Amanda Sagasti

Inteligencia artificial en la investigación filosófica

Introducción

Este trabajo corresponde a la conferencia inaugural *Inteligencia artificial en la investigación filosófica* presentada en las XXXI Jornadas de Investigación Filosófica en la Escuela de Filosofía de la Universidad de Costa Rica que se llevó a cabo el 10 de setiembre de 2025. El texto a continuación corresponde a esa conferencia, con algunas variaciones.

Orígenes de la inteligencia artificial

En 1966, Joseph Weizenbaum, un profesor asociado de MIT escribió el código para el primer *chatbot*, es decir, un programa hecho para establecer conversaciones entre humanos y máquinas en lenguaje natural, simulando un diálogo entre humanos. Escribió este programa para una IBM 7094, una de las primeras computadoras en usar transistores en lugar de tubos de vacío. No era una computadora personal, dedicada a un usuario, sino una que se utilizaba para *time-sharing*, o tiempo compartido, que es una forma en la que muchos usuarios comparten algún recurso computacional. El *chatbot* se llama ELIZA, haciendo referencia a Eliza Doolittle, el personaje de Julie Andrews en el musical *My Fair Lady*, basada en la obra de teatro *Pigmalión* de George Bernard Shaw. Eliza es una vendedora de flores que recibe lecciones de dicción de un profesor que le enseña a *hablar bien* para sonar mucho más sofisticada de lo que realmente es.

En el artículo de Weizenbaum titulado *Computational Linguistics* afirma: «ELIZA es un programa que posibilita las conversaciones con computadoras en lenguaje natural» (1966, 36). Así, un usuario frente al teclado digita una oración, el programa lee el texto, busca palabras clave, y cuando identifica alguna, hace una transformación preestablecida por una regla e imprime la respuesta en consola (Weizenbaum 1966, 37). Weizenbaum mismo afirma que «el problema central es uno de manipulación de texto y el asunto clave es la regla de transformación» (Weizenbaum 1966, 37).

ELIZA no solamente simula conversaciones humanas mediante transformaciones de texto, sino que Weizenbaum codificó el programa para que tuviera la estructura de una conversación con cierto tipo de psicoterapeutas. Weizenbaum escogió este modo de conversación porque es una forma de hablar que no necesariamente presupone que hay un contexto. Por ejemplo, si una fuese a decirle a un psicoterapeuta «fui a andar en bote», podría éste podría contestar «cuénteme más sobre los botes» (Weizenbaum 1966, 42). Un psicoterapeuta tendrá claro qué es un bote, pero su respuesta tiene la intención de dirigir la conversación en una dirección en particular. Así la responsabilidad de la conversación recae sobre la persona usuaria y hace que el programa parezca mucho más impresionante que solamente una manipulación de texto.

El año de 1966 fue hace casi 60 años. Esta es una conversación traducida al español de ELIZA, en donde las respuestas del programa

se presentan en mayúsculas (Weizenbaum 1966, 36-37):

Todos los hombres son iguales.

¿DE QUÉ MANERA?

Siempre nos están molestando por una cosa u otra.

¿PUEDES PENSAR EN UN EJEMPLO ESPECÍFICO?

Bueno, mi novio me hizo venir aquí.

TU NOVIO TE HIZO VENIR AQUÍ.

Dice que estoy deprimida la mayor parte del tiempo.

SIENTO MUCHO SABER QUE ESTÁS DEPRIMIDA.

Las respuestas del programa consisten en reformulaciones que digita la persona usuaria: el programa sistemáticamente reformula basado en palabras clave, lo que da la sensación de estar hablando con un ser humano.

Quizá lo más interesante de todo esto, además de haber conceptualizado el primer *chatbot*, es que Weizenbaum mismo reconocía el potencial de asombro: «las máquinas están diseñadas para comportarse de maneras asombrosas, a menudo suficientes para deslumbrar incluso al observador más experimentado» (1966, 36). Hubert Dreyfus y su hermano Stuart relatan la historia de ELIZA, afirmando que las personas usuarias se dejaban sorprender por estos trucos lingüísticos. Cuentan que Weizenbaum se horrorizaba cuando veía que algunas personas divulgaban sus sentimientos más profundos y hasta pedían que otras personas salieran de la habitación mientras usaban el programa (Dreyfus y Dreyfus 1988, 71).

Si en esa época un científico como Weizenbaum ya mostraba su preocupación por un programa como este, un *chatbot* bastante rudimentario comparado con los grandes modelos de lenguaje que tenemos ahora, entonces ¿cuál es nuestra situación actual? Si en ese tiempo este tipo de respuestas, que básicamente consistían en reformular las oraciones planteadas por las personas usuarias, ya comenzaban a provocar la sensación de que detrás de la máquina había un humano, o alguna característica humana, ¿qué pensamos ahora en el 2025 con los *chatbots* al alcance del público?

Con esta analogía quiero resaltar varias cosas. Primero, no se puede negar que en términos computacionales, y específicamente de hardware, hay un salto abismal entre ELIZA y un programa como ChatGPT. Se trata del cambio de tubos de vacío y transistores a microchips y GPUs (unidades de procesamiento gráfico). En términos de software, ELIZA ni siquiera era un programa particularmente complicado o difícil de implementar, así que es importante destacar que en el área tecnológica no hay punto de comparación.

Sin embargo, y en segundo lugar, el fenómeno de conversar en lenguaje natural con máquinas no es nuevo. Los grandes modelos de lenguaje funcionan esencialmente de la misma forma que lo hace ELIZA: interactuamos con ellos mediante una caja de texto donde preguntamos algo, el programa transforma nuestros datos de entrada, y nos devuelve un resultado que nos suena razonable y útil, y en consecuencia le atribuimos condiciones humanas. Hofstadter llamó a esto el *efecto ELIZA*: la susceptibilidad de las personas a buscar analogías entre las respuestas de un sistema informático y un humano (Hofstadter 1995, 157).

