspectivas fundamentales: la identidad narrativa, el conocimiento tácito y la experiencialidad narrativa.

Metodología

La metodología de este artículo se basa principalmente en la revisión bibliográfica a través de herramientas de búsqueda especializadas. Se consultaron una variedad de fuentes, incluyendo libros académicos, artículos de revistas revisadas por pares, informes técnicos y documentos de conferencias para obtener una comprensión amplia de la relación entre la IA y la narrativa, particularmente, en el contexto de la alineación de valores. Este artículo es parte de una investigación más amplia que se ha enfocado en los cruces entre la narrativa, los medios y la tecnología, con un interés particular en la experiencia como vínculo. Este marco conceptual proporcionó la base para la exploración de las relaciones entre la IA y la narrativa, de donde surge la temática explorada en este artículo.

Para estructurar la revisión bibliográfica, se utilizó un enfoque temático. Se estudiaron los conceptos fundamentales de la IA, tanto débil como general, y su relación con la ética. En ese punto, se profundizó en el problema de la alineación.

Lo anterior considerando las posibles conexiones entre la IA y la narrativa, con un énfasis particular en la identidad narrativa, el conocimiento tácito y la experiencialidad narrativa. Cabe señalar que la naturaleza exploratoria de esta metodología permite la inclusión de perspectivas diversas y, a veces, discordantes. Estas discrepancias sirven para ilustrar la complejidad y la riqueza de este campo de estudio emergente. Las conclusiones presentadas en este artículo son interpretaciones del autor basadas en la revisión bibliográfica realizada.

Inteligencia artificial e inteligencia (artificial) general

El 9 de marzo de 2016, en Corea del Sur, tuvo lugar un evento significativo tanto para la historia del Go, un milenario juego de mesa chino, como para el campo de la IA. Durante la competencia de cinco partidas conocida como The DeepMind Challenge Match, se enfrentaron AlphaGo — un programa de IA desarrollado por DeepMind Technologies— y Lee Sedol, el campeón mundial indiscutido del Go con 18 títulos a su nombre (retirado en la actualidad). El duelo prometía determinar si un humano o una máquina estaba más capacitado para uno de los juegos más complejos del mundo, cuyo número de posibles jugadas supera la cantidad de átomos en el universo[1]. Justamente, la complejidad inherente de este juego lo había convertido en un desafío para los desarrolladores de IA.

Uno de los aspectos más interesantes que emergen de los eventos de la partida relatados en el documental AlphaGo de 2017 es la distinción que emerge entre una inteligencia artificial dedicada o estrecha (ANI, por sus siglas en inglés), y lo que implicaría una inteligencia artificial general (AGI, por sus siglas en inglés). Los sistemas de IA dedicada, como AlphaGo, son extremadamente competentes en tareas específicas, pero no pueden aplicar sus habilidades fuera de estos dominios estrechos. Por otro lado, la IA general hace referencia a sistemas que serían capaces de realizar cualquier tarea que un humano pueda hacer, abordando simultáneamente problemas de naturaleza diversa en contextos distintos y aprendiendo nuevas habilidades. En otras palabras, la AGI tiene la capacidad de aplicar un conjunto central de recursos cognitivos a una amplia gama de tareas diferentes, al igual que los humanos[2].

1. Patrick Kiernan, "Which is greater? The number of atoms in the universe or the number of chess moves?", *Nationals Museums Liverpool* (página web), s.f. www.liverpoolmuseums.org.uk/stories/which-greater-number-of-atoms-universe-or-number-of-chess-moves

2. Henry Shevlin, *et al.*, "The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge", *EMBO Reports* 20 (2019): e49177, https://doi.org/10.15252/embr.201949177

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

La habilidad de Sedol para contemplar y evaluar varios aspectos de la situación corresponde justamente a las prácticas de una inteligencia general. Su mente no estaba concentrada en aspectos específicos del juego únicamente, lo que se evidenció en el estrés psicológico que experimentó durante la partida. Entre otras cosas, Sedol fue sometido a la presión de ser visto como un representante de la humanidad frente al avance de la IA, casi como si su victoria pudiera disipar, en alguna medida, el temor que esta tecnología despierta en términos de la obsolescencia del trabajo humano. Además, Sedol pudo sopesar otros factores, desde la inconveniencia de no poder "leer" a un oponente humano frente a él, hasta la preocupación por el impacto que este juego podría tener en su carrera.

Al igual que en la segunda partida de Garri Kasparov contra Deep Blue, otro duelo considerado un hito en la llamada segunda ola de la IA (y cuyos eventos también se narran en el documental *Game Over: Kasparov and the Machine*[3]), los aspectos psicológicos representan un enorme peso sobre la humanidad de un jugador profesional frente a una IA. A la importancia y atención que revisten este tipo de encuentros, se suma lo atípico de tener a un programa informático como rival. Lo anterior plantea un escenario donde los jugadores se cuestionan sobre sus capacidades, las de su oponente y sobre la naturaleza del juego al cual le han dedicado su vida. La capacidad de reflexión humana incluye la consciencia de la identidad e historia propias. Esto se hace evidente con una inesperada jugada hecha por AlphaGo, donde dejó atrás sus datos de estrategia humana, para ejecutar una movida completamente atípica, la célebre jugada 37 del segundo juego. Ante tal despliegue de capacidad, el campeón coreano declaró:

> I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative. This move was really creative and beautiful (...) This move made me think about Go in a new light. What does creativity mean in Go?[4]

La reflexión de Sedol, dotada de profundidad, claridad e incluso de sentido estético, es la manifestación de una inteligencia que está ocupada de múltiples elementos al mismo tiempo, y que puede atender a nuevas circunstancias, como evidentemente lo era esta partida. Dominar estas capacidades, inherentes a la inteligencia humana, es aún una meta para el campo de la inteligencia artificial. Proyectar las implicaciones que puede tener tanto para él como para la cultura del juego, requiere

3. Vikram Jayanti, dir., *Game Over: Kasparov and the Machine*, 2003.

4. Greg Kohs, dir., *AlphaGo*, 2017, min 52:14, YouTube, 13 de diciembre de 2017. https://www.youtube.com/watch?v=WXuK6gekU1Y

sopesar distintos escenarios, lo cual es crucial para construir una posición ética, como argumenta Mark Coeckelbergh[5]. Según este filósofo belga, la imaginación es fundamental para proyectar escenarios futuros, como las consecuencias de nuestras acciones, la forma de vida colectiva o tecnologías específicas. La imaginación nos permite realizar operaciones éticas básicas, como ponerse en el lugar de otros, proyectar ideales de vida personales y comunes, o formar una autoimagen moral y consciente. En la partida, tanto Sedol como los ingenieros de DeepMind y los espectadores hicieron uso de su inteligencia para explorar estos aspectos.

