INTELIGENCIA ARTIFICIAL Y EDUCACIÓN SUPERIOR. Posibilidades, riesgos aceptables y limites que no se deben traspasar

INTEL·LIGÈNCIA ARTIFICIAL I EDUCACIÓ SUPERIOR. Possibilitats, riscos acceptables i limitis que no s'han de traspassar.

ARTIFICIAL INTELLIGENCE AND HIGHER EDUCATION. Possibilities, acceptable risks and limits that should not be crossed.

Luis Miguel González de la Garza

Profesor Titular de Derecho Constitucional

Universidad Nacional de Educación a Distancia (UNED)

E-mail: lmdelagarza@der.uned.es

ORCID: https://orcid.org/0000-0002-8628-0079

Resumen: La IA Generativa ha llegado también al mundo de la universidad, tanto en el plano de los alumnos que acceden a los estudios superiores como en el del profesorado que se encuentra en su doble vertiente como docentes e investigadores con esta realidad tecnológica que no hace sino crecer. Sin embargo, es preciso conocer los límites de lo que la IA Gen puede ofrecer así como los riesgos que estas tecnologías pueden suponer para la enseñanza superior ya que frente a las ventajas que toda tecnología posee, surgen también importantes dudas sobre los riesgos que estas podrán acarrear en un futuro a medio y largo plazo. En este trabajo nos centramos esencialmente en los riesgos que diversos científicos sociales han detectado, fundamentalmente, para intentar aportar argumentos para una reflexión serena y esencialmente

prudente sobre la incorporación la IA Gen a la educación superior. Casi todos los males de pueblos e individuos dimanan de no haber sabido ser prudentes y enérgicos durante un momento histórico, que no volverá jamás, señalaba Santiago Ramón y Cajal, estas palabras de nuestro premio Nóbel de Medicina y padre de la neurociencia nos advierten de que hay momentos críticos en la regulación de los problemas antes de que estos se transformen o puedan transformarse en irresolubles como consideramos en este estudio, ese momento en el campo de la IA Gen ha llegado.

Palabras claves: inteligencia artificial, inteligencia artificial generativa, educación superior, investigación, riesgos.

Resum: La IA Generativa ha arribat també al món de la universitat, tant en el pla dels alumnes que accedeixen als estudis superiors com en el del professorat que es troba en el seu doble vessant com a docents i investigadors amb aquesta realitat tecnològica que no fa sinó créixer. No obstant això, cal conèixer els límits del que la IA Gen pot oferir així com els riscos que aquestes tecnologies poden suposar per a l'ensenyament superior ja que enfront dels avantatges que tota tecnologia posseeix, sorgeixen també importants dubtes sobre els riscos que aquestes podran implicar en un futur a mitjà i llarg termini. En aquest treball ens centrem essencialment en els riscos que diversos científics socials han detectat, fonamentalment, per a intentar aportar arguments per a una reflexió serena i essencialment prudent sobre la incorporació la IA Gen a l'educació superior. Gairebé tots els mals de pobles i individus dimanen de no haver sabut ser prudents i enèrgics durant un moment històric, que no tornarà mai, assenyalava Santiago Ramón y Cajal, aquestes paraules del nostre premi Nóbel de Medicina i pare de la neurociència ens adverteixen que hi ha moments crítics en la regulació dels problemes abans que aquests es transformin o puguin transformar-se en irresolubles com considerem en aquest estudi, aquest moment en el camp de la IA Gen ha arribat.

Paraules clau: intel·ligència artificial, intel·ligència artificial generativa, educació superior, recerca, riscos.

Abstract: Generative AI has also reached the world of the university, both at the level of students who access higher education and at the level of teachers who find themselves in their double aspect as teachers and researchers with this technological reality that only grows. . However, it is necessary to know the limits of what Gen AI can offer as well as the risks that these technologies can pose for higher education since, despite the advantages that every technology has, important doubts also arise about the risks that these technologies may pose. carry in the medium and long term future. In this work we focus essentially on the risks that various social scientists have detected, fundamentally, to try to provide arguments for a calm and essentially prudent reflection on the incorporation of Gen AI into higher education. Almost all the evils of peoples and individuals arise from not having known how to be prudent and energetic during a historical moment, which will never return, Santiago Ramón y Cajal pointed out, these words from our Nobel Prize winner in Medicine and father of neuroscience warn us that There are critical moments in the regulation of problems before they become or can become unsolvable, as we consider in this study, that moment in the field of Gen AI has arrived.

Keywords: artificial intelligence, generative artificial intelligence, higher education, research, risks

1 Introducción

Las tecnologías que bajo el concepto de Inteligencia Artificial se agrupan en la actualidad y las que en un futro se incorporarán, como la próxima revolución de la computación cuántica, suponen cambios relevantes en nuestras modernas sociedades de la información. Sin embargo, la expresión Inteligencia Artificial pudiera quizá traducirse mejor como "Habilidades de procesamiento electrónico de la información" expresión seguramente poco efectista y de poca utilidad para los departamentos de marketing de las empresas que desarrollan estas tecnologías pero que se compadecen de forma más precisa con la idea de lo que realizan realmente estas tecnologías en vez de con el concepto humano de "Inteligencia" la cual no está clara su definición entre los expertos y de la que no se conoce, por el momento, su funcionamiento desde una perspectiva neurobiológica.

Sin embargo y pese a lo anterior las grandes corporaciones que impulsan estas tecnologías como Microsoft, Google, Meta, Open AI, entre otras con finalidades comerciales internacionales de inmensa envergadura han logrado crear una ecología publicitaria viral a nivel internacional de la que es difícil para los Estados y para las organizaciones públicas y privadas sustraerse para no caer en la acusación de no seguir el ritmo de los tiempos o del progreso, pese a que en un futuro puedan derivarse graves consecuencias indeseables por actuaciones de seguimiento no debidamente reflexionadas ni previstos los posibles cursos anómalos de tales tecnologías. Es claro no obstante que nos encontramos en una corriente histórica prácticamente irresistible de promoción de unas tecnologías que avanzan a velocidad uniformemente acelerada y contra el principio de precaución y que podría desencadenar fenómenos equivalentes a los daños de la telefonía móvil en el caso de los menores que empezaron en el año 2010 y que se ha traducido en la generación de problemas educativos y de salud mental severos para la infancia y la juventud y para la sociedad en su conjunto tanto en los EE.UU., como en Europa y que en estos momentos se están evaluando. Hay que recordar como señala (Horwitz: 2024) que estas multinacionales operan con una absoluta falta de transparencia en sus políticas industriales y son conocedoras internamente de los efectos que tienen sus tecnologías sobre la sociedad.

