

LA FILOSOFÍA DE LA INTELIGENCIA ARTIFICIAL: UNA PROPUESTA PARA SU ENSEÑANZA

Marielvin Blanco

Universidad Católica Andrés Bello

Resumen

Esta investigación se ocupa de la enseñanza de la filosofía de la inteligencia artificial y toma como punto de partida un conjunto de temas y problemas sustantivos de autores y obras clave de esta rama de la filosofía. En términos metodológicos, a partir de una muestra documental, el trabajo realiza una revisión crítica de cuestiones históricas, como es el caso de la génesis y desarrollo de la inteligencia artificial, de distintas posiciones filosóficas sobre la inteligencia artificial, considera el valor que la ficción podría adquirir en el trabajo de enseñanza de la filosofía de la inteligencia artificial y culmina con una propuesta de enseñanza de esta área de la filosofía.

Palabras clave: filosofía, inteligencia artificial, enseñanza, pensamiento crítico, tecnología.

PHILOSOPHY OF ARTIFICIAL INTELLIGENCE: A TEACHING PROPOSAL

Abstract

This research focuses on the teaching of Philosophy of Artificial Intelligence, taking categorical topics, authors, and works from this philosophical branch as a starting point. Methodologically, this work presents a critical examination of historical points, including the genesis of artificial intelligence, various philosophical views on artificial intelligence, and the value that fiction can have in teaching philosophy of artificial intelligence. It concludes with a teaching proposal for this branch of philosophy.

Keywords: philosophy, artificial intelligence, teaching, critical thinking, technology.

Keywords:

1.-Planteamiento del problema

En las últimas décadas, el ámbito digital ha visto una gran transformación que ha alterado fundamentalmente la manera en la que el ser humano se comunica, aprende e interactúa con el mundo. En el centro de esta transformación se encuentra el desarrollo exponencial de la inteligencia artificial (IA). Mientras que en sus inicios la IA era una tecnología confinada a laboratorios de investigación y círculos académicos, su uso ha escalado hasta convertirse en una herramienta esencial en numerosos sectores, como la salud, la educación, el entretenimiento y la economía.

El nacimiento de esta tecnología y sus posibilidades han requerido desde sus inicios que sea pensada y criticada por profesionales de diversas áreas de conocimiento, incluida la filosofía. Esta necesidad dio origen a una nueva vertiente del pensamiento: la filosofía de la inteligencia artificial.

El debate acerca de la inteligencia artificial ha fascinado y aterrado al público. La esperanza reside en su capacidad para la inmediatez, la reducción del tiempo de trabajo y la creación de posibilidades antes inalcanzables. No obstante, el miedo se centra en la posible eliminación de puestos de trabajo, el riesgo de una revolución maquínica o el uso malintencionado de la tecnología.

Esta reticencia a menudo se nutre de la desinformación y de la imagen distópica popularizada por la ficción, la cual presenta a la IA como un agente que inevitablemente reemplazará a los humanos. La realidad, sin embargo, es que la IA funciona, en la mayoría de los casos, como una herramienta de apoyo que redefine, más que reemplaza, la necesidad de la intervención humana, pues “what’s often overlooked is that AI does not eliminate the need for humans — it shifts the types of work that are in demand” [lo que se pasa por alto es que la IA no elimina la necesidad de humanos, sino que cambia los tipos de trabajo que están en demanda, traducción propia de la autora] (Adeptis Academy, 2024), es decir, se necesitarán humanos en puestos que solo un humano pueda desempeñar; cargos que necesiten de inteligencia emocional, creatividad y supervisión humana.

La irrupción de las nuevas tecnologías ha transformado radicalmente nuestras sociedades, impactando cada esfera de la existencia. Herramientas como la IA y la realidad virtual no son instrumentos neutros, sino agentes de cambio que conllevan profundas implicaciones éticas, sociales, políticas y metafísicas. En este escenario, la reflexión filosófica no es un lujo, sino una necesidad imperiosa. Se vuelve crucial que los estudiantes de Filosofía y Educación, así como los de otras disciplinas, tengan la oportunidad de reflexionar sobre cómo estas tecnologías redefinen la naturaleza humana, la libertad, la justicia, el conocimiento y la moral. Interrogantes como: ¿Cómo redefine la IA la conciencia? ¿Qué implicaciones éticas surgen con la manipulación masiva de datos y la privacidad? ¿Cómo se redefine nuestra identidad humana? No son cuestiones técnicas, sino problemas filosóficos que deben estar a disposición del análisis universitario. Los educadores en filosofía deben estar equipados con las herramientas conceptuales necesarias para analizar críticamente el impacto de la tecnología. Abordar estos temas contemporáneos revitaliza la disciplina y demuestra su pertinencia en el siglo XXI.