Entre tanto, AlphaGo estaba completamente dedicada a un solo propósito: jugar y ganar la partida de Go. Sus acciones tuvieron que ver con la enorme cantidad de probabilidades que se desprenden de las jugadas que, por posibilidad estadística de éxito, contempla llevar a cabo. Esta tarea, que de por sí es tan compleja que sus creadores en algunos casos no entendían sus decisiones, es la única cosa que le ocupaba —y la única de la que puede ocuparse—. AlphaGo no tuvo conciencia de que jugó frente al mejor jugador del mundo, tampoco sintió emoción por haber ganado la serie de manera apabullante 4 partidas a 1, ni tampoco reclamó el orgullo de haberse probado más competente que los humanos en la tarea para la que fue diseñado. Todas estas cuestiones, por supuesto, implicarían la consciencia de sí mismo, del mundo y de su posición en él, algo para lo cual no está capacitado. AlphaGo no está en el mundo.

En la actualidad, y tras solo algunos años de investigación y desarrollo, estamos en medio de un panorama más complejo que el de aquellos duelos. Los recientes avances en el campo de la inteligencia artificial han desembocado en la emergencia de una nueva categoría conocida como inteligencia artificial generativa, que ha suscitado un creciente interés académico y público. Esta modalidad de IA se distingue por su abierta capacidad para generar contenido original, trascendiendo las funciones mayormente analíticas de los modelos de IA antecesores. Las aplicaciones de esta tecnología generativa abarcan desde la creación de imágenes y composiciones musicales hasta la generación de texto. Aunque considerarlas creativas es aún motivo de debate, estos modelos han empezado a tomar relevancia en terrenos que se consideraban exclusivos de la capacidad humana. El asombro de Sedol ante la jugada 37 de AlphaGo, es ahora un sentimiento compartido en distintos ámbitos.

Un ejemplo destacado de esta tecnología es ChatGPT, lanzado el 30 de noviembre de 2022. Se trata de un sofisticado modelo de procesamiento de lenguaje natural

5. Mark Coeckelbergh, *Imagination and Principles. An Essay on the Role of Imagination in Moral Reasoning* (Londres: Palgrave Macmillan, 2007), 11-20, https://doi.org/10.1057/9780230589803

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

desarrollado por OpenAI. Esta aplicación fue entrenada utilizando un método de aprendizaje por refuerzo con retroalimentación humana (RLHF, por sus siglas en inglés), el cual le ha permitido afinar sus capacidades a partir de la interacción con usuarios[6]. A través de un proceso de entrenamiento intensivo en un vasto *corpus* de texto, ChatGPT ha aprendido a generar respuestas que son coherentes y relevantes para el contexto proporcionado por las entradas de personas en todo el mundo, en distintos idiomas, estilos, niveles de detalle y en una gran variedad de tópicos.

Sin embargo, a pesar de las impresionantes capacidades de ChatGPT, aún está en discusión si puede ser considerado una inteligencia artificial general. Aunque este *chatbot* puede manejar una amplia gama de temas y adaptarse a diferentes contextos, su competencia se limita al dominio del procesamiento del lenguaje, lo que lo asemeja más a una IA dedicada. Aunque resulte obvio, es necesario tener en cuenta que ChatGPT no puede realizar tareas fuera de este ámbito, como conducir un vehículo o diagnosticar una enfermedad, lo que requeriría, por lo mínimo, el análisis de información multimodal y un entrenamiento completamente distinto. Además, aunque ChatGPT puede generar respuestas basadas en los patrones que ha aprendido de su *corpus* de adiestramiento, no tiene la capacidad de aprender nuevas habilidades o razonar de la misma manera que un humano.

Sobre este último punto, algunos investigadores afirman que ChatGPT presenta varias limitaciones técnicas[7]. Entre ellas, el conocido hecho de que "alucina", es decir, genera respuestas incorrectas o sin sentido que pueden pasar por razonables. OpenAI ha hecho esfuerzos para mitigar este problema, pero la cuestión persiste. Además, el modelo tiene dificultades para realizar inferencias espaciales, temporales o físicas, así como para predecir y explicar comportamientos y procesos psicológicos humanos. Otra limitación es su inconsistencia, ya que puede generar salidas contradictorias con la misma entrada.

Finalmente, los autores afirman que a pesar de que ChatGPT puede responder a diversas preguntas y generar texto coherente, no posee conciencia, emociones ni experiencias subjetivas. Por razonadas que parezcan sus respuestas, el contenido que genera se basa en los millones de textos que ha revisado en su entrenamiento. En este punto, es esencial enfatizar que nuestras expectativas y consideraciones sobre estas herramientas se anclan en la forma en que definimos y entendemos

6. "Introducing ChatGPT", *OpenAI* (blog), 30 de noviembre de 2022, https://openai.com/blog/chatgpt

7. Chaoning Zhang, *et al.*, "One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era", *arXiv:2304.06488v1* (2023), https://doi.org/10.48550/arXiv.2304.06488

el término "inteligencia". Este concepto, ambiguo y multidimensional, es aún motivo de debate en ámbitos científicos y filosóficos, y es normal que la primera intención sea relacionarla a la inteligencia humana.

Sin embargo, hay que contemplar que incluso si se logra desarrollar una inteligencia artificial general, existe la probabilidad de que esta nueva entidad cognitiva posea una cualidad de inteligencia divergente a la nuestra. Además, hay que señalar la confusión que se presenta comúnmente entre la capacidad para resolver tareas percibidas como complejas por los humanos y la noción de inteligencia[8]. Esta es una muestra de cómo nuestros prejuicios y presuposiciones humanas influyen en las expectativas para la inteligencia artificial.

En todo caso, es innegable que herramientas como ChatGPT representan un avance significativo en el camino hacia una eventual inteligencia artificial general. Además, este es el objetivo de investigación de algunas de las empresas tecnológicas más avanzadas del mundo, como IBM, DeepMind, OpenAI y la recientemente formada xIA de Elon Musk. Incluso, se ha empezado a usar el concepto de "superinteligencia artificial"[9], como una meta superior a la AGI, y a la cual se plantea llegar en cuestión de algunos años[10]. Aunque todavía se debate si estos objetivos son alcanzables, se ha vuelto una cuestión apremiante para la integridad de la especie humana el garantizar que, si se logra, una IA de estas características actúe siempre en beneficio de los seres humanos y en consonancia con sus valores y principios.