Precisamente la incorporación de la Inteligencia Artificial Generativa, IA Gen en adelante, a la Educación Superior es una apuesta de muy alto riesgo de la que se desconocen los posibles efectos tanto positivos como negativos que las mismas proyectarán tanto sobre las generaciones futuras como sobre la propia organización en la elaboración de la ciencia y su difusión en los centros de investigación superior como son las Universidades. Los efectos tras lo que actualmente parece una "*fiebre*" por incorporar estas tecnologías en todos los nichos de la sociedad, no sólo en la educación superior. En la actualidad hasta la elaboración de una barra de pan que no incorpore en su fase de producción la IA Gen es un producto sospechoso de haber sido formulado con técnicas retrogradas o primitivas.

La fuerza con la que la IA Gen se ha incorporado a nuestras sociedades ya ha originado un *nuevo y peligroso sesgo* que es preciso considerar en esta introducción porque ilustra algunos aspectos que veremos más adelante, este es el *efecto del Sesgo heredado de la IA*: Se basa en un exceso de confianza humana en la automatización inherente a la IA, en la tendencia a aceptar

excesivamente los resultados algorítmicos incluso cuando son notablemente incorrectos. Esto significa como señalan (Vicente y Matute, 2023) que los humanos no sólo están dispuestos a confiar en la IA porque somos *avaros cognitivos* -es decir no se quiere pensar si la máquina se supone que ya lo ha hecho por nosotros- sino también porque perciben que la IA es digna de confianza pudiendo inducir en las personas un efecto de *autoridad* en el cumplimiento de sus indicaciones o consejos con el que no se discrepa. Como señala (Carr, 2014) si realizamos una tarea o un trabajo por nuestra cuenta, usamos diferentes procesos mentales diversos de cuando confiamos en la ayuda de un ordenador. Si el software reduce nuestra implicación con el trabajo y, en particular, si nos empuja *a un rol más pasivo* como observador o controlador, eludimos el procesamiento cognitivo profundo que sostiene el *efecto generación*. Como resultado, obstaculizamos nuestra capacidad de acumular la clase de conocimiento rico y real que conduce a la sabiduría práctica. El *efecto generación* requiere precisamente el tipo de esfuerzo que la automatización con IA busca aliviar, reducir y en algunos casos eliminar. Este efecto obviamente es muy peligroso por la delegación no justificada en procesos de IA que puedan estar defectuosamente entrenados o sencillamente que como cualquier máquina pueda sufrir una avería en su procesamiento o sea alterada de forma maliciosa por un hackeo exitoso lo que es perfectamente posible.

2 Algunas consideraciones previas sobre la Inteligencia Artificial

Como hemos señalado más arriba no discriminamos muy bien que es la inteligencia artificial, porque no sabemos tampoco qué es o cuantas son las inteligencias naturales tanto humanas como no humanas, es decir, las de todas aquellas especies que nos rodean. Es claro que los insectos tienen inteligencia, como otras muchos y diversos órdenes de especies que conviven con nosotros, son inteligentes (Maher, 2022). Los psicólogos han cambiado sus criterios en los últimos 30 años para definir la inteligencia humana y sus teorías sobre las mismas. Por ejemplo para el neuropsicólogo y educador de la Universidad de Harvard, Howard Gardner en su obra *Frames of Mind: The Theory of Multiple Intelligences*. El autor describe ocho tipos de inteligencias en los niños. Las investigaciones realizadas identifican la existencia de zonas en el cerebro humano que corresponden a determinados espacios de conocimiento, todos ellos distintos y relativamente independientes entre sí. 1. *Inteligencia lógico-matemática*. Su

capacidad de resolución de problemas es muy llamativa y suele relacionarse con un tipo de inteligencia no verbal, es decir, que puede saber la respuesta a un determinado problema mucho antes de verbalizarla. A los niños que poseen este tipo de inteligencia se les da bien resolver misterios o pruebas de ingenio, hacer rompecabezas, realizar ejercicios de lógica, contar o hacer cálculos, los problemas informáticos y jugar a juegos de estrategia. 2. *Inteligencia lingüística*. Estos niños son hábiles y tienen preferencias por actividades como leer, conversar, contar chistes, escribir cuentos y poemas, aprender idiomas y jugar a juegos de palabras. 3. *Inteligencia espacial*. Este tipo de inteligencia tiene la capacidad de pensar en tres dimensiones. Las personas que la desarrollan son hábiles en la resolución de problemas espaciales como dibujar y pintar, leer mapas, contemplar cuadros, resolver laberintos o jugar a juegos de construcción. 4. *Inteligencia musical*. Es propia de niños con una habilidad innata para el aprendizaje de los diferentes sonidos, lo que se traduce en una gran capacidad para cantar, escuchar música, tocar instrumentos, componer canciones, disfrutar de conciertos y seguir diferentes ritmos. 5. *Inteligencia kinestésico-corporal*. Es la capacidad para usar todo el cuerpo en la expresión de ideas y sentimientos, y la facilidad en el uso de las manos para transformar elementos. A los niños que la manifiestan se les da bien bailar, actuar, imitar gestos o expresiones, hacer deporte, correr, moverse y saltar. *6. Inteligencia intrapersonal*. Distingue a aquellos que se conocen mejor a sí mismos. A estos niños les gusta trabajar de manera autónoma, establecen metas y se centran en alcanzarlas, comprenden sus sentimientos y saben cuáles son sus puntos fuertes y débiles. *7. Inteligencia interpersonal*. Opuesta a la inteligencia intrapersonal, es común entre las personas que se les da bien conversar, trabajar en equipo, ayudar a los demás, mediar en conflictos y conocer gente nueva. *8. Inteligencia naturalista*. Relacionada con el gusto por los temas medioambientales, plantas y animales. Estos niños disfrutan realizando actividades como ir de acampada, hacer senderismo, cuidar animales, conocer detalles de la naturaleza, reciclar y cuidar el medioambiente.

No vamos aquí a profundizar en los diversos "tipos de inteligencia" lo que es claro, es que los expertos en la materia de la cognición humana mantienen distintas teorías y no es un tema en absoluto cerrado, todo lo contrario. Siendo eso así, es sospechosamente llamativo que si en la investigación de la inteligencia humana natural existen muy diversas teorías en pugna acerca de lo que la inteligencia es y cuantos tipos hay o cuáles son sus características, un *reduccionismo*

artificial propio de la ignorancia (Burke, 2023) o de la era de posverdad que son términos en alguna medida equivalentes ha introducido en el circuito de la opinión pública la "*inteligencia artificial*" sin más adjetivos como un equivalente por completo erróneo. Ignorando lo que la inteligencia es con respecto al ser humano, que debe ser, supuestamente la referencia contra la que se contrasta esa inteligencia artificial que circula en la opinión pública como un tópico poderoso más de nuestra época y que parece invadirlo todo, más como un argumento de marketing o propaganda en muchas ocasiones, que como descriptiva de tecnologías que puedan definir y emular con precisión su marco de operaciones con respecto al ser humano, recordemos que "ninguna de las inteligencias artificiales disponibles" comprende o entiende lo que hace, lo que las diferencia de forma radical de la inteligencia humana.