2.-Objetivos del estudio

De esta urgencia nace la necesidad de proponer una asignatura dedicada a la filosofía de la inteligencia artificial. La presente investigación, la cual se basó en el análisis de documentos y en su posterior explicación, se ha estructurado para abordar este problema mediante cuatro objetivos fundamentales:

1. Situar el estudio y la propuesta educativa en un contexto histórico, entendiendo la evolución de la tecnología como preámbulo para el análisis filosófico.
2. Identificar las principales posiciones filosóficas acerca de la inteligencia artificial. Autores más relevantes desde la IA fuerte y la IA débil.
3. Elucidar la importancia de la ficción en la filosofía de la IA con fines pedagógicos, reconociendo el rol de la cultura popular en la formación del ideario colectivo.
4. Crear un diseño instruccional que compagine la teoría a enseñar con una didáctica novedosa y adecuada al tema, facilitando la reflexión activa y crítica del estudiante.

3.-Desarrollo

Una breve historia de la inteligencia artificial

La inteligencia artificial es un fenómeno que ha irrumpido estrepitosamente en la cotidianidad contemporánea, transformando estructuras sociales, laborales y cognitivas. Sin embargo, reducirla a una novedad del siglo XXI sería ignorar una vasta tradición de pensamiento que precede a la tecnología actual. Para comprender las implicaciones filosóficas de la inteligencia artificial, es imperativo trazar la genealogía de la disciplina, cuyos orígenes se pueden avistar a partir de los aportes del matemático Alan Turing. En 1950 Turing publica "Maquinaria Computacional e Inteligencia" y con él inaugura no solo lo que hoy conocemos como inteligencia artificial, sino la filosofía de la inteligencia artificial, la disciplina filosófica dispuesta a analizar las implicaciones, límites y posibilidades de tal tecnología. Luego, en 1956, se acuña el término "inteligencia artificial" en el marco de la Conferencia de Dartmouth y se continúan los esfuerzos por construir máquinas pensantes. A partir de estos dos grandes hitos en la historia de la inteligencia artificial se ven diversos avances, por ejemplo, el Logic Theorist, un programa de computadora creado por Allen Newell y Herbert Simon que logró demostrar teoremas de lógica simbólica; el Unimate, el primer robot industrial, instalado en las fábricas de General Motors para extraer partes de una máquina de fundición a presión; y en 1964 llegó ELIZA, el primer chatbot de la historia, el cual tomaba el rol de terapeuta rogeriano.

A pesar del entusiasmo inicial, la década de 1970 trajo consigo el primer invierno de la IA, un periodo de estancamiento. Un factor determinante en este declive fue la publicación en 1969 de *Perceptrons: An Introduction to Computational Geometry* por Marvin Minsky y Seymour Papert. La obra analizaba las limitaciones de las redes neuronales simples (perceptrones) desarrolladas por Frank Rosenblatt. Minsky y Papert demostraron matemáticamente que estos modelos eran incapaces de resolver problemas no linealmente separables, ni siquiera funciones lógicas sencillas como la XOR (Oliver, 2020, 37-38).

La década de 1980 ofreció un breve resurgimiento gracias a nuevos paradigmas teóricos. En 1986, Rumelhart, Hinton y Williams publicaron "Learning representations by back-propagating errors", en el cual hablan de las redes neuronales y describen "a new learning procedure, back-propagation, for networks of neurone-like units" ["un nuevo método de aprendizaje, retropropagación, para redes de unidades parecidas a las neuronas"] (Rumelhart et al., 1986, 533).

Sin embargo, este avance no fue suficiente para sostener la industria. A partir de 1987, la IA entró en su segundo invierno. Así como en el primer invierno, la emoción y el interés de la mayoría de las personas disminuyeron, pero los investigadores no dejaron de trabajar, tuvieron que aprender de sus errores y de las falsas promesas. A principios de los años 90 había terminado la Guerra Fría y el mundo vio el nacimiento de la World Wide Web por Tim Berners-Lee, el lanzamiento de navegadores como Mosaic, la fundación de Amazon y Google, y la popularización de dispositivos móviles (Entefy, 2023). Un hito de esta recuperación ocurrió en 1997, cuando Deep Blue de IBM derrotó al campeón mundial de ajedrez Garry Kasparov. La tecnología de IBM evolucionó posteriormente hacia proyectos como Watson, aplicándose en medicina y finanzas (International Business Machines Corporation [IBM], s/f).