El problema de la alineación en la inteligencia artificial

La eventual consecución de una IA que pueda aportar puntos de vista y tomar decisiones, con un razonamiento que exceda las habilidades intelectuales humanas, sin duda marcará un hito en la historia, aunque por ahora no es posible predecir de qué manera. Teniendo en cuenta el escenario actual, es válido afirmar que la inteligencia artificial es una tecnología que podría generar beneficios inmensos,

8. Johan Egbert (Hans) Korteling, *et al.*, "Human- versus Artificial Intelligence", *Frontiers in Artificial Intelligence* 4 (2021): 1-13, https://doi.org/10.3389/frai.2021.622364

9. Bill Hibbard, *Super-Intelligent Machines* (Nueva York: Springer, 2002), 99-109, https://doi.org/10.1007/978-1-4615-0759-8; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 52-57.

10. Marvie Basilan, "Elon Musk Predicts 'Digital Superintelligence' To Arrive In 5-6 Years, Launches AI Company", *International Business Times*, 13 de julio de 2023, https://www.ibtimes.com/elon-musk-predicts-digital-superintelligence-arrive-5-6-years-launches-ai-company-3704835; Jan Leike e Ilya Sutskever, "Introducing Superalignment", *OpenAI* (blog), 5 de julio de 2023, https://openai.com/blog/introducing-superalignment

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

pero al mismo tiempo alberga riesgos sin precedentes. Por tanto, a medida que los sistemas de IA se vuelven más potentes y prevalentes, su alineación con los valores humanos se convierte en un imperativo existencial.

Este asunto constituye la preocupación central de la alineación de IA, un campo que busca asegurar que los sistemas de IA actúen de manera beneficiosa, en lugar de perjudicial, para la humanidad. La alineación de la IA es un esfuerzo arraigado en la intersección de la investigación técnica de la IA y abordajes filosóficos, particularmente la ética. Sus esfuerzos se podrían condensar en la pregunta ¿cómo diseñar sistemas de IA que no solo comprendan, sino que también respeten y se adhieran a los valores humanos y a los marcos éticos? La anterior es una cuestión de una alta complejidad, sobre todo al analizar sus implicaciones. Por ejemplo, ¿qué constituye a los valores humanos? ¿Cómo pueden codificarse o enseñarse estos valores a un sistema de IA? Y adicionalmente, ¿cómo se puede asegurar que una IA avanzada no encuentre una manera de eludir estas restricciones?

Iason Gabriel, investigador en temas de política y ética para Google Deepmind, plantea en su trabajo la diferenciación entre las partes técnicas y las normativas de la alineación de la IA, una distinción útil para seguir avanzando en el tema[11]. Gabriel distingue primero la perspectiva técnica, que se relaciona con la creación de mecanismos y algoritmos que permiten a un sistema de IA actuar de acuerdo con ciertos principios o valores. Pero luego, desde el punto de vista normativo, el desafío es identificar y seleccionar estos principios. Los dos son dimensiones interdependientes del problema que necesitan una reflexión continua. Esto implica que el desarrollo de IA éticamente alineada es una tarea que requiere un intercambio interdisciplinario constante y consciente. La idea de que el desarrollo técnico y el normativo pueden ser tratados de manera independiente, como si fuese posible "cargar" los valores deseados en la IA al final del proceso, resulta simplista, dada la forma en que se entrenan muchos modelos de inteligencia artificial.

Por ejemplo, dentro del contexto de las técnicas de aprendizaje automático, el aprendizaje supervisado proporciona un marco viable para la alineación de la IA. Este método se enfoca en entrenar un modelo para identificar y responder a patrones mediante el uso de datos etiquetados, lo que permite a un humano evaluar la eficacia del modelo. En este contexto, el aprendizaje supervisado puede establecer un camino para alinear la IA con ciertos principios éticos, siempre y

11. Iason Gabriel, "Artificial Intelligence, Values, and Alignment", *Minds and Machines* 30 (2020): 411-437, https://doi.org/10.1007/s11023-020-09539-2

cuando estos principios puedan ser codificados en los datos de entrenamiento del sistema. Lo anterior abre la puerta a un abordaje utilitarista de la alineación de la inteligencia artificial, donde los sistemas se diseñan y programan con el objetivo de maximizar la felicidad o el bienestar general. Siguiendo los principios del utilitarismo, la acción moralmente correcta será la que dé como resultado la mayor felicidad para el mayor número de individuos en el futuro.

Este principio se podría manifestar en métodos como el aprendizaje por refuerzo, en el cual un agente se capacita para maximizar una recompensa numérica recibida del entorno. No obstante, este enfoque presenta una serie de dilemas, como cuál será el criterio se etiquetan los datos, cómo cuantificar conceptos abstractos como felicidad o bienestar, cómo equilibrar intereses de distintos individuos, o el dilema sobre la validez de los medios para conseguir el objetivo. El asunto se complica al sumar el enfoque deontológico, que consistiría en establecer ciertos principios morales inviolables que los sistemas de IA deben respetar, independientemente de las consecuencias. Estos principios, dados de "arriba hacia abajo" podrían estar relacionados con leyes, derechos humanos, o principios morales generales. Un ejemplo relacionado serían las "leyes de la robótica", una serie de normas que aparecen en las historias de ciencia ficción creadas por Isaac Asimov. No obstante, y tal como sucede en algunas historias del autor, este tipo de normas pueden entrar en conflicto o interpretarse de manera ambigua cuando se llevan al campo de lo particular. Otros investigadores exploran las dificultades de dictar principios explícitos a agentes inteligentes que no manejan la gran cantidad de sutilezas e información tácita que los humanos dan por sentado en un mundo altamente complejo[12].

Finalmente, Gabriel menciona otro cuerpo de posibilidades que eluden, hasta cierto punto, la necesidad de codificar explícitamente los principios morales en los sistemas de IA. Los investigadores en el campo han considerado métodos que implican un abordaje de "abajo hacia arriba" como el aprendizaje por refuerzo inverso (IRL), que busca extraer una función de recompensa a partir de comportamientos óptimos observados, y otros enfoques evolutivos que valoran la interacción de los agentes con su entorno para maximizar la recompensa. En otras palabras, los sistemas inferirían principios éticos mediante un proceso de aprendizaje a través de la observación, la experiencia y el refinamiento continuo.