3 La IA no entiende ni comprende absolutamente nada de lo que procesa y nos muestra como resultado de su procesamiento, información y desinformación con un alto componente de ruido.

Es importante tener presente la Primera Ley de (Kranzberg, 1986) La tecnología no es ni buena ni mala, ni neutral, efectivamente la tecnología no es en ningún sentido neutral, ya que posee un conjunto innato de caracteres bien definidos que tienen un impacto directo en las formas de lo que se puede hacer con ella, es decir, marca de forma estricta unas directrices de conformación social del tipo todo o nada que exigen adaptación a la misma y por esa razón carece de neutralidad.

Los defensores de la IA Gen, en lo que podemos definir como una epidemia sociogénica (Haidt, 2024) emplearán un argumento antiguo que consiste en señalar que hay que moderar o controlar los contenidos de la propia IA, por ejemplo, a través de los datos de entrenamiento de las redes generativas para limitar los problemas que veremos, lo que es una tarea seguramente imposible porque los modelos ya han sido entrenados con información masiva obtenida de Internet de la cual una buena parte es información basura: información falsa, información errónea o tendenciosa pero con la que se han entrenado muchos modelos de IA.

Marshall McLuhan observo con gran acierto por qué esa es una pista falsa, centrarse en el contenido. En su libro (McLuhan,1964) "Understanding Media: The Extensions of Man" el autor intentó explicar la profunda transformación de la sociedad provocada por las tecnologías de comunicación electrónicas de su época, desde el telégrafo y el teléfono hasta la radio y la televisión. Las tecnologías electrónicas nos brindaron un mundo muy diferente del mundo basado en la imprenta que comenzó cuando Johannes Gutenberg inventó la imprenta en la década de 1440. McLuhan intentó explicar a sus lectores coetáneos las formas en que estas tecnologías cambiaron los hábitos, las mentes, el yo y las sociedades de las personas, pero la gente sólo parecía capaz de pensar en los mensajes explícitos o implícitos transmitidos en los programas de televisión y los anuncios, lo que hoy sucede por ejemplo con los sesgos que operan en los modelos de aprendizaje de la IA Gen. ¿Ese anuncio contiene un mensaje subliminal sobre sexo o sobre política? ¿No podemos simplemente regular mejor el contenido de los anuncios y programas? No, dijo McLuhan, no entiendes la clave del problema, porque el medio es el mensaje. Cuando la televisión se hizo dominante, todo (incluidas las noticias y la cobertura de las elecciones) se convirtió en una forma de entretenimiento, para ser consumido pasivamente, y esto cambió la sociedad y la democracia en todo el mundo de maneras que son difíciles de articular pero seguramente disruptivas como señaló McLuhan .

En la actualidad una nueva tecnología disruptiva, la IA Gen, nos vuelve a recordar que el medio es el mensaje y que su forma es -precisamente- el mensaje pero debemos precavernos de que la moderación de contenidos sigue siendo una pista falsa propia de lo que se ha dado en llamar los insensibles idiotas tecnológicos porque el "contenido" de un medio sigue siendo el jugoso trozo de carne que lleva el ladrón para distraer al perro guardián de la mente y, una vez más, frente a no saber qué hacer frente a una disrupción social tecnológica de gran magnitud -la IA Gen- se dedican los esfuerzos a la regulación ingenua de los contenidos por dos razones, la primera es porque es lo más sencillo o socorrido que se puede hacer cuando no se sabe qué hacer dando la impresión de que se hace algo y lo segundo es porque los motores de los intereses económicos que impulsan estas tecnologías no encuentran frenos que adviertan de los riesgos que implican unas tecnologías que suponen beneficios astronómicos concentrados para quienes las impulsan y riesgos sociales difusos, no bien explicados o claramente identificables para la sociedad aunque estos puedan ser inmensos.

Debemos ser conscientes con (Jonas, 2015) de las heurísticas del temor: sólo la previsible desfiguración del hombre nos ayuda a alcanzar aquel concepto de hombre que ha de ser preservado de los peligros. Por ello no es inadecuado tener presente la ley de Murphy que señala que si algo puede salir mal, saldrá mal .En realidad la versión original dice que: "si hay dos o más maneras de hacer algo y una de ellas puede resultar en una catástrofe, alguien se decidirá por esta última". Edward Aloysius Murphy era ingeniero aeroespacial y formuló su ley en 1949 después de descubrir que estaban mal conectados todos los electrodos de un arnés para medir los efectos de la aceleración y deceleración en pilotos de aviación. La Ley de Murphy nos recuerda algo habitual en la ciencia experimental y es que "desconocemos lo que desconocemos y que por ello no hay nada más peligroso que la ignorancia en acción". Los expertos en Inteligencia Artificial, por ejemplo, conocen muy bien los límites de ésta, el problema se encuentra entre los no expertos y profanos que confunden "ciencia" con "ciencia ficción" y a partir de ahí se sumen en mundos puramente especulativos gravemente desconectados de la realidad científica y si lo anterior lo conectamos a una sociedad ansiosa de recibir novedades y noticias rápidas, fascinantes y superficiales a diferencia de la realidad científica que es pausada, metódica, prudente y reflexiva llegamos a lo que podríamos llamar: "efervescencias mediáticas sociales guiadas" en torno a temas como la IA Gen muy de moda en nuestra era de la posverdad.

Lo que debe preocuparnos seriamente no es una capacidad especialmente abrumadora de una superinteligencia artificial inexistente como señala correctamente (Larson, 2022) capaz de hacerlo todo o conocerlo todo propia de la ciencia ficción que pueda realizar acciones verdaderamente inteligentes e intencionales en un futuro a corto o medio plazo, sino el hecho de que estamos encomendado cada vez más decisiones realmente importantes a máquinas estúpidas que, en ningún momento son "conscientes de lo que hacen" eso es lo verdaderamente trascendente.