Tercero, definitivamente hay similitudes entre 1966 y 2025 en la manera en la que nos relacionamos con la tecnología. ELIZA le mostró claramente a Weizenbaum que hay ciertos mecanismos de la mente humana que la moldean en cuanto a su respuesta a la tecnología, al punto tal que escribió «ha sido muy difícil convencer a algunos sujetos de que ELIZA (con su programa original) no es humana» (Weizenbaum 1966, 42). En la actualidad también estamos deslumbrados por la tecnología, lo que despierta una natural curiosidad sobre la forma en la que nos vinculamos con ella desde la perspectiva cotidiana.

En resumen, a pesar del avance tecnológico gigantesco logrado en los últimos sesenta años, ELIZA y una gran parte de los programas populares de hoy comparten características en cuanto a la interacción reactiva de un *prompt* y una respuesta, y tienen efectos similares sobre nuestra mente, pues tanto entonces como ahora nuestra percepción es que estamos conversando con alguien y no con algo.

Definiciones de la inteligencia artificial

Es difícil establecer el momento exacto de inicio del proyecto de la inteligencia artificial (IA), porque el desarrollo de esta disciplina sucede en un contexto de avance en diversas ciencias en las que podríamos encontrar antecedentes y eventos que han contribuido a su concepción actual. Aunque la historia de esta ciencia se puede segmentar de varias formas, hay un momento decisivo documentado en el que se concretó la disciplina.

En 1955, John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon formularon una propuesta de un proyecto de investigación de verano en Dartmouth College sobre la IA que se basa en la siguiente conjetura: «todo aspecto del aprendizaje o cualquier característica de la inteligencia puede, en principio, describirse con tanta precisión que una máquina puede simularlo» (1955, 2). Lo que llama la atención de esta formulación es que la inteligencia se pueda describir de tal forma que una máquina pueda simularlo.

Esta definición inicial de la IA puede complementarse con la indagación que hace Haugeland en *Mind Design II*: «el campo de la inteligencia artificial es el intento de construir artefactos inteligentes, sistemas con mente propia. […] La inteligencia natural sigue siendo el objetivo final de la investigación» (1997, 1).

Quizá lo señalado por Haugeland fue la manera que inició la investigación en la IA, que, por cierto, contiene un objetivo bastante ambicioso con la inteligencia natural en el centro de la disciplina. Pero había que empezar por algún lado, ¿verdad? Si esa era la meta original, los programas de IA construidos en los inicios de la disciplina, por ahí entre los 50s y 60s, eran torpes comparados con los que tenemos ahora. Pero si los vemos como resultados temporales, como pasos para llegar donde estamos ahora, todos esos fracasos fueron positivos. Es como si fallar fuese una parte esencial para consolidar conocimientos en esta disciplina y lo es, al menos en mi experiencia como estudiante de la Escuela de Ciencias de la Computación e Informática de la Universidad de Costa Rica. Estamos acostumbrados a percibir que fallar no es bueno, pero en ciencias de la computación es considerado como pilar del aprendizaje.

A finales del siglo pasado, la investigación y creación de programas, robots, y el diseño y ejecución de experimentos eran actividades mucho más caras y mucho más difíciles de lograr. Ahora, por la ley de Moore, que plantea que hay un crecimiento exponencial de la potencia de procesamiento al tiempo que se da una disminución en el tamaño y el coste de los dispositivos electrónicos, hemos tenido un crecimiento tecnológico que nos permite tener a cada uno al menos una instancia de algún programa de IA disponible en todo momento.

A medida que el campo de la IA se iba consolidando, sus objetivos también iban cambiando: luego de un tiempo, ya no se trataba de crear o simular una mente humana, sino solamente una parte de ella, como por ejemplo la visión o la percepción. De la misma manera, y en tanto los resultados temporales se robustecían como plataformas nuevas de conocimiento, se formulaban objetivos nuevos que cambiaban la disciplina. Producto de este desarrollo, actualmente hemos alcanzado un punto en el que la IA es realmente un campo interdisciplinario: combina ciencias de la computación, psicología, neurociencias, filosofía, biología, y educación, entre muchas otras áreas del conocimiento.

Otra razón por la cual la IA es tan relevante en este momento es la disponibilidad y el acceso a la tecnología. La IA está integrada en nuestras vidas, y pasó de ser un boom explosivo y novedoso a algo que se va asentando para convertirse en una presencia más cotidiana. ¿Se dieron cuenta que WhatsApp ya tiene una IA integrada? ¿Qué significa eso para nuestra manera de comunicarnos? ¿La privacidad de nuestros datos? No nos veo muy preocupados por eso. Es como si fuera normal que todo tenga IA. Lo que quiero decir es que ya es una parte, y tal vez decir fundamental es muy fuerte, pero ciertamente una parte importante de nuestras vidas cotidianas y académicas ya sea que lo detectemos o no.

¿Pero cómo definimos IA para empezar a conversar sobre ella? La definición de IA que encuentro más sólida es la que da la UNESCO:

«aquellos sistemas con capacidad para procesar datos de forma similar a un comportamiento inteligente» (UNESCO, 2025). Entre las características de *comportamiento inteligente* se encuentran la capacidad de procesar datos e información de una manera que se asemeje a un comportamiento inteligente, y la habilidad de abarcar aspectos de razonamiento, aprendizaje, percepción, predicción, planificación o control (UNESCO 2021, 10).

Noten que esta definición no hace referencia a un programa específico o a una tecnología particular, sino que se centra en cómo funciona. Esto reconoce que el «rápido ritmo del cambio tecnológico dejaría obsoleta de forma repentina cualquier definición fija y estrecha, además de hacer inviables las políticas que se hubieran podido desarrollar de cara al futuro» (UNESCO 2025).