12. Nate Soares, "The Value Learning Problem", *Artificial Intelligence Safety and Security* (2018), 89-97, https://www.semanticscholar.org/paper/The-Value-Learning-Problem-Soares/f042e8db015e31b13a8aa2a8b814ff745b3f4f49

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Estas técnicas podrían aplicarse en la alineación de valores de IA de diversas formas, incluyendo el aprendizaje a partir de la observación de la conducta, el análisis de grandes conjuntos de datos para entender las preferencias de un gran número de personas, y el uso de procesos evolutivos para seleccionar agentes que muestren comportamientos "más morales" en un mundo social simulado. Y desde esta perspectiva, surge también la posibilidad en la que centra este documento: la comprensión de relatos. Expertos en el campo, como Mark Riedl y Brent Harrison, proponen que agentes de inteligencia artificial capacitados para analizar y decodificar relatos, tendrían el potencial de discernir los valores implícitos en diversos ejemplos narrativos de una cultura específica.

Dada la capacidad intrínseca de la narrativa para encapsular la experiencia humana, este enfoque podría ser la apertura a un amplio rango de posibilidades relacionadas con la ética, la experiencia y la consciencia: "We hypothesize that an intelligent entity can learn what it means to be human by immersing itself in the stories it produces"[13]. Esencialmente, estos investigadores argumentan que el conocimiento sociocultural implícito y explícito codificado en las historias puede servir para producir una señal de recompensa alineada con valores humanos para el aprendizaje por refuerzo en inteligencias artificiales. Esto permitiría aportar a una de las limitaciones de la alineación de valores discutidas anteriormente: no es posible contemplar todas las posibles encrucijadas y dilemas morales para dictar normas generales explícitas. Los investigadores exploraron las primeras bases sobre cómo se pueden usar ejemplos narrativos —que surjan de una colaboración colectiva de humanos— para conseguir una inteligencia artificial "aculturada", es decir, que ha adoptado los valores implícitos en una cultura o sociedad en particular. La conclusión fue que brindar la capacidad de leer y comprender historias a las inteligencias artificiales puede ser el medio más conveniente para cultivar inteligencias artificiales que se integren en las sociedades humanas y contribuyan a nuestro bienestar general.

No obstante, aprender valores de las historias también presenta nuevos desafíos, como también reconocen los investigadores. Por un lado, las historias escritas en lenguaje natural pueden contener eventos y acciones no ejecutables por una IA, o al ser escritas por humanos para un público humano hacen uso de conocimiento tácito, compartido o propio del sentido común, dejando muchas cosas sin mencionar, y en ese

13. Mark O. Riedl y Brent Harrison, "Using Stories to Teach Human Values to Artificial Agents", *AAAI Workshop: AI, Ethics, and Society* (2016), https://www.semanticscholar.org/paper/Using-Stories-to-Teach-Human-Values-to-Artificial-Riedl-Harrison/33b53abdf2824b2cb0ee083c284000df4343a33e

mismo sentido, se suelen omitir eventos que no impactan directamente en la historia, sin mencionar que operaciones propias de la narración, como *flashbacks* o *flashforwards*, podrían generar problemas de comprensión. Pero más allá de los primeros desafíos formales, es esencial reconocer el potencial intrínseco de la narrativa que opera como una herramienta esencial para conferir significado a la experiencia y moldear nuestra percepción del mundo. Integrarla en la IA podría transformar la forma en que las máquinas interactúan y entienden el contexto humano. Es importante destacar que esta confluencia entre narrativa e IA trasciende lo meramente técnico, introduciendo reflexiones filosóficas sobre la consciencia y la formación de la identidad en entidades inteligentes. Así, al aspirar a una "inteligencia narrativa" en la IA, se nos insta a reflexionar sobre las repercusiones de la narrativa en la conciencia, identidad y ética humanas, temas que se empezarán a considerar a continuación.

Elementos de la narrativa para la alineación de la IA

La narrativa, desde una perspectiva académica y cultural, puede ser definida como una estructura discursiva mediante la cual se representan y organizan secuencias de eventos y experiencias en una forma coherente y significativa, sirviendo como vehículo para la construcción y transmisión de conocimiento, valores y perspectivas. A través de la historia de la humanidad ha sido la principal herramienta con la que las civilizaciones han registrado, interpretado y compartido sus vivencias, mitologías y visiones del mundo[14]. Su importancia trasciende la mera exposición de hechos, ya que, en su esencia, la narrativa encarna la cosmovisión de un grupo o sociedad, reflejando y moldeando a la vez las normas, valores y aspiraciones culturales. Además, actúa como una brújula sociocultural que guía y orienta a las comunidades en su entendimiento del pasado, su interpretación del presente y su proyección hacia el futuro.

En este sentido, la narrativa puede ser entendida como un pilar fundamental en la construcción y consolidación de identidades, tanto personales como colectivas. Por consiguiente, la aptitud para interpretar narrativas se presenta como una posibilidad en el proceso de asimilación de principios éticos y en la profundización del entendimiento de la esencia humana en el campo de la IA. Conforme a la concepción propuesta por algunos autores la narrativa se manifiesta como un medio esencial para conferir coherencia y significado al mundo circundante: "As such, we characterise

14. Joseph Campbell, *The Hero with a Thousand Faces* (Princeton: Princeton University Press, [1949] 2004), 351-358; Yuval Noah Harari, *Sapiens. De animales a dioses: Una breve historia de la humanidad* (Barcelona: Debate, 2017), 37-54.

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

narrative as a mediation of human experience and understanding of the social world through the process of emplotment, which is the organisation of heterogeneous elements in a meaningful synthesis"[15]. Dicha narrativa incorpora en su composición aspectos fundamentales tales como causalidad, temporalidad, espacio y la vivencia de entidades conscientes interrelacionadas con los sucesos que se narran (personajes). Es por esto que las narraciones, desde las de índole personal hasta las míticas, desempeñan un papel crucial en la atribución de significado en varios niveles.

Cabe mencionar que el concepto de la inteligencia narrativa[16], es decir, la capacidad de entender y crear narrativas, se ha presentado antes en el campo de la inteligencia artificial. Los antecedentes datan de los años de 1970 y principios de los de 1980, cuando en el marco de investigaciones en IA, la narrativa textual fue un elemento de interés debido a la integración del procesamiento verbal y la comprensión de situaciones complejas. Las limitaciones de la tecnología y una desaceleración en el campo de investigación de la IA de la época hicieron que la narrativa se dejara de lado en pos de metas más verificables, pero los cruces entre la narrativa y la tecnología siguieron adelante en líneas como los relatos interactivos, el hipertexto, las interfaces o los videojuegos.