Es conocido el caso del usuario de Siri que le dijo: "Siri, apunta lo siguiente en la lista de la compra". A lo que Siri respondió: "Lo siguiente" apuntado en la lista de la compra". Las máquinas de procesamiento de datos actuales constituyen un ejemplo de lo que el filósofo de la mente Daniel Dennet llamaba "habilidad sin comprensión" y esa es, como recuerda entre nosotros (Mántaras, 2020) una buena definición de lo que la IA es hoy. Los sistemas más

avanzados de IA como los que se basan en el aprendizaje profundo -redes neuronales- detectan correlaciones pero no relaciones de causa y efecto. Por ejemplo, no pueden aprender que es la salida del sol lo que provoca el canto del gallo y no al revés, cuestiones que un niño de cinco años comprende perfectamente. Estos aspectos fueron detallada y correctamente examinados por (Searle, 2000) en efecto, el argumento fundamental que Searle aduce -expuesto gráficamente a través de la ya clásica imagen y argumento de la sala china- desarrollado en el trabajo "Mentes, cerebros y ciencia" y perfectamente perfilados en "El misterio de la conciencia" en España es que:

1) la sintaxis, no es lo mismo que, ni es por sí misma suficiente para la semántica. Pero el argumento más profundo contra el computacionalismo es que:

2) los rasgos computacionales de un sistema no son intrínsecos a la mera física de ese sistema, sino que necesitan de un usuario o interprete externo que proporcione una interpretación computacional al sistema, es decir, el ser humano consciente. Podríamos aquí formular la siguiente pregunta, si un humano o conjunto de estos no se encuentra presente ¿existiría la IA Gen? la respuesta obviamente es negativa, no existe autonomía de esa supuesta IA del ser humano ni en el sentido más elemental del concepto de autonomía.

En ese sentido deberían estar presididas todas las aproximaciones al fenómeno de la IA Gen desde el principio de "precaución" reconocido por la UE ya en el año 2000 para diversos entornos que podríamos considerar de riesgo y entre los que se encuentran los "usos de la IA Gen" considerada sólo como lo que es, una herramienta compleja. Estas ideas también han sido plasmadas, por ejemplo, en la "Declaración de Barcelona para el adecuado desarrollo de la IA" de 8 de marzo de 2017. Y recientemente se ha aprobado la Resolución del Parlamento Europeo de 20 de octubre de 2020 con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014 (INL) que será a nuestro juicio un instrumento de una relevancia extraordinaria para lograr que las empresas que desarrollen sistemas de IA que generen riesgos asuman la responsabilidad civil de los daños que se puedan generar.

La exigencia de responsabilidad civil es una herramienta poderosa si además opera en el marco de una agencia reguladora que analice cuidadosamente la "fiabilidad" de los sistemas en los que se haga uso de la IA Gen. Éste sería, precisamente, el punto 2 de la Declaración de Barcelona cuando señala que: "Todos los sistemas artificiales que se utilizan en nuestra sociedad deben someterse a pruebas para determinar su fiabilidad y seguridad". Por lo que es normal que se haga lo mismo con los sistemas de IA particularmente en dominios como la medicina o los robots autónomos que emplean IA. Aunque se desarrollaron procedimientos de verificación y validación para sistemas basados en el conocimiento en los años ochenta y noventa, todavía faltan para la IA Gen basada en datos. Por supuesto, en este momento las prácticas de aprendizaje automático hacen una distinción entre un conjunto de datos de ejemplo utilizado para el entrenamiento y un conjunto de prueba utilizado para medir hasta qué punto un sistema ha alcanzado niveles adecuados de rendimiento, pero todavía existe una diferencia significativa entre un conjunto de prueba y pruebas reales en condiciones del mundo real. Además, una vez que estén disponibles las metodologías adecuadas de verificación y validación, necesitaremos una red de agencias en países europeos (o una agencia centroeuropea) que las utilice. Deben convertirse en la autoridad independiente para certificar aplicaciones de IA antes de que se utilicen de forma generalizada. El Parlamento Europeo ha decidido recientemente crear una agencia de robótica e inteligencia artificial que potencialmente podría asumir esta tarea" y en España se ha creado recientemente la agencia AESIA que, pese a sus mejores intenciones, no podrá cumplir su finalidad . El control de fiabilidad en IA es muy importante y ello es debido a que ni los diseñadores de tales sistemas pueden prever cómo se comportara en muchas situaciones no previstas, ya que se trata de fenómenos emergentes, es decir, en los que la resultante es más y es cualitativamente diferente que la suma de las partes de que se compone y el comportamiento, conducta o resultado emergente surge de la interacción de los elementos constituyentes entre ellos y con su entorno con la producción de resultados inesperados o insospechados, es decir, nos situamos en el paradigma epistemológico VUCA (Volatilidad; Incertidumbre: Complejidad y Ambigüedad) en el que se producen los fenómenos mucho más habituales de lo que se podría suponer denominados "alucinaciones" de determinados tipos de IA Gen como precisan (Alkaissi y McFarlane, 2023) especialmente graves en el mundo académico y que para autores como (Emsley, 2023) no son alucinaciones sino sencillamente invenciones y falsificaciones con quien coincidimos porque en los grandes modelos de lenguaje

(LLM) es imposible evitarlos como han demostrado de forma consistente (Xu et al, 2024) pero existen muchos más argumentos en este sentido como sistematiza (Leffer, 2024) .

Una agencia como la propuesta a nivel europeo -Oficina de IA - podría, con ciertas reformas, ser una forma eficiente de verificar que un producto o un servicio basado en IA Gen antes de salir al mercado de consumo ha pasado por un proceso de verificación de fiabilidad y consistencia con los principios que debe cumplir la IA en diversos entornos muy específicos en los que esta pueda ser aplicada, pero repetimos, no porque la IA tenga capacidades extraordinarias, sino precisamente por lo contrario porque no las tiene y la IA puede ser perfectamente implementada a través de los algoritmos apropiados, por ejemplo, para que no sea neutral o para producir resultados económicamente favorables a las empresas y desfavorables para los usuarios, es decir resultados con trampa.

Es preciso señalar que los sistemas de IA aplicados a Educación son considerados de alto riesgo por el reciente Reglamento de IA de la UE, en ese sentido el considerando 56 del Reglamento.

Los sistemas de alto riesgo exigen procedimientos rigurosos para evitar algunas de sus posibles o probables consecuencias negativas, con carácter general estas han sido a nuestro juicio e inicialmente correctamente consideradas por el Reglamento de IA de la UE, pueden verse en ese sentido los considerandos 72 a 78 del citado Reglamento y que no reproducimos por su extensión o en diversos documentos de la UNESCO o del Consejo de Europa que también son sensbles a estos problemas.

Por último y volviendo al medular problema de la la calidad de la información que alimenta los modelos de IA Gen, es sumamente importante para el entrenamiento de modelos sin supervisión y con supervisión (Berzal: 2018) que la información procesada errónea o falsa opera de forma equivalente a los sesgos en el entrenamiento de modelos de aprendizaje. Esa información defectuosa mezclada sin cuantificación con la información veraz y real genera inevitablemente ruido en el resultado final de los modelos, en el sentido preconizado por (Khaneman et al, 2021)

Es importante observar que esa señal de entrada con ruido informativo forma la historia de las narrativas electrónicas que procesa la IA Gen y como ha demostrado recientemente (Garnier-Brun et al, 2024) en un contexto ligeramente diferente -el de modelos de decisión económica-

puede ser valido para los modelos de entrenamiento basados en la "experiencia" y sus modelos de toma de decisiones en los que los resultados del procesamiento de la información obtenidos pueden situarse muy por debajo de lo óptimo dando como resultado información con una alta tasa de error que pueden traducirse en modelos de toma de decisiones producidos por la IA Gen con márgenes de error elevados lo que puede producir grave confusión en los resultados obtenidos por esta y por lo tanto inadecuada para la educación y para ser integrada en procesos científicos rigurosos.