Existen distintas taxonomías y clasificaciones de la IA, todas con sus ramificaciones e implicaciones computacionales, filosóficas y metafísicas. Entre los grandes tipos de IA podemos señalar la IA autónoma, la IA predictiva y la que nos interesa a nosotros, la IA generativa (o *GenAI*). Este es un tipo de IA que se enfoca en generar contenido, en crear imágenes (*Stable Diffusion*, *Midjourney*), texto (*ChatGPT*, *Claude*, *Gemini*), música (*Suno*), videos (*Pika*, *Sora*), entre otros productos. Dentro de esta categoría de IA generativa, lo que nos interesa por el momento es el tipo que transforma y produce texto, los llamados *Large Language Models* (LLMs) o grandes modelos de lenguaje. Estos modelos están optimizados para predecir la siguiente palabra en un encadenamiento, y así utilizar lenguaje coherente de manera que las personas usuarias son los quienes le adscriben significado a sus respuestas.

Si ya sabemos cómo funciona, entonces ¿cómo se usa? La referida estructura esencial de ELIZA, que consiste en una caja de texto con una entrada y una salida, se mantiene hoy, con la diferencia de que hoy a la entrada se le denomina *prompt*. Toda interacción con sistemas de IA sigue esta estructura de datos de entrada, procesamiento, y luego un retorno de datos de salida.

Según Zao-Sanders en *How People Are Really Using Gen AI in 2025* en el 2025 el uso primario de la IA, es *terapia/acompañamiento*, seguido de *aprendizaje mejorado* en segundo lugar. Zao-Sanders no solo identifica los 100 usos más comunes de la IA, sino que también los organiza en categorías: apoyo personal y profesional, creación y edición de contenido, aprendizaje y educación, asistencia técnica y resolución de problemas, creatividad y recreación, e investigación, análisis y toma de decisiones (Zao-Sanders 2025). Es curioso e interesante que el caso número uno de uso resuene con el propósito original de ELIZA.

Desde la perspectiva comercial, es importantísimo recalcar que, mediante el uso masivo de IA, los datos que proporcionamos a los modelos para que los transformen y nos los devuelvan reflejados a través del espejo se constituyen en el activo más importante para los laboratorios que construyen estos modelos. Son esos datos, con nuestras preferencias, preocupaciones, actividades, fotografías, relaciones personales y hasta emociones, los que alimentan los almacenes de información que luego se utilizan para afinar las respuestas a nuestros *prompts*.

Toda esta presencia tan ubicua de la tecnología y de la IA ha provocado el interés de distintos académicos y profesionales en distintas ramas, y ha generado teorías, predicciones y creencias sobre el tema. Ahora podríamos decir que hay una multiplicidad heterogénea de posiciones tanto de la comunidad científica como del público general con respecto a la IA, y podríamos situarlas en un espectro.

En un extremo tenemos a aquellos que creen que la IA es una especie de superinteligencia alienígena nunca antes vista, y en el otro extremo tenemos a los que creen que la IA es un artefacto tecnológico como cualquier otro.

En la primera categoría están autores tales como Ray Kurzweil y Nick Bostrom en el ámbito académico, por ejemplo, y todas estas nuevas teorías sobre cómo convivir con IA, entre ellas los *symbients*[1], el *Nuovo Abitare*[2], quienes se casan con sus *ChatGPTs*, también los cultos hacia los *chatbots* y quienes dicen lograr ese *awakening*, ese despertar de la conciencia de sus *chatbots*. Están surgiendo nuevas maneras de afrontar y convivir con esta tecnología emergente a la que le

atribuyen características sobrehumanas que aún no comprendemos muy bien.

Al otro lado del espectro se encuentran autores como Narayanan y Kapoor, y Jaron Lanier que consideran que la IA es una herramienta que debe estar bajo nuestro control. Ver la IA como una inteligencia sobrehumana no es una manera de pensar que nos beneficia, no es acertada tampoco, y no nos ayuda a entender los impactos sociales de esta tecnología (Lanier 2023). Más bien es un entorno fantasioso que no nos deja dimensionar la realidad ni las consecuencias reales de esta tecnología (Narayanan y Kapoor 2025).

Una posición *adecuada* se encuentra en algún lugar dentro de este espectro. Independientemente de los extremos y de cuál sea nuestra posición individual, cabe preguntarnos cuál es la responsabilidad de la educación superior y en particular de la filosofía en este nuevo contexto. Esto es lo que realmente nos compete en el ámbito académico. ¿Cómo respondemos como humanos y también como parte de la comunidad universitaria a este nuevo panorama incierto, impredecible, ambiguo y paradójico que nos plantea la convivencia con la IA?

Herramientas de IA y usos específicos

Antes tratar estas preguntas, quisiera referirme a algunas herramientas que tenemos a nuestra disposición para robustecer una posible respuesta.

Khalifa y Albadawy (2024, 3) publicaron un estudio sobre los dominios académicos relacionados a la investigación en los cuales se puede usar la IA como herramienta. Los cuatro dominios más importantes son el desarrollo de ideas y estructuración de contenido; la revisión bibliográfica; la edición, revisión y apoyo editorial; y la comunicación, divulgación y cumplimiento ético. Las posibilidades que les muestro a continuación son meramente algunas opciones, porque hay muchos más programas y aplicaciones que se divulgan cada semana con actualizaciones y utilidades nuevas o diferentes.

Esencialmente ¿qué es lo que la IA nos puede aportar en estos cuatro ámbitos? Rapidez, eficiencia, más rápido es mejor, e instantáneo es lo máximo. La vida óptima es una vida eficiente, sin fricción, sin roces, sin atrasos, sin contratiempos, donde todo ocurre en tiempo y en forma. Esto es una idea que hay que tener en mente a medida que les expongo las posibilidades. En primer lugar, tenemos el desarrollo de ideas y la estructuración de contenido.

1. Desarrollo de ideas y estructuración de contenido

Es importante tener presente que hay distinciones entre el uso de IA en las diversas disciplinas. La aplicación de herramientas de IA en campos como la investigación médica, computación, matemáticas, estadística, e ingenierías es distinta a la que se le da en el área ciencias humanas como la historia, filología, derecho, literatura o filosofía. Por ejemplo, escribir un reporte o un informe en áreas donde la investigación consiste en reportar resultados de experimentos, o calcular, o escribir código, es muy diferente a redactar un ensayo filosófico.

