DESAFÍOS DE LA INTELIGENCIA ARTIFICIAL

BASES CONCEPTUALES, ASPECTOS ÉTICOS E IMPACTOS DE LA INTELIGENCIA ARTIFICIAL GENERATIVA

EL TÉRMINO

- El término "inteligencia artificial" se generalizó tras la famosa escuela de verano de Dartmouth de 1956, en Hanover, New Hampshire, USA, organizada por John McCarthy. En ella participaron Marvin Minsky, Nathaniel Rochester, Claude Shannon, Julian Bigelow, D.M. Mackay, Ray Solomonoff, John Holland, Nathanial Rochester, Oliver Selfridge, Allen Newell, Herbert Simon.
- Algunos consideran que el término "inteligencia artificial" nos predispone a pensar en ella de una determinada manera: que su objetivo es imitar la inteligencia humana, y creen que por esta razón debería evitarse el término. Muchos prefieren hablar más directamente de "aprendizaje automático". Esto plantea importantes cuestiones sobre cómo conceptualizamos la tecnología y qué esperanzas tenemos depositadas en ella, cuestiones que están en la base de los problemas que nos preocupan.

LA DEFINICIÓN

- Es difícil encontrar "la" definición correcta de inteligencia, pues se han encontrado hasta 70 definiciones diferentes. De hecho, puede haber definiciones apropiadas para diferentes propósitos, con enfoques comunes y divergencias considerables.
- Un enfoque amplio para definir la inteligencia es a través de sus componentes necesarios: el uso de la memoria, la comparación, la comprensión, el razonamiento, el pensamiento abstracto, la resolución de problemas y la adaptación a nuevos contextos. La capacidad de aprender de los errores y de adaptarse a nuevas circunstancias se incluye con frecuencia en las definiciones de inteligencia.
- Podemos definirla, de una forma imprecisa o genérica, como la capacidad de un agente para alcanzar objetivos en una amplia gama de circunstancias. Se trata de un enfoque habitual en la IA. Por ejemplo, se aproxima al planteamiento de Russell sobre los seres humanos, según el cual su inteligencia se mide por su **capacidad para alcanzar sus objetivos**.

DIFERENCIAS FUNDAMENTALES

- La inteligencia humana es una entidad compleja, encarnada en un cuerpo biológico, con funciones que van más allá de la programación.
- La inteligencia humana no sólo implica el procesamiento de información, sino también la percepción, la emoción, la conciencia y la capacidad de interactuar con el mundo físico. Aunque las computadoras y los humanos pueden procesar información, la forma en que lo hacen y el tipo de inteligencia que poseen son fundamentalmente diferentes, con diferentes cualidades en términos de **velocidad**, **conectividad**, **actualización**, **escalabilidad**, **consumo de energía**.
- Las computadoras utilizan algoritmos y lenguajes formales para operar, mientras que los humanos utilizan una inteligencia encarnada que interactúa con el mundo, con experiencias subjetivas que son, por su propia naturaleza, "difícilmente" expresables en un argumento formal que requiere precisión. Los ordenadores tienen "una inteligencia mecánica": los ordenadores sólo pueden hacer lo que los humanos les ordenan, son rápidos y precisos, pero no tienen capacidad de **creación**.
- Los animales, humanos y no humanos, necesitan comportamientos inteligentes para seguir vivos, pero los ordenadores no tienen ese deseo, sólo siguen instrucciones.

HITOS DE LA INTELIGENCIA ARTIFICIAL

- Se pueden encontrar las bases de la Inteligencia artificial en estos artículos:
  - (1943), “Behavior, Purpose and Teleology”, **Wiener, Rosenbluth y Bigelow** del MIT, establecieron un marco conceptual para entender el comportamiento dirigido a objetivos, demostrando cómo una máquina puede ajustar su comportamiento en función de la información que recibe sobre su entorno, corrigiendo errores y acercándose a un objetivo predefinido, y sentando las bases de la **cibernética.**
  - (1943), "Un Cálculo Lógico de Ideas Inherentes en Actividad Nerviosa" **McCulloch y Pitts** del MIT, demostraron que cualquier comportamiento que se pueda expresar como reglas "si-entonces" (por ejemplo, "si veo una luz roja, entonces freno") puede ser simulado por una **red neuronal**.