4 La educación superior

La educación de los seres humanos hasta ahora se ha basado fundamentalmente en el esfuerzo personal y en el deseo de aprender sistemáticamente los conocimientos necesarios para adquirir las competencias precisas para afrontar las tareas cualificadas de diversas áreas de conocimiento práctico y teórico de las diversas disciplinas académicas. Un sistema educativo en el que el esfuerzo personal corra el riesgo de erosionarse gravemente puede significar que el dominio de las habilidades que se pretende con la educación transforme sus cimientos en la ausencia de esfuerzo personal puesto que éste podría ser desempeñado por sistemas de procesamiento de información capaces si de ahorrar mucho tiempo a los alumnos, pero tanto ahorro como se produce puede ir anudado a una falta de conocimientos sólidos que han sido transferidos a los modelos de IA generativa. El esfuerzo personal puede suplantarse con sistemas informáticos que realicen el trabajo que hasta ahora venían realizando los alumnos en sus diversas fases de aprendizaje en los planes de estudio. Parece posible pensar en que el acceso a este tipo de tecnologías incidirá igualmente en el deseo de aprender cuando éste aprendizaje puede obtenerse con suma facilidad a través de estas herramientas con un previsible efecto de desmotivación, transformando la educación en una fórmula de unir piezas informativas quizás no debidamente comprendidas por quienes tienen como misión central aprender una correcta metodología de aprendizaje y una profunda comprensión de los conceptos que se deben integrar en la formación de los alumnos para la resolución de los problemas que finalmente son el objeto de muchos tipos de formaciones por ejemplo las de naturaleza técnica o sociales. Indudablemente estos aspectos tienen implicaciones psicológicas importantes de las que se desconoce actualmente su alcance tanto para el alumnado como para los docentes y sería importante disponer de estudios sobre el

sentimiento de utilidad de los alumnos en sus procesos formativos, como por ejemplo el sentimiento de transformar el centro de atención del estudiante hacia los medios de los que se sirve, siendo estos, es decir, las herramientas las imprescindibles y generando sentimientos de capacidades inferiores por parte de los propios estudiantes. Es decir percibir erróneamente que poco se puede aportar a unos modelos que disponen de toda la información. Estas tecnologías son *inhibidoras de la implicación personal en el descubrimiento autónomo del saber que es esencial en el aprendizaje*.

Parece razonable pensar que no se trata de comparar la IA Gen con los hombres y mujeres más capacitados sino, precisamente con las capacidades medias en las que, no ahora, pero en un futuro a medio plazo estas herramientas podrán sustituir a muchas personas que se encuentran precisamente en ese substrato medio.No existen, por el momento, en la literatura científica excesivos modelos tentativos de hasta qué punto debe penetrar la IA Gen en los modelos educativos, siendo el denominado (AIAS) y su escala de evaluación tal vez el más completo hasta el presente elaborado por (Perkins, Furze, Huevas y MacVaugh). *[Los avances recientes en la Inteligencia Artificial Generativa (IA Gen) han creado un cambio de paradigma en múltiples áreas de la sociedad, y es probable que el uso de estas tecnologías se convierta en una característica definitoria de la educación en las próximas décadas. IA Gen ofrece oportunidades pedagógicas transformadoras y, al mismo tiempo, plantea desafíos éticos y académicos. En este contexto, describimos una herramienta práctica, simple y suficientemente completa para permitir la integración de herramientas IA Gen en la evaluación educativa: la Escala de Evaluación de IA (AIAS).*
La AIAS permite a los educadores seleccionar el nivel apropiado de uso de IA Gen en las evaluaciones en función de los resultados de aprendizaje que buscan abordar. La AIAS ofrece mayor claridad y transparencia para estudiantes y educadores, proporciona una herramienta política justa y equitativa para que las instituciones trabajen y ofrece un enfoque matizado que aprovecha las oportunidades de IA Gen al tiempo que reconoce que hay casos en los que dichas herramientas pueden no ser pedagógicamente apropiadas o necesarias. Al adoptar un enfoque práctico y flexible que pueda implementarse rápidamente, la AIAS puede constituir un punto de partida muy necesario para abordar la incertidumbre y la ansiedad actuales con respecto a la IA Gen en la educación. Como objetivo secundario, es estudio se involucra con la literatura

actual y defiende un discurso reorientado sobre las herramientas IA Gen en la educación, uno que ponga en primer plano cómo las tecnologías pueden ayudar a apoyar y mejorar la enseñanza y el aprendizaje, lo que contrasta con el enfoque actual en IA Gen como facilitador del aprendizaje académico. https://doi.org/10.48550/arXiv.2312.07086].

En ese sentido señalan los autores que el discurso en torno a las herramientas IA Gen en la educación, particularmente desde la llegada de ChatGPT en noviembre de 2023, ha experimentado una rápida transformación. Inicialmente, la atención se centró en frenar el uso de estas herramientas mediante ajustes de políticas y modificaciones en las estrategias de evaluación, y la aparición de herramientas de detección de texto mediante IA reforzó aún más este enfoque, ya que las IES (instituciones de educación superior) las consideraron fundamentales para identificar los usos indebidos de esta tecnología. Sin embargo, estudios recientes han destacado limitaciones significativas de estas herramientas de detección, incluidos desafíos con la precisión, riesgos de acusaciones falsas y posibles sesgos contra hablantes no nativos.

Estos hallazgos han estimulado un creciente reconocimiento entre las instituciones educativas de la necesidad de reconsiderar la dependencia de las herramientas de detección. Existe un llamado cada vez mayor a un enfoque alternativo en el que la integridad académica siga siendo primordial, pero el uso de herramientas IA Gen se integre para fomentar el desarrollo de habilidades de los estudiantes, particularmente en preparación para futuros entornos de trabajo donde estas herramientas podrían prevalecer. El AIAS ha surgido como una herramienta en este contexto. Replantea la conversación con los estudiantes sobre IA Gen desde una postura prohibitiva a una más permisiva, guiándolos sobre cómo usar estas herramientas de manera efectiva, utilizando una escala de cinco puntos diseñada para respaldar un equilibrio entre simplicidad y claridad.