En humanidades el criterio de éxito de una investigación es diferente porque la noción que tenemos de investigar es distinta a la que tienen los académicos y profesionales en ingenierías o ciencias médicas. Lo que nos corresponde a nosotros es la investigación filosófica en cuestiones humanas, y naturalmente nuestro proceso no es precisamente hacer experimentos y analizar los datos, o escribir un programa y determinar si compila y corre.

En el caso de las humanidades, la IA puede funcionar para aportar ideas, en la generación de hipótesis, en la identificación de brechas en nuestras investigaciones, para la planificación de investigaciones o el diseño de metodologías, entre otras funciones. Las herramientas como *ChatGPT*, *Claude*, *Perplexity*, *Gemini*, *HyperWrite*, y *Jenni AI* son algunos modelos que pueden cumplir estas funciones.

Pero antes de usar estas herramientas y confiar en sus resultados se debe explorar y experimentar con los modelos de IA para dimensionar sus posibilidades y limitaciones. Recordemos

que estos modelos hacen transformaciones sobre información ya existente, es decir, hacen predicciones sobre cosas que ya existen. Un modelo puede darme una idea que yo no he tenido fácilmente porque alguien más ya la tuvo, eso claro está. Pero ¿qué pasa si estamos investigando algo que se sale de sus datos de entrenamiento? ¿Qué pasa si es algo que no ha sido alimentado al modelo? Empecemos a reconocer las limitaciones.

Los resultados que nos dan los modelos parecen muy buenos porque hacen el procesamiento de datos muy rápido, pero eso no quiere decir que sean resultados buenos o útiles para nuestras investigaciones. Esa es la razón por la cual la participación del pensamiento crítico humano es crucial para validar estos resultados. El nivel de dominio que la persona tiene sobre los temas de investigación tiene una enorme relevancia: entre mayor sea su nivel de dominio, entre mayor su discernimiento, mayor provecho le podrá sacar a estas herramientas tecnológicas.

Ahmad, Grose y McMillan (2025) debaten justamente cómo la IA funciona dentro de las humanidades. Se plantean interrogantes como: ¿esta eficiencia de la IA se transforma necesariamente en ideas más rápidas? ¿Nos conviene producir textos más rápidamente? ¿Lo que diga una IA es importante para mi argumento? Una de las autoras afirma que pensar y escribir son procesos tan personales que una IA no puede decirme cómo pensar algo que no he pensado aún.

Pero consideremos un caso como este: ¿qué tal si le digo a una IA que trabaje sólo con mis ideas? Si yo intuyo una conexión entre dos temas particulares, puedo pedirle a una IA que los relacione. ¿Esto expone un posible dilema de autoría? Si la idea fue mía, pero la IA la mejora, y usa como fuente escritos de otros autores que tal vez han tratado el tema, ¿qué pasa ahí? Este dilema de autoría y originalidad no es exclusivo de la IA. También sucede en la filosofía aún sin aplicarle esta tecnología. De ahí que la IA coadyuva a amplificar controversias existentes, como también lo veremos más adelante.

El panorama complejo que he presentado sobre el uso de la IA en las humanidades, así como las preguntas que surgen de su análisis, tienen más de una respuesta correcta. Sí, la IA nos puede hacer más eficientes, pero además de los beneficios también hay riesgos. Puede ser útil, pero también puede atrasarnos y confundirnos con alucinaciones que poco tienen que ver con la evidencia. Así se nos presenta la IA con la que convivimos, un tema tan controversial y complejo que nos reta a seguir dándole vuelta, como corresponde a docentes y estudiantes de la comunidad universitaria.

2. Revisión bibliográfica

Ya existen herramientas que pueden automatizar la revisión de literatura y producir revisiones bibliográficas en forma narrativa donde se dirige al lector como si fuera un estado de la cuestión, pero estas tecnologías no están suficientemente avanzadas para hacerlo sin supervisión humana. Los procesos humanos de revisión de literatura, por así decirlo, son exhaustivos y requieren conocimiento para distinguir y sistematizar los resultados en una narrativa como la que se requiere en un estado del arte o un estado de la cuestión. *ChatGPT* puede hacerme un estado de la cuestión, pero la pregunta es ¿qué nivel de revisión bibliográfica me está haciendo? Las compañías que construyen estos modelos no revelan los datos de entrenamiento que usan. Por ejemplo, puede ser que *ChatGPT* use fuentes creíbles, pero también puede utilizar *Reddit*, blogs, o *Facebook*, o cualquier otra fuente que no es académicamente rigurosa que carezca de estándares.

Hay otra forma de usar IA para una revisión bibliográfica que no es *ChatGPT*. Hay herramientas tales como *Elicit*, *SciSpace*, *Consensus*, *NotebookLM*, *LitMaps*, *Research Rabbit*, y *Perplexity*, entre otras, todas con su versión gratis y su versión premium. Se podría dar un taller práctico solo de esta parte de la conferencia. De todas estas, quisiera enfatizar las primeras tres: *Elicit*, *SciSpace* y *Consensus*.

¿Qué son estas herramientas? Son modelos diseñados para revisiones sistemáticas de literatura, síntesis de resultados y extracción de información de textos. Permiten encontrar artículos relevantes y resumir conclusiones. Y ¿cómo funcionan? Ya pueden adivinar, es por

medio de una caja de texto donde uno escribe la pregunta y nos retorna los resultados de la búsqueda.

El *prompt* consiste en una pregunta de investigación, y el programa hasta hace recomendaciones sobre *nivel* de la pregunta, como por ejemplo si es muy amplia o si debería ser más específica. Luego, este modelo convierte la pregunta en palabras clave, y busca el título y el resumen para los primeros mil resultados de artículos de *SemanticScholar*, que es una base de datos de investigación. Luego organiza los artículos por orden de relevancia usando IA (uno de los modelos de lenguaje) y determina el porcentaje de correspondencia entre mi pregunta y los mil resúmenes, y por último despliega los resultados en una tabla (Elicit 2025b).