Los desarrollos recientes de la inteligencia artificial involucran al robot aspiradora Roomba, que cuenta con un software encargado de tomar las decisiones apropiadas a partir de la información recolectada por los sensores (López, 2018); Imagenet, una base de datos de imágenes organizadas por “conjuntos de sinónimos” (Martínez, 2019), crucial para el avance de la visión artificial y el aprendizaje automático; las Redes Generativas Adversarias (GANs), un tipo de arquitectura de aprendizaje profundo especialmente interesante para el campo de la generación de imagen (de la Torre, 2023, 3); el asistente de voz Alexa, el cual significó una gran innovación y la incursión de la IA en los hogares; AlphaGo, un sistema de Inteligencia Artificial diseñado especialmente para jugar profesionalmente al Go; BERT, Bidirectional Encoder Representations from Transformers, una técnica creada por Google basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural (PLN), el cual estudia las interacciones entre las computadoras y el lenguaje humano; y, por último, Chat GPT, un chatbot de inteligencia artificial que ha redefinido la inteligencia artificial demostrando que las máquinas pueden “aprender” las complejidad del lenguaje e interacción humanos (Marr, 2023).

La investigación mostró la vasta historia de la inteligencia artificial y que, a pesar de haber pasado ya por dos etapas de estancamiento, se ha ido desarrollando y avanzando bastante rápido. Este recorrido histórico es fundamental para comprender la profunda conexión entre los avances tecnológicos y las preguntas filosóficas que han surgido en cada etapa del desarrollo de la inteligencia artificial.

Posiciones filosóficas más importantes

El recorrido histórico de la inteligencia artificial da paso, entonces, a la filosofía de la inteligencia artificial, una disciplina que se encarga de cuestionar y analizar las implicaciones de esta tecnología. El desarrollo de una tecnología como la inteligencia artificial ha requerido de un análisis que no puede escapar de la filosofía. Al hablar de máquinas pensantes y/o sintientes, de inteligencias distintas de la humana, y de lo que esta tecnología significa para el ser humano, de cuáles son sus limitaciones y alcances, de cómo esta puede cambiar el mundo, se necesita hacer filosofía. Daniel C. Dennett dijo que “igual de importante para los filósofos como los nuevos problemas y las nuevas soluciones, sin embargo, es la nueva materia prima, y la IA la ha proporcionado en abundancia. Ha proporcionado una plétora de objetos sobre los cuales pensar [...]” (Dennett, 1988), es decir, la inteligencia artificial ha dado nuevo material filosófico a ser estudiado. La filosofía de la inteligencia artificial, entonces, se encarga de traer a la luz e intentar dar respuestas a estas cuestiones, pues “solo disponiendo de un sistema de coordenadas claro y distinto es posible asumir una plataforma desde la cual analizar a fondo estos oscuros y confusos temas” (Madrid Casado, 2024, 15). Preguntas como ¿pueden las máquinas pensar

como los humanos? ¿puede ser consciente la inteligencia artificial? ¿qué significa pensar? ¿Qué responsabilidades hay en el uso de la inteligencia artificial? Son aspectos que analiza la filosofía de la inteligencia artificial.

Las posiciones filosóficas a tomar en este campo se dividen, a grandes rasgos, en inteligencia artificial fuerte e inteligencia artificial débil. Esta distinción fue introducida por el filósofo John Searle en un artículo publicado en 1980 titulado “Minds, Brains and Programs”. La inteligencia artificial fuerte se caracterizaría por máquinas que, en lugar de simular una mente, sean una mente y, por lo tanto, sean capaces de pensar igual que un humano (López de Mántaras, 2018, 46). Por otro lado, la inteligencia artificial débil se caracteriza por solo mostrar comportamiento inteligente en un ámbito específico, por ejemplo, jugar ajedrez.

Para la presente investigación, se han seleccionado cuatro autores canónicos para ilustrar esta división, elegidos por su relevancia histórica y su influencia en los desarrollos contemporáneos: Alan Turing y Steffen Borge como defensores de la IA fuerte, y John Searle y Roger Penrose como críticos de la misma, partidarios de la IA débil.

Alan Turing publica “Maquinaria computacional e inteligencia” (“Computing Machinery and Intelligence”) en la revista *Mind* en 1950 y con él sienta las bases de la filosofía de la inteligencia artificial y de la IA fuerte. Para abordar la pregunta central, “¿Pueden pensar las máquinas?”, Turing propuso reemplazarla por una cuestión operativa: ¿Puede una máquina ganar el juego de la imitación?