No obstante, persisten muchos desafíos, en particular en relación con el acceso al software IA Gen, las aplicaciones prácticas en diversos entornos y la gestión de la brecha digital. El acceso a las herramientas de IA Gen puede variar según la ubicación, y las versiones premium de algunas herramientas comunes, como GPT-4, que exigen pagos mensuales. Si bien herramientas como AIAS demuestran una forma en que las IES pueden integrar eficazmente las herramientas

de IA gen para garantizar la participación continua de los estudiantes y el desarrollo de habilidades, esto debe equilibrarse con los requisitos institucionales más amplios de mantener la integridad académica, cambiar las expectativas de los estudiantes y los requisitos futuros de la industria. Esto requiere una cantidad considerable de agilidad institucional y un diálogo abierto con los estudiantes para equipar mejor a las nuevas generaciones para que aprovechen las herramientas de IA Gen como parte de sus futuras trayectorias profesionales y de aprendizaje.

Como vemos, en parte, el cambio de modelo se ha debido a que los instrumentos de reconocimiento de plagio -pensemos en herramientas como Turnitin, Grammarly o Plag entre otras herramientas- han llegado prácticamente al límite de su utilidad y ello es debido que como los modelos de IA Gen son entrenados con inmensas bases de datos tomadas sin filtrado y en bruto de Internet la información que generan las herramientas anti plagio detectan el plagio en la propia información de entrenamiento de los modelos de IA Gen. De esa forma las configuraciones para detectar plagios se han vuelto cada vez más complejas con los problemas ya señalados por los autores.

4.1 Tecnologías que pueden debilitar la atención y el pensamiento reflexivo y crítico característico de la educación superior.

Señala (Carr, 2020) que docenas de estudios a cargo de psicólogos, neurobiólogos, educadores y diseñadores web apuntan a la misma conclusión: cuando nos conectamos a la Red, entramos en un entorno que fomenta una lectura somera, un pensamiento apresurado y distraído, un *pensamiento superficial*. Es posible pensar profundamente mientras se navega por la Red, como es posible pensar someramente mientras se lee un libro, pero no es éste el tipo de pensamiento que la tecnología promueve y recompensa. Una cosa está clara: si, sabiendo lo que sabemos hoy sobre la plasticidad del cerebro, tuviéramos que inventar un medio de reconfigurar nuestros circuitos mentales de la manera más rápida y exhaustiva posible, probablemente acabaríamos diseñando algo parecido a Internet. No es sólo que tendamos a usar la Red habitualmente, incluso de forma obsesiva. Es también que la Red ofrece exactamente *el tipo de estímulos*

sensoriales y cognoscitivos —*repetitivos, intensivos, interactivos, adictivos*— que han demostrado capacidad de provocar alteraciones rápidas y profundas de los circuitos y las funciones cerebrales.

La "*superficialidad*" que apunta Carr es una consecuencia natural de la forma en la que las tecnologías se muestran en los sistemas de reproducción informática. Obtenemos una enorme cantidad de información cuando navegamos por internet o cuando usamos el teléfono móvil o cuando usamos sistemas de IA Gen, pero nos llega de manera muy fragmentada; muchos segmentos de información multimedia (sonidos, fotos, imágenes en movimiento, textos) que compiten entre sí, solapándose mutuamente por obtener la atención de un cerebro y un sistema nervioso al que esos efectos le producen placer y en muchos casos adicción a esa fuente ininterrumpida de estímulos positivos. En adultos es problemático y genera múltiples disfunciones en diversos ordenes de la conducta humana que aquí no abordamos, pero *en la educación y en los menores* los efectos se muestran dramáticos, razón por la que el legislador debe ser consciente de ello e impedir que estas tecnologías se usen en la educación de forma totalmente indiscriminada.

Lo anterior, como hallazgo, se sitúa en la línea de las investigaciones de los psicólogos sociales como (Haidt, 2024) es decir, podemos inicialmente pensar que estas tecnologías son inevitables, pero debemos ser conscientes de que no lo son, y que se pueden regular y de hecho se deben regular frente a un impulso que se realiza por las grandes multinacionales de la información como señalamos más arriba. De igual forma que se está dando marcha atrás de forma acelerada por sus nefastas consecuencias perfectamente acreditadas sobre los daños objetivos en la salud mental de los menores y jóvenes en el uso del uso de pantallas o telefonía inteligente, no es necesario esperar que surjan las peores consecuencias para desarrollar regulaciones tuitivas y prudenciales sobre el uso de las tecnologías en el ámbito de la educación superior. Si se permite que los negocios estén por encima de los intereses de la sociedad los perjuicios pueden ser extraordinariamente graves.

5 Los sesgos que acompañan a la penetración de la IA gen en la educación.

Los sesgos son errores mentales como precisa (Matute: 2019) que cometemos sistémicamente todas las personas y que se pueden predecir dado que en todos nosotros ocurren en las mismas situaciones y funcionan de la misma manera y en la misma dirección. Una descripción general (Phol: 2004) de estos sesgos son los que veremos seguidamente, y que son extraordinariamente relevantes por dos razones. Primera, en la información que procesan los sistemas de IA Gen basados en Big Data y Machine Learning estos datos masivos incorporan tales sesgos de forma nativa (Glauner et al: 2018) y pueden ser identificados. Segundo, en el desarrollo de la programación de IA se puede instruir al sistema experto para identificar -con ciertos límites- algunos de tales sesgos mediante técnicas de correlación estadística lo que pondría permitir -con muchas dificultades y en el futuro- su compensación, circunstancia no exenta -hay que añadir- de problemas éticos en educación importantes (Nguyen et al: 2022) ya que pueden conducir desde su uso en forma de un paternalismo ético siempre cuestionable, pensemos en los Nudges -empujoncitos conductuales- de (Sunstein y Thaler: 2009) a un totalitarismo político incompatible frontalmente con la libertad de pensamiento de una sociedad democrática abierta. Precisemos que estos sesgos pueden ser tratados como incentivos, pero difieren de los tradicionales tales como la coacción, la persuasión o la negociación (Grant: 2021) por su base neurobiológica y conductual, es decir, escapan a una primera revisión racional ya que operan en un marco motivacional donde la emoción y los procesos de toma de decisiones subconscientes y automáticos van primero y la razón opera después sólo si se examina con detalle la conducta tipo realizada o por realizar de forma reflexiva lo que no será lo habitual.