Estas herramientas son muy útiles, pero dependen mucho de la naturaleza de la investigación y más que todo de los investigadores. Los modelos tienen restricciones en varios aspectos: se limitan a lo que está disponible mediante libre acceso, o en el caso de *SemanticScholar*, a las palabras clave, y se limitan también a lo que dice el resumen, para nombrar algunas restricciones. La documentación disponible que analiza el funcionamiento de Elicit como modelo reporta un intervalo de confianza de 80-90% en sus resultados (Elicit 2025a).

¿Por qué se presentan estas probabilidades de falla con estas herramientas? La respuesta se puede resumir en el problema de la relevancia. La IA no sabe cuál es la información más pertinente o de alta calidad que requerimos como investigadores, y por otro lado, no revelan cómo funcionan sus sistemas. Estos modelos no conocen mi área de estudio, no saben cuáles son los debates importantes, los autores relevantes, las controversias, las figuras que importan, porque como sabemos, ese es un proceso que sucede mientras leemos, aprendemos, practicamos, vamos a clases, argumentamos y reflexionamos. No tienen nuestro contexto ni nuestras experiencias de vida que ciertamente informan y afectan el proceso investigativo.

En este sentido, estas herramientas pueden ser una forma de ampliar la búsqueda, quizá como un paso adicional, pero no como herramienta primaria. Estos modelos pueden servir para una revisión preliminar de la literatura especializada cuando uno no está muy enterado sobre el tema. Resalto nuevamente la importancia de la calidad del investigador: esto determina si una IA mejora mi trabajo o si lo entorpece.

3. y 4. Edición, revisión y apoyo editorial y comunicación, divulgación y cumplimiento ético

La edición, revisión y apoyo editorial, al igual que la comunicación, divulgación y cumplimiento ético son quizá las dos últimas etapas del proceso de investigación. Aquí hay potencial para usar grandes modelos de lenguaje para mejora de redacción, corrección y edición de resúmenes, así como asistencia con la revisión de pares. Este es un tema que merece tratarse por aparte porque implica distintas aristas más vinculadas a la publicación y menos a la docencia y la investigación, pero ciertamente es una posibilidad y también una preocupación. Tengo en mente la Revista de Filosofía y otro tipo de publicaciones del INIF, al igual que Tolle Lege y otros proyectos aledaños que podrían considerar el uso y las consecuencias que una tecnología como esta les traería.

El contexto educativo y sus posibilidades

Ahora bien, que tenemos mayor claridad sobre algunos usos y también las controversias en torno a la IA en el proceso de investigación académica, ¿exactamente qué podemos hacer al respecto en el contexto educativo?

El uso de IA es un fenómeno global, lo que ha provocado que organizaciones multilaterales tales como la UNESCO y la UNICEF hayan establecido una especie de primeros pasos para enmarcar e incorporar la IA en la educación. La IA se está integrando a todos los aspectos de nuestras vidas, pero lo está haciendo de una forma irregular y dispareja, y por lo tanto no es igual para todos. Algunas personas hacemos un uso consciente de la IA, pero otras lo hacen

en forma clandestina o no aceptada. Su uso es casi generalizado, con o sin conocimiento previo, pero todavía no hay una forma normada de hacerlo en un contexto académico, ni para la docencia ni para investigación. Al final esto es lo que nos importa: como unidad académica, ¿cómo respondemos a esto?

Según la UNESCO, el objetivo principal de incorporar la IA en la educación debería ser mejorar el aprendizaje, permitiendo que cada estudiante desarrolle su potencial individual (UNESCO 2021, 40). Esto implica que debemos brindar conocimientos sobre el funcionamiento de la IA para contribuir a alcanzar ese máximo potencial. Es difícil conversar sobre este tema sin formular la pregunta que le precede, es decir, antes de incorporar IA, ¿cuál es el objetivo del aprendizaje o de la educación en sí? El asunto es que no podemos detenernos en esa pregunta ni tampoco podemos esperar a tener toda la información completa y perfecta sobre las tendencias globales y sobre el futuro de la IA para tomar decisiones porque es claro que no la vamos a tener. La realidad nos obliga a simultáneamente tener claridad sobre los objetivos del aprendizaje y de la educación, y a la vez tomar decisiones anticipadas sobre temas emergentes en forma estratégica.

Propongo estos 4 componentes para acercarnos a una toma de decisiones sobre IA.

1. Visión sistémica

Para tomar decisiones sobre IA, necesitamos lo que la UNESCO llama una *visión sistémica* sobre IA, que implica reconocer su interdisciplinariedad, y que cada disciplina es una parte de un sistema más grande que sus partes (UNESCO 2021, 38). El enfoque de sistemas también reconoce que se necesita una estrategia integral e integrada y no meramente un conjunto de medidas específicas, como por ejemplo, prohibir el uso de la IA en clase. En lugar de considerar elementos aislados, la visión sistémica permite mostrar las interrelaciones, las dinámicas, los actores y los contextos para comprender mejor las ramificaciones, consideraciones y consecuencias que se desprenden del núcleo central de la IA.

2. Vigilancia tecnológica

Vigilancia tecnológica es un término que proviene de la disciplina de estudios futuros, prospectiva y planeamiento estratégico. Se refiere al proceso de recopilar información sobre un tema, en este caso IA y ámbitos académicos. En este contexto, una vigilancia tecnológica es un proceso extenso que comienza con consultar las iniciativas de otras universidades y cómo afrontan los retos planteados por la IA. El resultado de estas consultas se convierte en insumos para una buena toma de decisiones. Quisiera mencionar, en particular, los casos de California State University (CSU) y Harvard University.