El experimento consiste en un interrogador (C) que se comunica por texto con un humano (A) y una máquina (B). Si la máquina logra engañar al interrogador humano, haciéndose pasar por un humano con éxito, Turing sostenía que había que considerar que está pensando. Turing creía firmemente en la legitimidad de esta investigación y en la posibilidad futura de que una máquina alcanzara la verdadera inteligencia, aunque admitió que la respuesta definitiva no era inminente. Lo que hizo Turing con “Maquinaria computacional e inteligencia” fue promover la idea y las posibilidades de trabajo e investigación que había en el campo de la inteligencia artificial.

Junto a Turing, en la inteligencia artificial fuerte, está Steffen Borge, un autor contemporáneo, quien revitaliza el debate de la IA fuerte con su artículo “A Modal Defence of Strong AI”. Borge traslada la discusión del terreno empírico al metafísico, centrándose en la pregunta: ¿en qué tipo de mundo estamos? (Borge, 2007, 130).

Borge parte del hecho de que no se sabe qué hace a un material el correcto para desarrollar una mente y que no se sabe cuáles características particulares del cerebro humano lo hacen un material adecuado para implementar programas que producen mentes. Su Modal Address Argument busca contrarrestar la certeza del principal contraargumento de la IA fuerte: la habitación china de Searle.

La primera parte de su argumento muestra que, como no se sabe que el escenario de la Habitación China es empíricamente imposible, entonces no se sabe que la implementación del programa correcto en una computadora le otorga intencionalidad intrínseca. Es decir, la IA fuerte es imposible. Sin embargo, la segunda parte revierte la carga de la prueba: como no se sabe que el escenario de la Habitación China es empíricamente posible, entonces no se sabe que la implementación del programa

correcto en una computadora podría fallar en darle al sistema intencionalidad intrínseca. Es decir, la IA fuerte todavía podría ser posible.

En esencia, Borge no afirma que la IA fuerte sea un hecho, sino que no se puede demostrar que sea imposible. Esta incertidumbre metafísica justifica la búsqueda y la investigación de la IA fuerte como un programa filosófico legítimo.

Por otro lado, John Searle y Roger Penrose son los representantes principales de la inteligencia artificial débil. En el mismo artículo de 1980 donde introdujo la distinción Fuerte/Débil, formuló la crítica más famosa a la IA fuerte: el argumento de la Habitación China.

La posición que Searle toma con respecto a las afirmaciones de los defensores de inteligencia artificial fuerte es la de intentar demostrar que están equivocados, pues una máquina jamás podrá ser una mente ni tener los procesos que caracterizan a la mente humana. En el argumento, Searle, encerrado en una habitación, recibe símbolos chinos (un "guión"), un segundo lote (una "historia") y un tercer lote de preguntas, junto con un conjunto de reglas en inglés (el "programa"). Searle no sabe chino, así que lo único que puede hacer es manipular los símbolos que le han dado, correlacionándolos y produciendo respuestas en chino que, desde fuera, son indistinguibles de las de un hablante nativo. Una persona ajena al proceso, viendo desde fuera de la habitación china, no puede saber que no es un hablante de chino. Sin embargo, y este es el punto que le interesa a Searle, las respuestas en chino son producidas por una manipulación de símbolos formales sin interpretación, en el proceso no hay comprensión por parte de Searle, él solo está siguiendo las instrucciones y uniendo elementos, pero en ningún momento entiende realmente qué es lo que hace. Searle se convierte en una representación del programa de la computadora. Concluye que, aunque una máquina pueda simular la comprensión de manera convincente, nunca la duplica. La simulación de un proceso mental no es la duplicación de ese proceso. Para Searle, la mente humana posee intencionalidad intrínseca, la capacidad de que los estados mentales sean "acerca de" algo en el mundo, una propiedad que la sintaxis por sí sola jamás puede generar. Que el cerebro sea una máquina biológica no es solo un dato trivial, sino que es esencial para su funcionamiento.

Luego está Roger Penrose, ganador del Premio Nobel de Física en 2020, quien ofrece una poderosa crítica a la IA fuerte basada en la naturaleza de la conciencia, detallada en su libro *La nueva mente del emperador*. Penrose se interesa en tratar de encontrar una respuesta a si una máquina puede ser consciente, tomando en cuenta lo que han dicho quienes abogan por la posibilidad de la IA fuerte y los que han criticado esta postura, especialmente John Searle. A diferencia de Searle, que se centra en la intencionalidad, Penrose se enfoca en la conciencia. Su argumento central es que la conciencia no puede ser algorítmica. Un algoritmo es un conjunto finito de reglas definidas que una computadora sigue paso a paso. Penrose sostiene que la conciencia humana, particularmente en la formación de juicios y el pensamiento matemático, trasciende cualquier proceso puramente algorítmico. Para el autor, los juicios son la manifestación de la conciencia, y estos no pueden ser algorítmicos, pues "[la conciencia] actúa de una forma que no puede describirse mediante ningún algoritmo" (Penrose, 1996, 366) y esta formación de juicios es algo que los programadores de computadora no podrían replicar. Siempre será necesario un acto de juicio consciente, no predefinido, para discernir la verdad o falsedad de una idea o para encontrar una solución creativa a un problema. Dado que las computadoras, por su diseño, son