5.1 Los sesgos que podemos inicialmente identificar en la introducción de la IA GENS SESGOS

Aversión a la desposesión o aversión a la pérdida de la IA (efecto dotación): es la tendencia a preferir evitar las pérdidas en lugar de la posibilidad de adquirir ganancias. Así, se demandará más dinero para renunciar a un objeto que ya poseemos de lo que se estaría dispuesto a pagar para adquirirlo. Vender bienes que uno normalmente usa activa regiones del cerebro asociadas al disgusto o la pena. Comprar también activa esas áreas, pero solo cuando los precios se consideran demasiado elevados, cuando se siente que un vendedor está recibiendo un dinero que

excede el valor de cambio. Las imágenes de resonancia magnética funcional del cerebro también indican que comprar a precios particularmente bajos es algo placentero. Hoy en día sabemos que el efecto dotación nos afecta incluso con sólo haber tenido el objeto en la mano durante unos instantes. Por eso mismo, muchas empresas invitan a probar productos y dejan que se prueben y posean todo lo que sea necesario. El vínculo emocional que se establece con el objeto pone en marcha el efecto dotación, un poseedor se siente como propietario y se activa inmediatamente la aversión a perderlo, a no tenerlo. Deshacerse de algo, aunque se haga de forma voluntaria se puede interpretar como una pérdida y nadie desea perder lo que tiene aunque realmente no lo tenga. La IA Gen como herramienta cuando empieza a utilizarse e integrarse en los diversos sistemas o procesos desde los que pueden obtenerse beneficios forman ya parte de las posesiones del ser humano que las emplea y desposeerse de ellas causa una sensación de pérdida por lo que la tendencia será a su mantenimiento es una circunstancia no racional sino neurobiológica.

Efecto Bandwagon o efecto de arrastre: Es la tendencia a hacer (o creer en) algo porque muchas personas lo hacen (o lo creen). Está relacionado con el pensamiento de grupo o el comportamiento gregario. El efecto bandwagon (que tomó su nombre de la carreta que lleva la banda musical de un tren de circo) dicta que la probabilidad de que una persona adopte una creencia o conducta es directamente proporcional a cuántos otros ya la tengan, lo que significa que existe una tendencia psicológica a seguir o imitar las acciones y pensamientos de los demás, porque preferimos ajustarnos a lo preexistente, ya que es imposible no derivar nueva información de lo que otros piensan y hacen. Este sesgo es tan poderoso que es capaz, como ha demostrado (Edelson et al, 2011) de lograr que se llegue a la distorsión de la memoria de las personas individuales para que se adapte a la memoria del grupo en base a su capacidad de influencia. La memoria humana es sorprendentemente susceptible a las influencias sociales, pero se sabe poco sobre los mecanismos subyacentes. En los experimentos realizados se analizó cómo los errores de memoria inducidos socialmente se generan en el cerebro al estudiar la memoria de individuos expuestos a recuerdos de otros. Los participantes exhibieron una fuerte tendencia a conformarse con recuerdos erróneos del grupo, produciendo errores tanto duraderos como temporales, incluso cuando su memoria inicial era fuerte y precisa. Las imágenes cerebrales de resonancia magnética funcional revelaron que la influencia social modificó la

representación neuronal de la memoria. Específicamente, una firma cerebral particular de actividad mejorada de la amígdala y conectividad mejorada entre amígdala e hipocampo predijo alteraciones de la memoria duraderas, pero no temporales. Los hallazgos revelan cómo la manipulación social puede alterar la memoria y extender las funciones conocidas de la amígdala para abarcar las distorsiones de la memoria mediadas socialmente. Este efecto está ligado a la propaganda de las grandes multinacionales cuando impulsan una tecnología que adquirirá con el tiempo el carácter de necesidad social como la IA Gen.

Ilusión del control: Es la tendencia a sobreestimar el grado de influencia sobre otros eventos externos. De esta manera, los seres humanos tienden a creer que pueden controlar o al menos influir en las consecuencias o resultados que claramente no pueden controlar ni influir.

Sesgo de confirmación: Es la tendencia de las personas a favorecer la información que confirme sus propios presupuestos o hipótesis, sin importar si la información es verdadera.

Sesgo del optimismo irreal frente a datos contrarios al optimismo: El optimismo irreal es un rasgo humano omnipresente que influye en dominios que van desde las relaciones personales hasta la política y las finanzas o la adopción de nuevas tecnologías como la IA Gen. Cómo las personas mantienen un optimismo poco realista, a pesar de encontrar información que desafíe esas creencias sesgadas, no es del todo conocido, autores como (Tali Sharot, 2012) señalan que ello parece ser debido a que mostramos una asimetría llamativa, por la que las personas actualizan sus creencias más en respuesta a información que fue mejor de lo esperado en comparación con información que fue peor. Esta selectividad fue mediada por una falla relativa en la codificación de errores que deberían reducir el optimismo. Regiones distintas de la corteza prefrontal rastrearon errores de estimación cuando requerían una actualización positiva, tanto en individuos altamente optimistas como poco optimistas. Sin embargo, los individuos altamente optimistas exhibieron un seguimiento reducido de los errores de estimación que requerían una actualización negativa dentro de la circunvolución prefrontal inferior derecha. Estos hallazgos muestran que el optimismo está ligado a una falla de actualización selectiva y una codificación neuronal disminuida de información no deseada con respecto a lo que sucederá en el futuro. Es claro que esa forma de procesar la información es esencial para el autoengaño

ya que impide ponderar la información relevante debido a que el optimista extremo valora mentalmente de forma inadecuada los datos de la realidad.

Prejuicio de desconfirmación o sesgo de disconformidad: Es la tendencia a realizar un crítico escrutinio de la información cuando contradice sus principales creencias y aceptar sin criterio aquella información que es congruente con sus principales creencias.

Percepción selectiva: Tendencia en la cual las ansias, esperanzas o ilusiones afectan a la percepción.

Efecto del Sesgo heredado de la IA: Se basa en un exceso de confianza humana en la automatización inherente a la IA, en la tendencia a aceptar excesivamente los resultados algorítmicos incluso cuando son notablemente incorrectos. Esto significa como señalan (Vicente y Matute, 2023) que los humanos no sólo están dispuestos a confiar en la IA porque son avaros cognitivos -es decir no se quiere pensar si la máquina se supone que ya lo ha hecho por nosotros- sino también porque perciben que la IA es digna de confianza pudiendo inducir en las personas un efecto de autoridad en el cumplimiento de sus indicaciones o consejos.

5.2 IA GEN: SESGOS, ERRORES Y FALSEDADES.

Los programas de IA Gen procesan los datos a través de sus algoritmos que incorporan sesgos como venimos señalando que pueden discriminan a ciertos grupos de la población tales como mujeres, personas con discapacidades, minorías raciales, minorías religiosas o étnicas, etc., porque los algoritmos carecen de una ética definida adoptan sus decisiones entre conjuntos de datos cuantificables y probabilidades y tratan de optimizar las soluciones de los problema que se les plantean generando incrementos en las respuestas de tendencia que se observan con mayores frecuencias estadísticas en las base de datos que los alimentan que toman de datos abiertos y no filtrados. Y en muchas ocasiones esta circunstancia no premeditada provoca que los sesgos formen parte de los algoritmos porque se encuentran en las bases de datos de procesamiento. Los algoritmos son programas informáticos, es decir “planes sofisticados de trabajo codificados informáticamente o secuencias de trabajo de lo que se va a hacer con los datos de entrada hasta producir una salida estructurada con arreglo a diversos criterios” y la ética de un algoritmo es la que se defina en su programación o en esos criterios, es decir, hay que

"diseñar la ética algorítmica" para que pueda identificar, valorar y en su caso compensar los sesgos y errores reproductivos aunque no se tarea sencilla como argumenta (Roselli et all, 2019). Este aspecto es correctamente señalado por la reciente Orden ejecutiva sobre el desarrollo y uso seguro y confiable de la inteligencia artificial de 30 de octubre de 2023 del Presidente norteamericano Joe R. Biden en su sección primera cuando expresa "Al final, la IA refleja los principios de las personas que la construyen, las personas que la utilizan y los datos sobre los que se construye". Por esa razón en la letra a) de su apartado segundo se proponen algunas de las medidas fundamentales de control.