CONCEPTOS BÁSICOS: MÉTODO CIENTÍFICO

- Podemos desarrollar máquinas que muestren un comportamiento inteligente siguiendo una serie de reglas.
  - Las reglas pueden ser deductivas (método deductivo), lo que significa básicamente que, partiendo de premisas verdaderas y siguiendo ciertas reglas, se puede estar seguro de que las conclusiones extraídas son ciertas.
  - También pueden ser inductivas (método inductivo), basadas en generalizaciones a partir de datos, que pueden ser más o menos precisas.
- Las estrategias para lograr la IA pueden pasar por modelar la mente, y algunas se centran en modelar el cerebro.

MÉTODO ABDUCTIVO

- En *El mito de la inteligencia artificial*, de Erik Larson, explica que una condición del comportamiento inteligente es la capacidad de formular conjeturas sensatas: formular hipótesis razonables que traten de explicar los fenómenos que tenemos ante nosotros.
- Larson se basa en la concepción de la abducción de **Charles Peirce**. Peirce afirmaba que, para explicar la razón, **necesitamos añadir la abducción al método deductivo e inductivo**.
- El **razonamiento deductivo** se ocupa de lo que puede concluirse con seguridad y validez a partir de un conjunto de premisas conocidas. La deducción no nos aporta nuevos conocimientos. Lo que sí puede decirnos es que, si las premisas son verdaderas y un argumento es formalmente válido, entonces la conclusión debe ser verdadera. **La deducción se basa en certezas**.
- El **razonamiento inductivo** trata de extrapolar la experiencia pasada al futuro. La inducción puede aportarnos nuevos conocimientos y permitirnos confiar en ciertas afirmaciones, aunque nunca podremos conocer sus afirmaciones con el grado de certeza que permite la deducción. Siempre podemos encontrar un caso observable incompatible con la norma: un cisne negro.

- Peirce argumentaba que había algo muy básico que faltaba en cualquier explicación basada simplemente en la deducción y la inducción. Por sí solas, estas dos formas de razonamiento no pueden interpretar la experiencia.
  - La deducción por sí sola no nos dice nada sobre el mundo. Necesita la inducción para verificar sus premisas. La inducción busca regularidades y patrones en la experiencia, pero **¿cuáles van a ser interesantes?** Ante una multitud de experiencias del mundo, ¿cómo sabemos por dónde empezar a buscar la explicación más probable?
- Necesitamos "sentido común" y podemos confiar en corazonadas o conjeturas razonables para "imaginar" los patrones que se esconden tras la observación empírica. El detective astuto prestará atención a los detalles que, juntos, forman la explicación que desvela el enigma y tiene buen ojo para saber **qué detalles son relevantes**. De hecho, la capacidad de deducción de Sherlock Holmes era mucho más que mera deducción y no habría sido nada de no ser por su capacidad de observación, que incluía no sólo acumular toneladas de datos, sino **dar sentido a los datos**.

MÉTODO ABDUCTIVO

THEORY LANDEN

MÉTODO ABDUCTIVO

- También el razonamiento abductivo se explica como inferencia a la mejor explicación. Es decir, identificar entre todos los datos disponibles de un fenómeno no explicado, la explicación más razonable o plausible.
- Larson sostiene que los casos de éxito de la IA implican en realidad la aportación de los humanos que diseñan el sistema basándose en sus inferencias abductivas. El famoso trabajo del equipo de Bletchley, que descifró el código Enigma, consistió en un proceso de eliminación de muchos códigos por considerarlos imposibles porque entrañaban contradicciones, lo que aceleró la tarea de descifrar los mensajes Enigma y facilitó al equipo la búsqueda de la respuesta. También ayudó la arrogancia de los nazis al suponer que su código era indescifrable. También señala que el equipo de DeepMind que trabajaba en AlphaGo programó en él conocimientos previos sobre el juego, inferencias suministradas desde fuera del marco inductivo.

MÉTODO ABDUCTIVO

- Supongamos que una máquina de aprendizaje automático está entrenada para reconocer imágenes de gatos. Se le proporciona un gran conjunto de datos de imágenes, algunas de las cuales son gatos y otras no. La máquina analiza estas imágenes y busca patrones y características comunes en las imágenes de gatos (por ejemplo, la forma de los ojos, las orejas, la cola, etc.). A partir de estos patrones, la máquina genera una "regla general" o modelo que puede usar para identificar gatos en nuevas imágenes. Este es un ejemplo de inferencia inductiva: la máquina va de casos específicos (las imágenes individuales de gatos) a una regla general (el modelo de reconocimiento de gatos).
- Mientras una máquina se entrena con datos para establecer reglas generales que se utilizan para predicciones o toma de decisiones cuando aparecen nuevos datos, los humanos no se limitan a aplicar reglas generales a casos específicos. También somos capaces de utilizar nuestra creatividad para generar nuevas ideas. Es decir, vamos más allá de los datos disponibles para generar nuevas explicaciones. Incluso si no hemos visto nunca a nuestro gato romper un jarrón, podemos imaginar a nuestro gato rompiéndolo e inferir lo que sucedió como hipótesis explicativa del fenómeno observado. El conocimiento observado (el jarrón roto y un gato travieso) y una hipótesis o suposición nuestra, produce una **imagen creativa en nuestra mente nunca observada**.