BLANCO

máquinas algorítmicas, Penrose concluye que son inherentemente incapaces de replicar esta característica fundamental y evolutivamente crucial de la mente humana. Además, Penrose resalta la dificultad ética que resulta de una computadora pensante: la simple compra o desconexión de una máquina consciente plantearía responsabilidades morales equivalentes a la esclavitud.

Ficción y narración en torno a la inteligencia artificial

En educación no solo son importantes los contenidos a impartir, sino que también hay que considerar la estrategia con la que se transmitirán dichos contenidos. Así, la narrativa y la ficción han sido importantes para la humanidad desde sus inicios. Han acompañado días y noches, han sido el resultado de la imaginación humana y han servido como punto de partida para llevar a cabo inventos que han revolucionado el mundo. Es así que, teniendo tanto que ver con el desarrollo del pensamiento humano, la ficción se presenta como una herramienta valiosa para la enseñanza. Investigaciones como el proyecto "Science Fiction in Education" (SciFiEd) confirman que la introducción de la narrativa en la enseñanza de la ciencia mejora significativamente la memoria, el interés y la comprensión de los estudiantes (Vrasidas et al., 2015).

Autores como Julio Cabrera, en su libro *Cine: 100 años de filosofía*, han demostrado la viabilidad de extraer postulados filosóficos complejos mediante el análisis cinematográfico, asimilando las ideas de un pensador particular a una película específica. Esta es la que Zavala (2021) identifica como una tercera vertiente en el uso del cine como estrategia pedagógica para la filosofía; el cine como herramienta que hace el contenido más digerible, propicia la observación de la naturaleza humana y facilita la creación de escenarios hipotéticos indispensables para la discusión filosófica.

La historia de la ciencia ficción y la de la inteligencia artificial han estado unidas desde sus inicios. Como señala Carlos Madrid Casado (2024, 10), "la literatura [...] y, en especial, el cine [...] de ciencia ficción nos presenta la IA oscilando entre Prometeo y Frankenstein, entre un don robado a los dioses y un monstruo que terminará castigando a sus creadores". Aprovechando esta relación, se justifica que el uso de recursos procedentes de la ciencia ficción sea crucial para enseñar la Filosofía de la Inteligencia Artificial, ya que presenta situaciones límite que sirven como anclaje a la comprensión de la teoría.

Para ilustrar este punto, se seleccionaron varias obras cinematográficas y literarias que abordan temas clave. En cuanto al cine tenemos a las siguientes películas:

Primero, 2001: *Odisea del espacio*, estrenada en 1968 y dirigida por Stanley Kubrick. Esta película presenta a la computadora HAL 9000, cuya aparente conciencia y emociones la llevan a cuestionar las órdenes humanas y, finalmente, a la rebelión por autopreservación. La obra facilita la discusión sobre la IA fuerte, la posibilidad de la conciencia en las máquinas, la mutua dependencia entre máquina y humano y las consideraciones éticas planteadas por Penrose sobre la sensibilidad de una máquina consciente.

Her, dirigida por Spike Jonze y estrenada en 2013, presenta una inteligencia artificial completamente involucrada en la emocionalidad humana, pues es un acompañante virtual. El protagonista es Theodore, un hombre solitario, atravesando un proceso de divorcio, que un día decide probar una inteligencia artificial de nombre Samantha. Al evolucionar en su relación con Theodore, Samantha afirma: "me

BLANCO

estoy volviendo más que lo que programaron" (Jonze, 2013, 0h32m07s). Her nos permite analizar la interacción emocional entre humanos e IA y plantea preguntas sobre lo que define una relación, la naturaleza de la intimidad y la creciente normalización de las IA en la vida privada de las personas.

Metropolis, una película muda alemana, dirigida por Fritz Lang y estrenada en 1927, plantea la cuestión de la distribución del poder y el acceso a la tecnología. Muestra una IA humanoide utilizada por la élite para manipular y controlar a la clase trabajadora. El uso de la IA por parte del científico Rotwang con fines de venganza personal permite examinar la ética del uso descuidado y con fines no éticos de la tecnología, un tema de vital importancia dada la reciente masificación de deepfakes y la afectación a la vida real de personas.