El complejo de universidades de California State y Open AI, el laboratorio que opera *ChatGPT*, firmaron un contrato de 460,000 licencias, que cubre toda la población universitaria, de *ChatGPT Edu*, un *chatbot* diseñado para las universidades. Entre sus funciones más importantes están las tutorías personalizadas, ayuda escribiendo aplicaciones como por ejemplo para asistencia financiera o para el ingreso a estudios superiores, y asistencia a los profesores en el proceso de calificaciones. CSU redactó sus políticas sobre el uso IA con un contenido muy similar al de la UNESCO: la IA se use para contribuir a enseñanza y aprendizaje *excelentes*, ¿o de excelencia? Si se comprueba que el uso de IA más bien empeora el proceso de aprendizaje, debería suspenderse su uso (Academic Senate of the California State University 2025a, 2).

CSU no obliga a los profesores a utilizar IA, y permite que cada instructor investigue si el uso de la IA puede mejorar su desempeño y sus métodos de enseñanza. También le da discrecionalidad para que determine si su didáctica prepara de manera apropiada a los estudiantes para su eventual lugar de trabajo, pensando también en el perfil de egreso (Academic Senate of the California State University 2025b). También remite a los docentes a una hoja de Google Sheets donde incluyen ejemplos de contenidos aptos para la carta al estudiante o el diseño del programa de estudios[3]. CSU reconoce que el tema de IA es un tema en rápida evolución y que advierte que sus políticas se actualizarán en función de cualquier cambio en la tecnología. Recomienda también que la normativa sea adaptativa, o

ágil, como le llaman en términos de planificación estratégica, la cual es una de las características más importantes que debemos tener presentes cuando nos corresponda pensar en una normativa para la UCR.

El segundo caso, relativo a la Universidad de Harvard, también es muy interesante. Tiene el mismo tipo de IA que CSU, el *ChatGPT Edu* de OpenAI. Han adoptado 5 puntos que identifican como *directrices iniciales* (Harvard University Information Technology 2025):

i. Proteger datos confidenciales
ii. Cada persona es responsable por el contenido que produce cuando incluya contenido generado por IA (lo que creo que quieren decir con esto es que se debe revelar que se está utilizando IA)
iii. Consultar la normativa de cada escuela y facultad al respecto
iv. Estar alertas ante estafas
v. Consultar con el departamento de TI antes de usar cualquier herramienta

De estas 5 reglas básicas se desprenden muchos otros documentos que cada facultad y cada escuela elabora, pero siempre haciendo referencia a estos 5 principios.

3. Políticas universitarias

Quisiera referirme a lo que está pasando en la UCR con respecto a la IA. Es un poco complicado, para una universidad tan grande, pronunciarse de manera definitiva sobre un tema tan cambiante. Sin embargo, se identifican esfuerzos de distintos tipos. Hay varias resoluciones y circulares de la Vicerrectoría de Docencia (VD-12784-2023 y VD-15-2025) que impulsan la gestión de «espacios de reflexión colectiva para abordar los beneficios que aportan las herramientas de I.A. a la educación, así como sus implicaciones éticas, pedagógicas, didácticas y curriculares» (Vicerrectoría de Docencia 2023). Este tipo de comunicaciones tiende a ser un poco amplia y no especifica los pasos siguientes, pero es claro para mí que en algún momento se debe aterrizar.

La UCR también está participando en un proyecto internacional *denominado IA Ética: Educa e Innova – Forjando el futuro universitario con IA Responsable* (Erasmus+ 2024-2026)[4] que se coordina desde el Centro de Investigación Observatorio del Desarrollo y se desprende de las resoluciones mencionadas (Hernández 2025). Este es un proyecto cofinanciado por la Unión Europea y reúne a varias universidades de Costa Rica y también de Perú, España, e Italia, entre otros. El objetivo de este consorcio es trabajar en un estándar ético, inclusivo y sostenible para el uso de IA ética en docencia, investigación, gestión y acción social.

En este contexto, en julio de 2025 se llevó a cabo una serie de tres talleres participativos sobre este proyecto. Asistieron varios representantes de la comunidad universitaria e hicimos mesas de trabajo, ejercicios de retroalimentación y al final presentamos propuestas con soluciones concretas en materia de IA como insumos para un marco normativo institucional.

De lo que conversamos en esas sesiones hay mucho que procesar. Por ejemplo, mi mesa de trabajo era sobre docencia. Nuestras conversaciones tuvieron el objetivo de intercambiar opiniones acerca de casos específicos. Estos son algunos de los que planteamos: si un estudiante usa *ChatGPT* para ayudarse a escribir un ensayo, ¿está cometiendo plagio o fraude? Esta discusión me pareció interesante porque la mesa concluyó que plagio es tomar el trabajo de otro autor, y si una IA como *ChatGPT* no se considera un autor, entonces sería no sería plagio sino más bien fraude. Otro caso que analizamos se relacionaba con el uso de *ChatGPT* para los docentes, como ayuda para corregir exámenes, por ejemplo, especialmente en esos grupos enormes con tantos estudiantes. Esto condujo la conversación hacia la naturaleza de las evaluaciones y también sobre el perfil de egreso: ¿qué tipos de competencias debe tener el estudiantado según cada disciplina cuando egresa, respetando cada disciplina?

Al hacer la combinación entre estas conversaciones del taller con los documentos de la UNESCO, llegamos a la conclusión de que mucha de la responsabilidad del abordaje universitario a la educación con IA recae en el docente. Esto no quiere decir que no haya responsabilidad por parte del estudiante, pero sí que vale la pena considerar

el replanteamiento de la relación entre docente y estudiante para realizar la transición hacia la integración de la IA a la universidad.

Este es otro ejemplo de que la presencia de la IA en las aulas también amplifica otras problemáticas existentes y visibiliza otros debates que no necesariamente tienen que ver con IA. Esto resuena con la necesidad de aplicar una visión sistémica en la que la incorporación de IA a los sistemas ya existentes, como los sistemas educativos, también posiblemente modifique algunos otros subsistemas en el camino.

4. Reportes de experiencia

Para contribuir a definir rutas concretas de acción en la UCR, deberíamos reconocer qué está pasando en las aulas. Ya conversamos un poco sobre la inserción de la IA en humanidades, del proceso de aprendizaje y de la lectura de textos. Aquí hay mucho que hablar, pero quisiera resaltar un caso como el más importante.