MÉTODO ABDUCTIVO

La programación de la inteligencia abductiva, es decir, la capacidad de generar nuevas hipótesis o ideas que van más allá de los datos disponibles es un desafío significativo en el campo de la inteligencia artificial.

Aunque se han hecho algunos avances en la programación del razonamiento abductivo, estos esfuerzos aún están lejos de replicar la creatividad y la imaginación que caracterizan al pensamiento humano

TIPOS DE IA

- IA débil: intento de crear programas que demuestren capacidades de inteligencia, sin ser necesariamente "inteligentes". Aprendizaje automático.
- IA fuerte: intento de construir programas que tengan inteligencia en forma de comprensión y/o que sean conscientes.
- Inteligencia Artificial General (IAG): intento de crear programas que demuestren todas las capacidades de inteligencia de los seres humanos.
- Superinteligencia: noción de inteligencia artificial que supera con creces la inteligencia humana y que, por tanto, puede resultarnos difícil o imposible de comprender. Se supone que la IA superinteligente es capaz de superarse a sí misma de forma recursiva y de crear tecnologías cada vez más inteligentes. El "acabose".

SINGULARIDAD

- La singularidad es un término acuñado por John von Neumann en la década de 1950 para describir la aceleración del progreso tecnológico.
- Vernor Vinge popularizó el término en 1993 en su ensayo "The Coming Technological Singularity: How to survive the post-human era". Vinge imaginó una época de “breakthroughs” sobre los que no tenemos control, derivados de la aceleración de las mejoras tecnológicas, ya que el progreso no lo impulsamos nosotros, sino las tecnologías inteligentes que hemos producido.
- Esto creará una "singularidad" en la que nuestras antiguas formas de entender el mundo dejarán de ser aplicables, un horizonte de sucesos más allá del cual no podemos ver actualmente, porque el mundo será tan diferente de nuestro mundo actual que **no podremos imaginar cómo es**.

- En 1948, el matemático Norbert Wiener introduce la idea de "cibernética" en su obra "Cybernetics or control and communication in the animal and the machine". Wiener perseguía una técnica de control que permitiese automatizar muchos comportamientos de las máquinas sin necesidad de que ese control, en muchísimas fases del proceso, requiriera la intervención humana.
  - El control se refiere a la capacidad de un sistema para mantener un estado deseado o alcanzar un objetivo.
  - La comunicación es esencial para el control, ya que el sistema necesita recibir información (retroalimentación) sobre su estado y el entorno.
- Wiener no veía una distinción fundamental entre los sistemas de control biológicos (como el cerebro humano) y los artificiales (como un termostato). Consideraba que ambos podían ser analizados y comprendidos utilizando los mismos principios.

CIBERNÉTICA IMPACTOS NEGATIVOS

- La cibernética de Wiener se centraba en los complejos bucles de retroalimentación interactivos, que utilizaba para analizar muchos problemas y comprender el comportamiento y los procesos inteligentes, tanto en los seres vivos como en las interacciones entre humanos y máquinas.
- En *The Human Use of Human Beings*, Wiener expresó muchos temores sobre cómo la tecnología puede participar en bucles de retroalimentación con efectos nocivos para los humanos, por ejemplo, al desplazar puestos de trabajo y controlar a los humanos.
- Al incluir la idea de **los bucles de retroalimentación,** podemos entender el problema del control de la IA no simplemente como una batalla directa entre los seres humanos y la potencia de cálculo, sino como la necesidad de un conjunto más complejo de preguntas sobre la interacción entre ambos.