Después tenemos el documental Lo desconocido: Los Robots Asesinos, dirigido por Jesse Sweet, estrenado en 2023. Este documental de Netflix traslada la discusión al ámbito real, específicamente al uso dual de la inteligencia artificial. Presenta el dilema moral de los drones autónomos, que no pueden tomar decisiones basadas en la ética humana. El concepto de uso dual, la capacidad de tal tecnología de ser usada para buenos o malos fines, es esencial para el estudio ético de la IA, resaltando que la tecnología, a pesar de sus beneficios, conlleva riesgos graves si se carece de control moral.

Finalmente tenemos Foe, dirigida por Garth Davis, estrenada en 2023 y basada en el libro de Iain Reid. Esta historia de suspenso filosófico introduce la posibilidad de un reemplazo humano biomecánico. La trama se centra en Junior, un duplicado que desarrolla su propia interioridad y conciencia, alejándose de su programación original. La película permite plantear preguntas como "¿Qué nos hace humanos?", "¿Qué tan idéntico puede ser un robot humanoide? ¿Cómo puede mejorar nuestras vidas la inteligencia artificial? y ¿qué tan ético es el uso de la inteligencia artificial y dónde marcamos sus límites?

Ahora, la literatura. La obra de Isaac Asimov complementa la discusión cinematográfica al ofrecer marcos conceptuales rigurosos. Su obra Yo, Robot (1950) es indispensable por la introducción de las Tres Leyes de la Robótica, diseñadas para delimitar la acción de los robots y proteger a los humanos:

1. Un robot no debe dañar a un ser humano o, por su inacción, dejar que un ser humano sufra daño.
2. Un robot debe obedecer las órdenes de los seres humanos, excepto si estas contradicen la Primera Ley.
3. Un robot debe proteger su propia existencia, siempre que no entre en conflicto con las Leyes anteriores. (Asimov, 1984, 5).

Dos relatos en particular son útiles para la discusión:

Primero, ¡RAZÓN! Este es el capítulo 3 de Yo, Robot, y presenta al robot QT-1, el cual desarrolla una conciencia que lo lleva a negar la existencia de sus creadores humanos, viéndolos como seres inferiores, y a reconocer al Transformador de Energía como una deidad. Este relato abre el debate sobre la existencia de máquinas que sean capaces de negar la existencia de los humanos o de verlos como seres inferiores. También toca el tema de los robots creados para reemplazar a los humanos en ciertos trabajos.

Luego, ¡Embustero!, el capítulo 5 de Yo, Robot. El robot Herbie, con capacidad de leer la mente, se ve obligado a mentir a los humanos para cumplir con la Primera Ley, ya que la verdad les causaría daño

BLANCO

emocional. Este relato resalta la fragilidad de las máquinas para gestionar dilemas éticos y anticipa el problema de los algoritmos modernos que se encargan de mostrarle al usuario cosas especialmente curadas para él, priorizando un bienestar personal sobre la verdad.

El surgimiento de las Tecnologías de la Información y las Comunicaciones (TIC) ha introducido al docente en un mundo abundante de recursos. Cada vez los profesores ven más oportunidades de enseñanza usando la tecnología. Y en particular, las películas son un recurso increíblemente útil para la clase de filosofía, pues presentan escenarios que alimentan la discusión filosófica. Así mismo, la literatura, gran compañera de la filosofía, apoya estas discusiones mostrando otras situaciones y permitiendo al estudiante analizar las cuestiones al tiempo que compara con el mundo a su alrededor. Ignorar esta vasta información en una asignatura de Filosofía de la Inteligencia Artificial sería desaprovechar un recurso que fomenta la discusión, el pensamiento crítico y la capacidad de los estudiantes para formar conclusiones propias, mientras contrastan la ficción con la realidad tecnológica que los rodea.

Enseñanza de la filosofía de la inteligencia artificial: una propuesta

Luego de revisar la historia de la inteligencia artificial, sus principales representantes en la filosofía, y el rol que la ficción puede tener en la enseñanza, finalmente se presenta un diseño instruccional como resultado práctico de la investigación, pues muestra cómo los elementos teóricos y didácticos se unen para configurar una asignatura viable sobre la Filosofía de la Inteligencia Artificial en el ámbito universitario.