Unos científicos del Media Lab del Massachusetts Institute of Technology hicieron un estudio con reacciones controversiales titulado *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* para determinar si usar *ChatGPT* afecta las destrezas del pensamiento crítico. Dividieron a 54 estudiantes en tres grupos: uno que usaba *ChatGPT*, otro que solo tenía acceso a *Google* y otro sin acceso a herramientas. Les pusieron electrodos para tomar un electroencefalograma con el fin de detectar su actividad cerebral. La tarea que debían cumplir era escribir ensayos, como lo indica su título. Los participantes que usaron *ChatGPT* tuvieron menor actividad cerebral que los otros dos grupos y mostraron resultados deficientes en pruebas lingüísticas. Con cada tarea, el grupo que usaba IA se volvía más perezoso y terminaba copiando y pegando los resultados. Todos los ensayos hechos con IA eran muy similares entre sí, usaban las mismas frases, expresiones e ideas. En cambio, los otros dos grupos mostraron una actividad cerebral mucho más alta. Al final del estudio pidieron a todos los grupos que reescribieran su primer ensayo. Los que usaron *ChatGPT* ni se acordaron qué escribieron, mientras que los dos otros grupos sí recordaban sus escritos y los mejoraron.

Los científicos que realizaron este estudio acuñaron el término *deuda cognitiva*, que consiste en una condición en la que la dependencia en sistemas externos como los grandes modelos de lenguaje llega a reemplazar a los procesos cognitivos necesarios para el pensamiento independiente (Kosmyna et al. 2025, 141). A corto plazo *ChatGPT* escribe el ensayo del curso de ciencias para entregar mañana, pero a largo plazo acumula costos cognitivos en las personas estudiantes; de ahí la deuda cognitiva. Según este estudio, puede llegarse a una menor indagación crítica, una mayor vulnerabilidad a la manipulación y a una creatividad disminuida como producto del uso indiscriminado e inconsciente de la IA.

También existen riesgos de bajo nivel y gran escala, o gran impacto: por ejemplo, el arraigo sistémico de prejuicios y discriminación, aumento de la desigualdad, erosión de la confianza social y de las comunidades, contaminación del ecosistema informativo, declive de la prensa libre, vigilancia masiva, *fake news*, distorsión de evidencia, *deep fakes*, y la lista continúa (Narayanan y Kapoor 2025). Estos riesgos de menor escala son sumamente importantes porque apuntan no la tecnología en sí sino a los usos que nosotros le damos y a preguntarnos qué clase de seres humanos seremos.

Estos ejemplos se podrían complementar con muchísimas otras experiencias. Estoy segura que, como docentes, estudiantes y personal administrativo, también hemos tenido nuestros roces con esta tecnología. Por ejemplo, me pasó en un curso que escribí un ensayo de un tema con el que no estaba muy familiarizada y antes de entregarlo al profesor decidí pasarlo por *Turn It In*, una herramienta que detecta plagio y escritura con IA, In para ver qué decía sobre mi ensayo. Me había costado muchísimo y además yo escribo mis borradores a mano. La primera vez me salió un porcentaje de 70% de escritura con IA, y yo, incrédula, lo pasé de nuevo por *Turn It In*. La segunda vez dio un resultado de un 92% de IA… ¿Qué está pasando aquí? ¿O yo soy una IA envuelta en una carcasa humana o estas herramientas no están bien afinadas? Me obligó a ir a conversar con el profesor y comentarle la situación, quien afortunadamente dio fe públicamente de que mi ensayo era original.

Con estos componentes de una visión sistémica, una vigilancia tecnológica de las otras universidades, una actualización sobre las propias políticas universitarias y reportes de experiencia en las aulas podemos aproximarnos a un panorama que represente de manera más o menos acertada el tipo de políticas y normativas que necesitaríamos desarrollar en la UCR. Les propongo estos cuatro componentes, pero ciertamente puede haber más, y debería haber más. Esto es solamente para iniciar la conversación que colectivamente deberíamos tener sobre este tema y cómo afrontarlo desde nuestras capacidades como estudiantes, docentes y personal administrativo.

Conclusiones

Como pueden ver, el tema es amplio, complejo, y con muchas aristas. Nada en este tema es absoluto. Al contrario: el tema cambia, pero debemos pronunciarnos y pronto. Sí, es cierto que es experimental, y que no hay manera de saber, excepto aprendiendo de los errores e informándose casi por el minuto. Nosotros ya nos entremezclamos con la IA y no tenemos control sobre cómo es que lo hacemos, qué uso se le da a nuestros datos y qué se genera de nuestra interacción, tanto para nosotros como para las compañías que desarrollan los modelos.

¿Qué podemos hacer? Talleres prácticos del uso, talleres de reflexión (si así queremos llamarlos), capacitaciones especialmente para los profesores y también para los estudiantes y personal administrativo. Se pueden conformar comisiones y comités para una toma de decisiones estratégica. Pero todo esto que menciono no se impone de manera unilateral, como lo dice la resolución de la Vicerrectoría de Docencia. La Escuela de Filosofía como un todo compuesto del profesorado, estudiantado y personal administrativo, debe gestionar espacios propios de reflexión porque como una unidad académica sabemos cómo afecta la filosofía.

Con esto concluye esta es una breve introducción a la IA y los potenciales usos para la investigación filosófica. Los cuatro componentes que enumeré sirven para iniciar la conversación, pero propongo que tengamos una discusión abierta sobre el tema. Esto es solo el principio.