IMPACTOS NEGATIVOS DE ARTEFACTOS HUMANOS: INTERNET

- Del mismo modo, **Tim Berners-Lee** inventó la World Wide Web en 1989, pero ha llegado a lamentar profundamente cómo se está utilizando y el impacto que está teniendo. Casi 30 años después de su creación, Berners-Lee empezó a expresar su grave preocupación por el hecho de que la web sólo estuviera cumpliendo parcialmente sus ideales como plataforma abierta que operaría más allá de las fronteras geográficas y culturales y que permitiría a cualquiera colaborar y compartir información, ideas y oportunidades.
- Sus preocupaciones se centran en la pérdida de control sobre los datos personales, la fácil difusión de información errónea en la red, la falta de transparencia y el uso de la red con fines de desinformación política.

RIESGOS Y BENEFICIOS DE LA IA

- En 2015, se publicó en el sitio web del Future of Life Institute una carta abierta titulada "Research priorities for robust and beneficial artificial intelligence", firmada por miles de personas, entre las que se encontraban muchas personalidades como S. Hawking y E. Musk y muchos expertos en el campo de la IA como Stuart Russell o Peter Norvig.
- La carta, que se aparta notablemente de los pronósticos catastrofistas, está vinculada a un documento detallado sobre **las prioridades de investigación para maximizar los beneficios sociales de la IA**. En la carta se afirma: "Los beneficios potenciales son enormes, ya que todo lo que la civilización tiene que ofrecer es producto de la inteligencia humana; no podemos predecir lo que podríamos lograr cuando esta inteligencia se vea magnificada por las herramientas que la IA pueda proporcionar, pero la erradicación de la enfermedad y la pobreza no son inimaginables".

NEURALINK

- Neuralink afirma que está superando los límites de la ingeniería y declara que su primer objetivo es hacer posible la independencia de las personas que sufren parálisis permitiendo la comunicación y la interacción con la tecnología a través de **interfaces implantadas en el cerebro**.
- Llama la atención que a menudo las nuevas tecnologías **se justifiquen en un primer momento por su posible aplicación en medicina**, o que su aplicación en medicina se destaque como ventaja justificativa.
- Estas nuevas tecnologías experimentales a menudo se justifican por referencia al "caso desesperado". Cualquier aplicación de la tecnología relacionado con la medicina puede considerarse urgente o de especial importancia, en comparación con la mayoría de los demás usos. A menudo se plantea la cuestión de si estos usos médicos **son la pendiente resbaladiza hacia una aceptación más amplia de una tecnología que, de otro modo, podría haber recibido una acogida menos entusiasta**.
  - ¿Debería permitirse que un voluntario sano consintiera una intervención de este tipo?
  - ¿Existe algún peligro potencial de "pendiente resbaladiza" por el uso de esta tecnología desde casos médicos graves a un uso más generalizado?

PENDIENTE RESBALADIZA

- Se trata de argumentos basados en el “efecto dominó”. Muchas veces los sucesos se encadenan los unos a los otros y, antes de tomar una decisión, es preciso valorar la cadena de consecuencias que se seguirán de ella. Pero otras veces se utilizan estos argumentos de un modo tramposo, afirmando una conexión entre A y D sin aportar pruebas.
- Quién presente un argumento de este tipo tiene obligación de probar que existe un nexo causal, pues de lo contrario está usando un argumento falaz.
- Si haces A, se seguirá B, de B se seguirá C… D.
- D no es deseable, por lo que no hagas o decidas A.

EL CONSECUENCIALISMO Y LA INFORMÁTICA

Una teoría consecuencialista afirma que lo mejor que se puede hacer, desde el punto de vista moral, es aquello que produzca los mejores resultados generales. Por lo tanto, necesitamos, como mínimo, lo siguiente: una explicación de los resultados a los que hay que aspirar; una explicación de cómo aspirar a esos resultados; alguna forma de medir los resultados para comparar las distintas líneas de actuación; y una explicación de quién o qué importa a la hora de evaluar y aspirar a los resultados.

Por lo general, una explicación básica del utilitarismo puede sostener que el resultado al que hay que aspirar es la felicidad y la ausencia de infelicidad, o el placer y la ausencia de dolor; una variante es que las consecuencias a las que hay que aspirar son el cumplimiento de las preferencias o la satisfacción de los deseos.

Los planteamientos consecuencialistas pueden parecer idóneos para su aplicación a la informática. Una ética consecuencialista podría especificar los objetivos que debe alcanzar el software, y el rendimiento podría evaluarse fácilmente. A menudo, la introducción de nuevas tecnologías se evalúa en función de los beneficios y perjuicios globales.