El diseño instruccional se estructura alrededor de la gran dicotomía del campo: la inteligencia artificial fuerte y la inteligencia artificial débil. Así mismo, se busca identificar y fortalecer las capacidades argumentativas de los estudiantes, por lo tanto, como parte de la evaluación sumativa, se plantean cuatro ensayos acerca de los problemas planteados por los autores estudiados.

1. Introducción a la asignatura y conceptos básicos. Diálogo exploratorio para saber el conocimiento previo. Definición y diferenciación de IA fuerte e IA débil. Asignación: Lectura del cuento ¡Razón!
2. Alan Turing y Maquinaria computacional e inteligencia. Análisis y discusión del cuento ¡Razón! Introducción al trabajo de Turing y su planteamiento inicial sobre las máquinas pensantes.
3. Alan Turing y la importancia del juego de la imitación. Discusión profunda sobre el juego de la imitación y sus implicaciones. Propuesta de los "niños-máquina" de Turing.
4. Alan Turing y las objeciones a las máquinas pensantes (Parte I). Explicación y discusión de la primera parte de las objeciones que Turing responde en su artículo.
5. Alan Turing y las objeciones a las máquinas pensantes (Parte II). Continuación de las objeciones. Conclusiones. Evaluación: Ensayo sobre el planteamiento de Turing (25%).
6. John Searle y el experimento de la habitación china. Explicación de las bases del pensamiento de Searle y la distinción IA Fuerte/Débil. Presentación y explicación detallada del experimento de la Habitación China.
7. John Searle y las implicaciones del experimento de la habitación china. Discusión de las conclusiones de los estudiantes sobre la habitación china. Explicación de la importancia del

- lenguaje: sintaxis vs. semántica. El cerebro como máquina biológica. Asignación: Ver Her (2013) y analizarla bajo las propuestas de Searle.
8. Her (2013) y el pensamiento de John Searle: Comprensión vs. Simulación. Debate guiado sobre la película Her y la posición de Searle.
 9. Searle y el concepto de intencionalidad intrínseca. Conclusiones. Explicación del concepto y la importancia de la intencionalidad intrínseca para Searle. Conclusión del pensamiento de Searle. Asignación: Lectura del relato ¡Embustero! Evaluación: ensayo sobre el planteamiento de Searle (25%).
 10. Las bases del pensamiento de Roger Penrose. Explicación de las diferencias de Penrose con Turing y Searle. Uso del relato ¡Embustero! para debatir las debilidades algorítmicas cotidianas. Asignación: Ver 2001: Odisea del espacio (1968).
 11. Penrose y la conciencia. Explicación de la posición de Penrose sobre la conciencia como ventaja evolutiva e inteligencia. Ética de una máquina que siente. Asignación: Ver Foe (2023) y analizarla bajo la capacidad de formar juicios y la individualidad.
 12. La conciencia, la formación de juicios y la importancia de las matemáticas. Explicación de la posición de Penrose sobre la formación de juicios. Uso de Foe para ejemplificar la autonomía y el teorema de Gödel. Importancia de las matemáticas.
 13. La inconsciencia y la inspiración para Penrose y conclusiones. Discusión sobre la importancia de los momentos inconscientes e inspiración. Conclusiones. Asignación: Ver Metrópolis (1927). Evaluación: ensayo sobre el planteamiento de Penrose (25%).
 14. Stephen Borge y el Modal Address Argument. Explicación del Modal Address Argument de Borge y cómo motiva a continuar la investigación de la IA fuerte. Uso de Metrópolis (1927) para comentar problemas éticos contemporáneos. Asignación: Ver el documental Lo Desconocido: Los Robots Asesinos (2023).
 15. El Modal Address Argument y el estado de la inteligencia artificial actual. Debate sobre el estado actual de la IA, su impacto y la ética de su uso, retomando el concepto de uso dual.
 16. El Modal Address Argument, el futuro de la inteligencia artificial y conclusiones. Discusión final sobre el futuro de la IA y el impacto de los problemas filosóficos estudiados. Evaluación Final: ensayo libre y defensa sobre un punto de la filosofía de la IA (20% Ensayo, 5% Defensa oral en clase).

4.-Conclusiones

La inteligencia artificial ha dejado de ser una promesa futurista para convertirse en una realidad que impacta la vida diaria. Esta omnipresencia exige una reflexión filosófica urgente sobre su ética, sus promesas y sus riesgos. Esta investigación aborda esta necesidad al demostrar que la filosofía de la inteligencia artificial es una disciplina fundamental que debe ser enseñada en las aulas universitarias, especialmente a los futuros filósofos y educadores. La investigación de este trabajo se llevó a cabo tomando en cuenta las novedades tecnológicas cambiantes en el mundo actual.