Actualmente, las conexiones emergentes en el campo nos llevan a reflexionar acerca de las características de la narrativa y su contribución potencial al avance de la inteligencia artificial. Específicamente, surgen interrogantes acerca de lo que una inteligencia narrativa podría significar en el contexto de las relaciones cognitivas que se derivan de la capacidad humana para crear y entender historias. Para enriquecer esta discusión, a continuación, se explorarán tres aspectos clave: la identidad narrativa, el conocimiento tácito y la experiencialidad narrativa.

La identidad narrativa

La identidad narrativa es una idea central en la filosofía de Paul Ricoeur, que se enmarca en el contexto más amplio de su ontología hermenéutica[17]. El autor argumenta que las historias que contamos sobre nosotros mismos y sobre otros son cruciales para la formación de nuestra identidad. La narrativa serviría como

15. Wessel Reijers y Mark Coeckelbergh, *Narrative and technology ethics* (Cham: Palgrave Macmillian, 2020), 42.
16. Michael Mateas y Phoebe Sengers, "Narrative Intelligence", en *Narrative Intelligence*, eds. Michael Mateas y Phoebe Sengers (Ámsterdam y Filadelfia: John Benjamins Publishing Company, 2003), 1-25
17. Paul Ricoeur, "La vida: un relato en busca de narrador", *Ágora: Papeles de filosofía* 25, no. 2 (2006): 9-22.

un mediador entre el cambio y la continuidad, y mediante un proceso de "entramado", la narrativa también permite dar sentido y coherencia a las eventualidades de la vida. Es a través del acto de narrar que llegamos a comprendernos y reconocernos a nosotros mismos, y por extensión, a los demás.

Por tanto, la identidad narrativa no es meramente una representación de nosotros mismos, sino una acción que forma y configura nuestra identidad en un proceso dialéctico. En su trilogía "Tiempo y narrativa", Ricoeur argumenta que la identidad no puede ser entendida como una entidad estática, sino como algo que se desenvuelve y cambia a lo largo del tiempo, siendo afectada por los eventos, decisiones y experiencias que se entrelazan en la narrativa de la vida de una persona. Además, plantea que la identidad tiene dos componentes interrelacionados pero distintos: el *idem* y el *ipse*. El *idem*, o "la mismidad", es la parte de la identidad que permanece constante a lo largo del tiempo, que nos permite reconocernos a nosotros mismos a pesar de los cambios que experimentamos. Por otro lado, el *ipse*, o "la ipseidad", representa la parte de la identidad que puede cambiar y transformarse a lo largo del tiempo.

El concepto de identidad narrativa de Ricoeur, entrelazado con la comprensión que tenemos de obras literarias y las historias culturales que valoramos, arroja luz sobre nuestra percepción de autonomía y responsabilidad moral. Ricoeur argumenta que no podemos reclamar una autonomía total, ya que nuestra identidad está intrínsecamente vinculada a las narrativas que tejemos junto con otros y las que repercuten desde el entorno cultural. Estas narrativas, por ende, influyen en cómo evaluamos nuestras acciones y las de quienes nos rodean. Nuestra interpretación y reinterpretación constante de estas historias refleja nuestra búsqueda de significado y lugar dentro del vasto entramado de relatos humanos. No obstante, y a pesar de esta profunda conexión con las narrativas externas, Ricoeur enfatiza nuestra responsabilidad individual[18]. Tenemos el poder de elegir cómo interpretamos y construimos nuestras propias narrativas. De este modo, nuestra responsabilidad moral se convierte en un reflejo directo de nuestra identidad narrativa, formada e instruida tanto por nuestra agencia personal como por las historias que nos rodean.

Lo anterior resuena con algunos aspectos de la psicología constructivista, cuyos postulados son útiles para seguir argumentando que nuestro pasado como individuos es algo que reconstruimos y resignificamos constantemente. Al igual que en una narrativa, las experiencias memorables se reconstruyen al mismo tiempo que se les otorga causalidad y sentido general. Como sostiene el psicólogo Charles

18. Ricoeur, *La vida*, 19.

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Fernyhough en lugar de poseer un recuerdo fijo de nuestro pasado, como si fuera una instantánea, lo construimos de nuevo cada vez que lo recordamos. Para el autor, recordar es tanto un acto narrativo como el producto de un proceso neurológico: "We are all natural-born storytellers; we engage in acts of fiction-making every time we recount an event from our pasts. We are constantly editing and remaking our memory stories as our knowledge and emotions change"[19].

Del mismo modo, investigaciones más recientes siguen desarrollando el concepto de identidad narrativa como central en una concepción de una identidad propia desde una perspectiva psicológica[20]. A pesar de que la naturaleza multidimensional del ser humano puede ser abordada desde múltiples perspectivas, la narrativa permite discernir componentes esenciales y constitutivos de la identidad personal, ubicándolos en una trayectoria causal y temporal. Tal estructuración no solo facilita la emergencia de una autoconsciencia coherente, sino que contextualiza al individuo dentro del vasto entramado sociocultural en el que se halla inmerso. Es pertinente señalar la resonancia de este principio en manifestaciones culturales, como es el caso de la saga cinematográfica "Blade Runner". En este contexto fílmico, se postula que para que un replicante —un humanoide artificial— pueda asimilarse eficazmente a la condición humana, garantizando su estabilidad cognitiva y emocional, se le debe dotar de una historia previa en su memoria, incluso cuando el mismo es plenamente consciente de la ficcionalidad de dicho pasado. Lo mismo podría llegar a contemplarse para un agente de inteligencia artificial.