LA DEONTOLOGÍA Y LA INFORMÁTICA

Los planteamientos deontológicos de la ética se basan en reglas y se centran en los tipos de acción que deben o no deben realizarse, a menudo expresados en el lenguaje del deber y la obligación.

Se trata de intentar incorporar la ética a la IA programándola como un conjunto de reglas. Puede parecer que, si un ordenador puede seguir una serie de reglas, puede comportarse moralmente, igual que un ser humano.

De hecho, como los ordenadores no se van a distraer por la emoción o el cansancio, quizá, una vez que hayamos averiguado con precisión cuáles deben ser las reglas, una IA sea mejor (moralmente) que nosotros.

LA POLÍTICA DE LA IA

- Las funcionalidades de una tecnología —que parece mágica— ocultan, entre otras problemáticas: el posible robo de datos, la vulneración de la privacidad o la concentración del poder en pocas manos (Microsoft-Open AI, Google), algo que aceptamos por la utilidad de esta revolucionaria tecnología. En palabras de Timnit Gebru:
  - Para empezar, se está produciendo un robo masivo de datos. He hablado con muchos artistas y es descorazonador oír lo que cuentan. Imagina que te pasas toda la vida intentando mejorar tus habilidades, y que compartes todo eso en internet, hablas sobre todo el proceso por el cual llegas a perfeccionar tu técnica. O un programador, hablando sobre cómo usa Stack Overflow y colgando videos y artículos sobre ello. Eso es lo que se supone que era internet, un lugar para compartir conocimiento y aprender de otros. Mucha gente lleva años haciendo esto y, de repente, llega una compañía con miles de millones, coge todo eso gratis, y te lo vende diciendo que eso va a hacer tu vida mucho más fácil. En ese sentido, sí creo que va a haber una revolución brutal porque nos están robando a todos de forma masiva [...] El arte no es eso, no es reemplazar a una persona con un generador de imágenes. El arte va de la expresión humana. (Gebru en Fernández Mateo, 2023)

LA SOSTENIBILIDAD DE LA INTELIGENCIA ARTIFICIAL

- The computational power of AI and energy use Data centres are among the world's largest consumers of electricity.
- The carbon footprint of the computational infrastructure of the world is set to greatly exceed that of the aviation industry. Much of this computational power is generated from technology which has been embedded into our lives. The particular demands of processing power of AI contribute to this and are receiving attention. Early research indicates that the carbon footprint of natural language processing (NLP) models is extremely high. This indicates that increases in speed, accuracy, and other improvements to AI may come at a high environmental cost. The case of Timnit Gebru's departure from Google is mentioned in Chap. 1. The paper she had been working on also discussed the energy costs of language processing models.

SOSTENIBILIDAD MATERIALES Y RECURSOS NATURALES

- La sostenibilidad también es un problema en lo que respecta a los seres humanos que contribuyen a la producción de la IA o son explotados por ella. Esto incluye no sólo a los trabajadores que extraen las materias primas necesarias y a los que fabrican y ensamblan los productos, sino también a los trabajadores a distancia que participan en la creación de inteligencia "artificial", como los moderadores de contenidos y los que etiquetan los datos (con consecuencias para su salud mental).
- Rastrear las enrevesadas cadenas de suministro, las complejidades del proceso de producción y los materiales y seres humanos implicados demuestra vívidamente cómo debemos tener en cuenta la complejidad y el alcance global de la IA cuando consideramos todas las cuestiones de impacto humano y moral..
- Los centros de datos se encuentran entre los mayores consumidores de electricidad del mundo. La huella de carbono de la infraestructura informática mundial superará con creces la de la industria aeronáutica. Las primeras investigaciones indican que la huella de carbono de los modelos de procesamiento del lenguaje natural (PLN) es extremadamente alta. Esto indica que los aumentos de velocidad, precisión y otras mejoras de la IA pueden tener un alto coste medioambiental: son los costes energéticos de los modelos de procesamiento del lenguaje.

REFERENCIAS

Fernández Mateo, J. (2021). ¿ Cuarta Revolución Industrial? El reto de la digitalización y sus consecuencias ambientales y antropológicas. Revista Diecisiete.

Fernández-Mateo, J. (2022). Realidad artificial: Un análisis de las potenciales amenazas de la Inteligencia Artificial. VISUAL REVIEW. International Visual Culture Review/Revista Internacional de Cultura Visual, 9(2), 235-247.

Fernández Mateo, J. (2024). Meta-literatura en el Algoriceno: explorando las fronteras de la inteligencia artificial. La Palabra, (48), 1-17.