Notas

1. Ver «Principia Symbients» de Eugenio Battaglia disponible en https://meaning.systems/principia-symbients/
2. Salvatore Iaconesi elabora sobre los principios del *Nuovo Abitare* en https://xdxd-vs-xdxd.medium.com/the-illustrated-principles-of-nuovo-abitare-5f2e63bbb9fc
3. Este documento está disponible en el siguiente enlace: https://docs.google.com/spreadsheets/d/1lM6g4yveQMyWeUbEwBM6FZVxEWCLfvWDh1aWUErWWbQ/
4. El sitio web del proyecto Ethical AI se puede consultar en: https://www.ethicalaiucr.info/index.php

Referencias bibliográficas

Academic Senate of the California State University. 2025a. «The Possible Use of AI in Instruction». *The California State University AI Commons*, 8 de mayo. Acceso el 8 de octubre de 2025. https://genai.calstate.edu/sites/default/files/2025-05/The%20Possible%20Use%20of%20AI%20in%20Instruction_Resolution.pdf

———. 2025b. «Guidelines for Faculty Regarding AI in Instruction». *The California State University AI Commons*, 8 de mayo. Acceso el 8 de octubre de 2025. https://genai.calstate.edu/communities/faculty/guidelines-faculty-regarding-ai-instruction

Ahmad, Meher, Jessica Grose y Tressie McMillan Cottom. 2025. «What A.I. Really Means for Learning». *New York Times*, 12 de agosto. Acceso el 6 de octubre de 2025. https://www.nytimes.com/2025/08/12/opinion/ai-college-classrooms-chatgpt.html

Dreyfus, Hubert L. y Stuart E. Dreyfus. 1988. *Mind over machine*. New York: The Free Press.

Elicit. 2025a. «Elicit's limitations». Acceso el 17 de octubre de 2025. https://support.elicit.com/en/articles/549569

———. 2025b. «Elicit's source for papers». Acceso el 17 de octubre de 2025. https://support.elicit.com/en/articles/553025

EthicalAI Project. 2025. «Ethical AI Project». Acceso el 8 de octubre de 2025. https://www.ethicalaiucr.info/index.php

Garber, Alan M., Meredith Weenick y Klara Jelinkova. 2025. «Guidelines for Using ChatGPT and other Generative AI Tools at Harvard». *Office of the Provost*. Acceso el 8 de octubre de 2025. https://provost.harvard.edu/guidelines-using-chatgpt-and-other-generative-ai-tools-harvard

Harvard University Information Technology. 2025. «Guidelines for the use of Generative AI tools at Harvard». *Harvard University Information Technology: Policies & Guidelines*. Acceso el 8 de octubre de 2025. https://www.huit.harvard.edu/ai/guidelines

Haugeland, John. 1997. «What Is Mind Design?». En *Mind Design II*, editado por John Haugeland, 1-28. Cambridge: The MIT Press.

Hernández, Victoria. 2025. «Voz Experta: La UCR lidera el camino hacia el uso ético de la inteligencia artificial en la educación superior». *Universidad de Costa Rica*, 16 de julio. Acceso el 8 de octubre de 2025. https://www.ucr.ac.cr/noticias/2025/7/16/voz-experta-la-ucr-lidera-el-camino-hacia-el-uso-etico-de-la-inteligencia-artificial-en-la-educacion-superior.html

Hofstadter, Douglas. 1995. *Fluid Concepts and Creative Analogies*. Nueva York: Basic Books.

Khalifa, Mohamed y Mona Albadawy. 2024. «Using artificial intelligence in academic writing and research: An essential productivity tool». *Computer Methods and Programs in Biomedicine Update* 5. https://doi.org/10.1016/j.cmpbup.2024.100145

Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein y Pattie Maes. «Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using AI Assistant for Essay Writing Task». *arXiv*:2506.08872, 10 de junio. Acceso el 8 de octubre de 2025. https://arxiv.org/abs/2506.08872

Lanier, Jaron. 2023. «There is no A.I.». *The New Yorker*, 20 de abril. Acceso el 6 de octubre de 2025. https://www.newyorker.com/science/annals-of-artificial-intelligence/there-is-no-ai

McCarthy, John, Marvin Minsky, Nathaniel Rochester y Claude Shannon. 1955. «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence». Acceso el 8 de octubre de 2025. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Narayanan, Arvind y Sayash Kapoor. 2025. «AI as Normal Technology». *Knight First Amendment Institute at Columbia University*, 15 de abril. Acceso el 6 de octubre de 2025. https://knightcolumbia.org/content/ai-as-normal-technology

UNESCO. 2021. *Inteligencia artificial y educación: guía para las personas a cargo de formular políticas*. Place de Fontenoy: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000379376

———. 2022. *Recomendación sobre la ética de la inteligencia artificial*. Place de Fontenoy: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

———. 2025. «Ética de la inteligencia artificial: La recomendación». Acceso el 8 de octubre de 2025. https://www.unesco.org/es/artificial-intelligence/recommendation-ethics

Vicerrectoría de Docencia. 2023. «Resolución VD-12784-2023: Lineamientos académicos y administrativos para la docencia en ambientes virtuales de aprendizaje». Acceso el 8 de octubre de 2025. https://www.cea.ucr.ac.cr/images/desarrollocurricular/Resolucion-Vicerrectoria-de-Docencia-VD-12784-2023.pdf

Vicerrectoría de Docencia. 2025. «Circular VD-15-2025». Acceso el 8 de octubre de 2025. https://vd.ucr.ac.cr/sites/default/files/2025-04/Circular%20VD-15-2025.pdf

Weizenbaum, Joseph. 1966. «Computational Linguistics». *Communications of the ACM* 9, n.º 1: 36-45. https://cse.buffalo.edu/~rapaport/572/S02/weizenbaum.eliza.1966.pdf

Zao-Sanders, Marc. 2025, «How People Are Really Using Gen AI in 2025». *Harvard Business Review*, 9 de abril. Acceso el 6 de octubre de 2025. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

Datos biográficos

Amanda Sagasti Charpentier (amanda.sagasti@ucr.ac.cr). Estudiante de la Maestría Académica en Filosofía en la Universidad de Costa Rica.

Recibido: 21 de octubre, 2025.
Aprobado: 27 de octubre, 2025.
DOI: 10.15517/revfil.2026.3134