El conocimiento tácito

Hay un momento clave en el único juego ganado por Lee Sedol, la jugada 78, un movimiento que AlphaGo solo calculaba en un 0,007 % de probabilidad de ejecutarse por un humano. La experiencia y talento del campeón coreano le permitieron ver con claridad la jugada, describiéndola como la "única posible", así no hubiera llegado a esa conclusión a través de cálculos o probabilidades. Lo anterior se relaciona con que mucho del saber humano es un "conocimiento tácito", un

19. Charles Fernyhough, *Pieces of Light: The New Science of Memory* (Londres: Profile Books, 2012), 15-31.

20. Benjamin A. Rogers, *et al.*, "Seeing your life story as a Hero's Journey increases meaning in life", *Journal of Personality and Social Psychology* 125, no. 4 (2023): 752-778, https://doi.org/10.1037/pspa0000341; Dan P. McAdams, "'First we invented stories, then they changed us': The Evolution of Narrative Identity", *Evolutionary Studies in Imaginative Culture* 3, no. 1 (2019): 1-18, https://doi.org/10.26613/esic.3.1.110; Vinicio Busacchi, "On Narrative Identity", en 5th *International Multidisciplinary Scientific Conference on Social Sciences and Arts. Conference Proceedings, vol. 5. Ancience Science, vol 2.2* (Sofía: SGEM, 2018), 555-563.

término acuñado por el polímata húngaro-británico Michael Polanyi para referirse a un tipo de conocimiento difícil de hacer explícito en palabras o criterios, y que atiende a la experiencia o la intuición[21]. Haciendo uso de este concepto para argumentar la imposibilidad de una IA general, Ragnar Fjelland aporta como ejemplos las habilidades de nadar o el andar en bicicleta, donde pocas personas podrían hacer explícitos todos los aspectos físicos involucrados, y aun si pudieran hacerlo, poco o nada influiría en su desempeño[22]. Según el autor, este es un modo de conocimiento que no se presenta en los programas automáticos.

Fjelland resalta que, para Polanyi, el conocimiento tácito, dado por la experiencia, es necesario en los humanos para cimentar el conocimiento especializado[23], un proceso por el cual, hasta este punto, no pasa una IA. Esto es particularmente problemático para la alineación de la IA, ya que la manera en que los seres humanos aprenden y aplican determinados principios morales involucra un desarrollo temprano de conocimientos tácitos que pueden crecer y convertirse en elementos comunes a nivel intersubjetivo (basta considerar que muchas normas de tipo ético compartidas de manera colectiva, no son necesariamente explícitas en leyes o normativas escritas). Lo anterior resuena con la crítica del filósofo Hubert Dreyfus quien argumentó que las computadoras jamás podrían adquirir inteligencia, ya que no estaban dotadas de cuerpo, infancia, ni práctica cultural[24].

Esto no es un punto menor, porque hace una referencia directa a la manera de "estar en el mundo" que profundizan filósofos de la fenomenología como Merleau-Ponty, quien trabaja al cuerpo como un elemento central de la experiencia humana[25]. Sin cuerpo, experiencia acumulada, consciencia y percepción es muy difícil imaginar la cognición humana. No obstante, la comprensión de relatos podría concentrar el esfuerzo de una IA en discernir la información tácita que se encuentra embebida en los relatos. Las narraciones, como reflejos de la experiencia humana, actúan como depósitos de información, donde se encapsulan tanto conocimientos explícitos como tácitos. Una historia, por ejemplo, puede contener lecciones aprendidas a partir de fracasos y éxitos, intuiciones sobre la naturaleza humana, o reflexiones sobre las relaciones interpersonales.

---

21. Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy* (Londres: Routledge, [1958] 1998), 90-99.
22. Ragnar Fjelland, "Why general artificial intelligence will not be realized", *Humanities and Social Sciences Communications* 7, no. 10 (2020), https://doi.org/10.1057/s41599-020-0494-4
23. Fjelland, *Why*, 3.
24. Hubert L. Dreyfus, Stuart E. Drey-fus y Lotfi A. Zadeh, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (Nueva York: The Free Press, 1986), 5.
25. Maurice Merleau-Ponty, *Fenomenología de la percepción* (Barcelona: Planeta DeAgostini, [1945], 1993), 100.

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Aunque las narrativas a menudo no explicitan directamente estas lecciones, las sugieren a través de metáforas, conflictos, y desenlaces. A través de los relatos, las experiencias del pasado, las expectativas del futuro y las vivencias del presente se entrelazan en una compleja trama de significado. Si una IA fuera capaz de procesar y comprender este entrelazamiento, tendría acceso a un nivel profundo de entendimiento sobre la cognición y percepción humanas, un nivel que va más allá de la mera información y se sumerge en la esfera de la experiencia vivida. En ese sentido, la capacidad de una IA para interpretar y aprender de los relatos no solo podría enriquecer su comprensión sobre la condición humana, sino que, más allá de eso, podría ser una vía para que estos agentes artificiales desarrollen una forma de "empatía computacional"[26] donde, aunque no tengan experiencias vividas ni emociones como tal, podrían llegar a entender y anticipar las respuestas y necesidades humanas a partir de la vasta reserva de sabiduría encapsulada en nuestras narrativas. Por tanto, mientras que la experiencia tácita es una característica innatamente humana y, por lo tanto, es difícilmente replicable en su totalidad por las máquinas, las narrativas podrían servir como un puente entre la IA y este tipo de conocimiento. A través del estudio y análisis de relatos, la IA podría acercarse, aunque sea parcialmente, a la rica fuente de experiencia y entendimiento humano que se encuentra codificada en nuestras historias.

La experiencialidad narrativa

La experiencialidad, término introducido y definido por la narratóloga Monika Fludernik como "the quasi-mimetic evocation of real-life experience"[27], emerge como un pilar central en la narratología contemporánea —sobre todo en la denominada narratología posclásica[28]— reflejando la relación intrínseca entre la representación narrativa y la experiencia humana. Esta relación se enmarca en la capacidad del texto narrativo para evocar y resonar con la familiaridad cognitiva y experiencial del lector. El concepto de experiencialidad ha sido central en el trabajo de Fludernik y ha influido en otros investigadores, que han elaborado reflexiones propias sobre él[29].

26. Özge Nilay Yalçın y Steve Dipaola, "Modeling empathy: building a link between affective and cognitive processes", *Artificial Intelligence Review* 53, no. 1 (2020): 2983-3006, https://doi.org/10.1007/s10462-019-09753-0

27. Monika Fludernik, *Towards a natural Narratology* (Londres: Taylor & Francis e-Library, 1996), 12.

28. David Herman, "Scripts, Sequences, and Stories: Elements of a Postclassical Narratology", *Publications of the Modern Language Association of America* 112, no. 5 (1997): 1046-1059, https://doi.org/10.2307/463482

29. Marco Caracciolo, "Experientiality", en *The living handbook of narratology*, eds. Peter Hühn *et al.* (Hamburgo: Hamburg University Press, 2014), https://www-archiv.fdm.uni-hamburg.de/lhn/node/102.html

La experiencialidad se arraiga en la activación de parámetros cognitivos "naturales" que reflejan la experiencia humana, como la encarnación de facultades cognitivas, la comprensión de la acción intencional, la percepción de la temporalidad y la evaluación emocional[30]. Estas estructuras cognitivas actúan como puentes entre las vivencias reales y las representaciones semióticas de dichas vivencias en los textos narrativos. En otras palabras, para interpretar historias hacemos uso de capacidades cognitivas que empleamos en la experiencia vivida. Un aspecto esencial de esta discusión se centra en el papel del cuerpo en la narrativa. Fludernik enfatiza la preeminencia del concepto de encarnamiento (surgido de la tradición fenomenológica), argumentando que este concepto engloba todas las demás categorías cognitivas y experiencialidad, ya que evoca la existencia en un marco específico de tiempo y espacio[31].