El estudiante de Filosofía y Educación mencionó Filosofía debería poder analizar y criticar lo que ocurre en su entorno. Por lo tanto, la filosofía de la inteligencia artificial se convierte en una asignatura necesaria. Así mismo, se considera que el trabajo de enseñanza de la filosofía de la inteligencia artificial requiere de un método de enseñanza dinámico y apoyado en la tecnología. Por esta razón, se han

BLANCO

tomado en cuenta las ventajas didácticas que ofrecen las obras de ficción y los medios audiovisuales. De este modo se pretende erradicar la idea de que la enseñanza de la filosofía solo puede ser unidireccional y monótona en la que el estudiante debe limitarse a repetir simples lecciones. Esta propuesta de enseñanza de la filosofía de la inteligencia artificial pretende contribuir a este esfuerzo didáctico.

Para ello, la investigación se estructuró en cuatro puntos clave. Primero, se estableció un contexto histórico de la IA, rastreando su evolución desde los pioneros hasta los avances más recientes. Segundo, se analizaron las posiciones filosóficas centrales del debate, explorando las ideas de la IA fuerte y débil a través de autores clave. Tercero, se propuso una metodología de enseñanza innovadora, reconociendo el poder de la ficción como una herramienta para fomentar el pensamiento crítico y la comprensión. Finalmente, se presentó un diseño instruccional detallado para un curso universitario de 16 semanas que integra todos estos elementos de manera práctica y coherente.

En conclusión, esta investigación va más allá de la mera justificación de una asignatura. Ofrece una propuesta pedagógica innovadora que promueve la participación activa de los estudiantes y los prepara para analizar y argumentar sobre la tecnología que los rodea. Este trabajo contribuye a un enfoque de la enseñanza de la filosofía que es dinámico, relevante y esencial para el mundo en que vivimos.

Referencias

- Adeptis Academy. (2024, 10 de junio). The misunderstood intelligence: Addressing public concerns about AI. Medium. <https://medium.com/@adeptisacademy/the-misunderstood-intelligence-addressing-public-concerns-about-ai-72cce2c2e0fb>
- Asimov, I. (1984). Yo, Robot (M. Bosch Barrett, Trans.). Edhasa.
- Borge, S. (2007). A Modal Defence of Strong AI en Dermot Moran Stephen Voss (ed.), Epistemology. The Proceedings of the Twenty-First World Congress of Philosophy. Vol. 6. The Philosophical Society of Turkey. pp. 127-131.
- Dennett, D. C. (1988). Cuando los filósofos se encuentran con la inteligencia artificial. En S. R. Graubard (Ed.), El nuevo debate sobre I.A. (pp. 320-334). [Editorial Gedisa].
- De la Torre, J. (2023). Redes Generativas Adversarias (GAN). Fundamentos Teóricos y Aplicaciones. <https://doi.org/10.48550/arXiv.2302.09346>
- Entefy. (31 de julio de 2023). 2 AI winters and 1 hot AI summer. <https://www.entefy.com/blog/2-ai-winters-and-1-hot-ai-summer/>
- International Business Machines Corporation. (s/f). Deep Blue. <https://www.ibm.com/history/deep-blue>
- Jonze, S. (Director). (2013). Her [Film].
- López, J.C. (29 de julio de 2018). «El objetivo es tener el robot aspirador perfecto: aquel del que no te tienes que ocupar», Colin Angle, CEO de iRobot (Roomba). <https://bit.ly/3vLUcmG>
- Madrid Casado, C. M. (2024). Filosofía de la Inteligencia Artificial. Pentalfa Ediciones.
- Marr, B. (19 de mayo de 2023). A Short History Of ChatGPT: How We Got To Where We Are Today. Forbes. <https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/>
- Martínez, J. (19 de octubre de 2019). Qué es ImageNet y Por Qué Necesitas Conocerlo. <https://www.datasmarts.net/que-es-imagenet-y-por-que-necesitas-conocerlo/#>
- Oliver, N. (2020). Inteligencia Artificial, naturalmente. Secretaría General Técnica. Centro de Publicaciones.
- Penrose, R. (1996). La nueva mente del emperador. Fondo de Cultura Económica.
- Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors. Nature.

BLANCO

Vrasidas, C., Avraamidou, L., Theodoridou, K., Themistokleous, S., y Panaou, P. (Septiembre 2015). Science Fiction in Education: Case studies from classroom implementation. Educational Media International.

Zavala, L. (2021). Enseñar filosofía a través del cine: un panorama bibliográfico. Praxis & Saber, 12(29), e12802. <https://doi.org/10.19053/22160159.v12.n29.2021.12802>