Además, en el contexto de la experiencialidad, la narrativa aborda la temporalidad humana dinámica, que está siempre acompañada por procesos evaluativos emocionales. Estos procesos dan lugar a la representación de experiencias que son memorables y emotivas, y que a menudo reflejan las evaluaciones de un narrador o un personaje[32]. Lo anterior, por supuesto, abre un dilema en el tema que se viene trabajando en este documento. Marie-Laure Ryan reafirma la idea de que, para entender las narraciones, son necesarias diferentes actividades cognitivas que cotidianamente usan los seres humanos para relacionarse con su entorno.

La autora postula que, para que la inmersión en la narrativa pueda darse, se requiere la extrapolación de las cualidades de nuestra experiencia del mundo para aplicarlas al mundo de ficción, de modo que podamos entender lo narrado y relacionarnos con ello de manera imaginativa. Ryan sintetiza la idea diciendo que "A brain in a vat would be unable to narrate or to understand stories; it is our embodied life experience that enables us to do so"[33]. Si se piensa, la idea de un cerebro que flota en un contenedor, sin corporeidad ni experiencia previa, se parece conceptualmente a la idea de una inteligencia artificial general, sin cuerpo o experiencia, en consonancia con la crítica de Dreyfus.

---

30. Monika Fludernik, "Natural Narratology and Cognitive Parameters", en *Narrative Theory and the Cognitive Sciences*, ed. David Herman (Stanford: Center for the Study of Language and Information Publications, 2003), 243-267.

31. Fludernik, *Towards*, 30.

32. Fludernik, *Towards*, 15.

33. Marie-Laure Ryan, "Narrative mapping as cognitive activity and as active participation in storyworlds", *Frontiers of Narrative Studies* 4, no. 2 (2018): 232-247, https://doi.org/10.1515/fns-2018-0020

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

Por tanto, emerge la pregunta: ¿son la experiencia y la corporeidad condiciones sin las cuales no puede haber una verdadera comprensión de relatos? Justamente, la reflexión sobre el papel de la experiencia y la corporeidad en la comprensión narrativa conduce a una discusión más amplia sobre la posibilidad de una inteligencia narrativa artificial. El análisis de Fludernik, Ryan, y otros, sugiere que la corporeidad y la experiencia de vida son elementos cruciales para entender y participar en la narrativa. Por lo tanto, cualquier desarrollo de una inteligencia narrativa en el campo de la inteligencia artificial tendría que considerar cómo se pueden incorporar estos elementos en un sistema que, en su esencia, carece de experiencia vital y corporeidad.

Este desafío se extiende más allá de la mera programación de rutinas cognitivas o habilidades lingüísticas y plantea preguntas filosóficas profundas sobre la naturaleza de la experiencia y la cognición. Relacionado estrechamente con la cuestión previa sobre el conocimiento tácito, la discusión que se deriva es si una auténtica inteligencia narrativa debe estar necesariamente precedida por la capacidad de experimentar de manera subjetiva, o si, por el contrario, las propias narrativas pueden actuar como un instrumento fundamental que facilite a los agentes de inteligencia artificial la comprensión de tal experiencia subjetiva. Por el momento, estas cuestiones permanecerán abiertas, a la espera de una resolución a través del progreso futuro en este ámbito.

Conclusiones

El avance en el campo de la inteligencia artificial no solo posee un carácter técnico, sino que también suscita interrogantes constantes respecto a la cognición, consciencia e inteligencia humanas. Lo mismo sucede cuando se busca la alineación de valores en este campo, cuando se suman preguntas éticas que han sido parte de la discusión filosófica durante siglos. De ahí que las consideraciones sobre la narrativa se conviertan en una perspectiva viable de incluir en la discusión, dada su profunda relación con la manera de concebir el mundo, la identidad personal y las relaciones humanas.

La narrativa, al ser una forma esencial de estructurar y dar sentido a la experiencia humana, ofrece una vía prometedora para la alineación de valores en la IA. Este enfoque puede llevar a una "empatía computacional", donde las IA, aunque carentes de experiencias vividas, puedan anticipar y responder adecuadamente a las necesidades humanas a partir del conocimiento derivado de nuestras historias. Esta capacidad no solo enriquecería su funcionamiento, sino que también fortalecería su alineación con valores humanos.

No obstante, la inteligencia narrativa requiere de elementos de experiencia para la comprensión de historias, lo cual presenta un interesante dilema. Por un lado, la comprensión y formulación de historias hace parte de una constitución de la identidad, así como una concepción ética del propio ser, pero al mismo tiempo, su comprensión requiere de experiencia acumulada. Por tanto, el desarrollo de una inteligencia narrativa, en el marco de la inteligencia artificial, podría llegar a ser una prueba de definitiva para un agente artificial situado en el mundo.

Por otro lado, es vital no perder de vista que el proyecto de una inteligencia artificial general podría no culminar en el resultado esperado. A pesar de los avances significativos que alimentan un entusiasmo generalizado hacia estas tecnologías, posturas como la de Fjelland son esenciales para evidenciar el camino aún por recorrer. Otros escenarios, como la concepción de una forma de inteligencia diametralmente distinta a la humana, también son posibilidades viables. En tanto, la contribución multidisciplinar realizada hasta el momento refleja interesantemente el progreso tecnológico y cultural humano, capaz de convocar múltiples campos del saber y generar discusiones en torno a un proyecto de una envergadura sin precedentes.

Bibliografía

Fuentes secundarias

[1] "Introducing ChatGPT". *OpenAI* (blog), 30 de noviembre de 2022. https://openai.com/blog/chatgpt

[2] Basilan, Marvie. "Elon Musk Predicts 'Digital Superintelligence' To Arrive In 5-6 Years, Launches AI Company". *International Business Times*, 13 de julio de 2023. https://www.ibtimes.com/elon-musk-predicts-digital-superintelligence-arrive-5-6-years-launches-ai-company-3704835

[3] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

[4] Busacchi, Vinicio. "On Narrative Identity". En *5th International Multidisciplinary Scientific Conference on Social Sciences and Arts. Conference Proceedings, vol. 5. Ancience Science, vol 2.2*, 555-563. Sofía: SGEM, 2018.

[5] Campbell, Joseph. *The Hero with a Thousand Faces*. Princeton: Princeton University Press, 2004.

[6] Caracciolo, Marco. "Experientiality". En *The living handbook of narratology*, editado por Peter Hühn, Jan-Christoph Meister, John Pier y Wolf Schmid. Hamburgo: Hamburg University Press, 2014. https://www-archiv.fdm.uni-hamburg.de/lhn/node/102.html

[7] Coeckelbergh, Mark. *Imagination and Principles. An Essay on the Role of Imagination in Moral Reasoning*. Londres: Palgrave Macmillan, 2007. https://doi.org/10.1057/9780230589803

[8] Dreyfus, Hubert L., Stuart E. Drey-fus y Lotfi A. Zadeh, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Nueva York: The Free Press, 1986.

[9] Fernyhough, Charles. *Pieces of Light: The New Science of Memory*. Londres: Profile Books, 2012.

[10] Fjelland, Ragnar. "Why general artificial intelligence will not be realized". *Humanities and Social Sciences Communications* 7, no. 10 (2020). https://doi.org/10.1057/s41599-020-0494-4

[11] Fludernik, Monika. *Towards a natural Narratology*. Londres: Taylor & Francis e-Library, 1996.

[12] Fludernik, Monika. "Natural Narratology and Cognitive Parameters". En *Narrative Theory and the Cognitive Sciences*, editado por David Herman, 243-267. Stanford: Center for the Study of Language and Information Publications, 2003.

[13] Gabriel, Iason. "Artificial Intelligence, Values, and Alignment". *Minds and Machines* 30 (2020): 411-437. https://doi.org/10.1007/s11023-020-09539-2

[14] Harari, Yuval Noah. *Sapiens. De animales a dioses: Una breve historia de la humanidad*. Barcelona: Debate, 2014.

[15] Herman, David. "Scripts, Sequences, and Stories: Elements of a Postclassical Narratology". *Publications of the Modern Language Association of America* 112, no. 5 (1997): 1046-1059. https://doi.org/10.2307/463482

[16] Hibbard, Bill. *Super-Intelligent Machines*. Nueva York: Springer, 2002. https://doi.org/10.1007/978-1-4615-0759-8

[17] Jayanti, Vikram. dir., *Game Over: Kasparov and the Machine*, 2003.

[18] Kiernan, Patrick. "Which is greater? The number of atoms in the universe or the number of chess moves?". *Nationals Museums Liverpool* (página web), s.f. www.liverpoolmuseums.org.uk/stories/which-greater-number-of-atoms-universe-or-number-of-chess-moves

[19] Kohs, Greg, dir. *AlphaGo*. 2017. YouTube, 13 de diciembre de 2017. https://www.youtube.com/watch?v=WXuK6gekU1Y

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078

[20] Korteling, Johan Egbert (Hans), Gillian Van De Boer-Visschedijk, Romy A.M. Blankendaal, Rudy Boonekamp y Aletta Eikelboom. "Human- versus Artificial Intelligence". *Frontiers in Artificial Intelligence* 4 (2021): 1-13. https://doi.org/10.3389/frai.2021.622364

[21] Mateas, Michael y Phoebe Sengers, "Narrative Intelligence". En *Narrative Intelligence*, editado por Michael Mateas y Phoebe Sengers, 1-25. Ámsterdam y Philadelphia: John Benjamins Publishing Company, 2003.

[22] McAdams, Dan P. "'First we invented stories, then they changed us': The Evolution of Narrative Identity". *Evolutionary Studies in Imaginative Culture* 3, no. 1 (2019): 1-18. https://doi.org/10.26613/esic.3.1.110

[23] Merleau-Ponty, Maurice. *Fenomenología de la percepción*. Barcelona: Planeta DeAgostini, [1945], 1993.

[24] Polanyi, Michael. *Personal Knowledge: Towards a Post-Critical Philosophy*. Londres: Routledge, [1958] 1998.

[25] Reijers, Wessel y Mark Coeckelbergh. *Narrative and technology ethics*. Cham: Palgrave Macmillian, 2020.

[26] Ricoeur, Paul. "La vida: un relato en busca de narrador". Ágora: Papeles de filosofía 25, no. 2 (2006): 9-22.

[27] Riedl, Mark O. y Brent Harrison. "Using Stories to Teach Human Values to Artificial Agents". *AAAI Workshop: AI, Ethics, and Society* (2016). https://www.semanticscholar.org/paper/Using-Stories-to-Teach-Human-Values-to-Artificial-Riedl-Harrison/33b53abdf2824b2cb0ee083c284000df4343a33e

[28] Rogers, Benjamin A., Herrison Chicas, John Michael Kelly, Emily Kubin, Michael S. Christian, Frank J. Kachanoff, Jonah Berger, Curtis Puryear, Dan P. McAdams y Kurt Gray. "Seeing your life story as a Hero's Journey increases meaning in life". *Journal of Personality and Social Psychology* 125, no. 4 (2023): 752-778. https://doi.org/10.1037/pspa0000341

[29] Ryan, Marie-Laure. "Narrative mapping as cognitive activity and as active participation in storyworlds". *Frontiers of Narrative Studies* 4, no. 2 (2018): 232-247. https://doi.org/10.1515/fns-2018-0020

[30] Shevlin, Henry, Karina Vold, Matthew Crosby y Marta Halina. "The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge". *EMBO Reports* 20 (2019): e49177. https://doi.org/10.15252/embr.201949177

[31] Soares, Nate. "The Value Learning Problem", *Artificial Intelligence Safety and Security* (2018). https://www.semanticscholar.org/paper/The-Value-Learning-Problem-Soares/f042e8db015e31b13a8aa2a8b814ff745b3f4f49

[32] Yalçın, Özge Nilay y Steve Dipaola. "Modeling empathy: building a link between affective and cognitive processes". *Artificial Intelligence Review* 53, no. 1 (2020): 2983-3006. https://doi.org/10.1007/s10462-019-09753-0

[33] Zhang, Chaoning, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon y Choong Seon Hong. "One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era". *arXiv:2304.06488v1* (2023). https://doi.org/10.48550/arXiv.2304.06488

Revista Colomb. Pensam. Estet. Historia arte
Edición 18 (Julio - diciembre de 2023) / pp. 53-76 / E-ISSN 2389-9794 / DOI: https://doi.org/10.15446/rcpeha.n18.106078