Es éticamente deseable que el algoritmo no reproduzca los sesgos sociales discriminatorios, pero acontece que los sesgos se encuentran incrustados en los datos que las personas y la sociedad en su conjunto produce y los algoritmos los reproducirán y en ocasiones los magnificarán.

El otoño del año 2020 Google presentó una tecnología de inteligencia artificial innovadora llamada BERT (Atkinson, 2023) que cambió la forma en que los científicos en lingüística construyen sistemas que aprenden cómo las personas escriben y hablan. BERT es una red neuronal de código abierto que ha sido entrenada para procesar el lenguaje natural. BERT procesa las palabras en el contexto de una oración, en lugar de palabra por palabra.

Pero BERT, que ahora se está implementando en servicios como el motor de búsqueda de Internet de Google, tiene un problema: podría estar detectando y asimilando sesgos en la forma en que un niño imita el mal comportamiento de sus padres. BERT es uno, de una serie de sistemas de inteligencia artificial que aprenden de mucha información digitalizada, tan variada como libros antiguos, entradas de Wikipedia y artículos de noticias, páginas web, redes sociales de toda especie. Las décadas e incluso siglos de sesgos, junto con algunos nuevos se encuentran en todo ese material de entrenamiento de la IA Gen, es decir, dado que el sistema BERT opera sobre textos que contienen cientos de millones de datos con sesgos, estos necesariamente se recogerán y perdurarán.

BERT y sus sistemas equivalentes tienen más probabilidades de asociar a los hombres con la programación de computadoras, por ejemplo, y generalmente no otorgan suficiente crédito a las

mujeres como ha señalado (Kurita et al, 2021). Un programa informático de procesamiento de lenguaje natural decidió que casi todo lo escrito sobre el presidente Trump era negativo, incluso si el contenido real era positivo, es decir, cuando el algoritmo de procesamiento de la información se retroalimenta de información sesgada se genera un proceso en el que la información verdadera es omitida. Pero lo más significativo del sistema BERT u otros sistemas equivalentes es su capacidad para replicar o reproducir sesgos allí donde los encuentre y reproducirlos en formulaciones actualizadas de resultados que los recogen. Además, BERT tiene la capacidad de realizar evaluaciones emocionales haciendo posibles clasificaciones de esta naturaleza, lo que en el ámbito de los sesgos cognitivos emocionales facilita el uso de estas clasificaciones con las más diversas finalidades siendo singularmente peligrosas en el ámbito de la propaganda cognitiva virtual, la transformación social basada en emociones y el uso de datos emocionales con finalidades comerciales o políticas en el nuevo paradigma que ha transformado una sociedad de clases a una sociedad clasificada.

6 Imparcialidad de los algoritmos de la IA vs exactitud: un grave problema en la educación.

Como señalan (Soleve y Matsumi, 2023) es incorrecto considerar que las máquinas deciden como lo hacen los seres humanos y que son mejores porque supuestamente están libres de prejuicios. Las máquinas deciden de manera fundamentalmente diferente de los humanos y, a menudo, persisten los sesgos y los errores. Estas diferencias son especialmente pronunciadas cuando las decisiones deben realizar un juicio moral o de valor o involucran vidas y comportamientos humanos. Algunas de las dimensiones humanas de la toma de decisiones que causan grandes problemas también tienen grandes virtudes. Además, los sistemas de IA Gen a menudo se basan demasiado en datos cuantificables y excluyen los datos cualitativos. Mientras que ciertas cuestiones pueden reducirse fácilmente a datos cuantificables, como el clima, por ejemplo, las vidas humanas son mucho más complejas. Por ello comparar la toma de decisiones humana con las de una máquina es similar a comparar manzanas y naranjas, no manzanas podridas con otras frescas, desde una perspectiva complementaria (Kearns y Roth, 2020)

6.1 La cuantificación es transformadora.

Los algoritmos se centran en datos que pueden cuantificarse fácilmente y este tipo de datos distorsiona el resultado. Los problemas surgen cuando se confía en demasiados datos cuantitativos y se excluyen los datos cualitativos mucho más difíciles de obtener, la razón es que no todo es fácilmente cuantificable. La cuantificación ciertamente puede conducir a conocimientos que de otro modo no reconoceríamos al estructurar los datos cuantitativos. Pero el hecho de que podamos observar ciertas cosas a través de la cuantificación no significa que la cuantificación sea una forma superior de conocimiento o que deba ser la única forma de examinar realidades complejas, singularmente en la educación. Desafortunadamente, es fácil dejarmos seducir por los grandes conjuntos de datos cuantificados a través de BigData porque nos ayudan a ver dimensiones que no se aprecian de otra forma. Lambert Adolphe Jacques Quetelet (1796-1874) uno de los primeros pioneros de la estadística, proclamó con entusiasmo en 1835: "Cuanto mayor es el número de individuos observados, más se borran y dejan en un lugar prominente las particularidades individuales, ya sean físicas o morales desde el punto de vista de los hechos generales, en virtud de los cuales la sociedad existe y se preserva." Quetelet creía que la estadística era una manera más refinada de entender a la humanidad que considerar las irregularidades de individuos específicos. Pero centrarse en los "hechos generales" omite el rico entramado de peculiaridades de los individuos y la naturaleza humana.

Ciertamente, las estadísticas pueden ser muy útiles y los intentos particulares de clasificar, calificar o inferir basándose en datos estandarizados agregados pueden ser valiosos. Pero estas prácticas pueden estar plagadas de peligros porque los sistemas algorítmicos no se limitan a ver el mundo: lo simplifican. La filósofa (Martha Nussbaum, 2001) sostiene acertadamente que "las emociones están impregnadas de inteligencia y discernimiento" e implican "una conciencia de valor o importancia". Las emociones "son" parte integrante del sistema de razonamiento ético. Los algoritmos no experimentan emociones. Los algoritmos de la IA Gen pueden imitar lo que las personas podrían decir o hacer pero no comprenden las emociones ni las sienten y es cuestionable si algún día los algoritmos podrán incorporar las emociones en su procesamiento lo que es completamente dudoso.

(Green, 2022) señala un conflicto aún más fundamental entre la toma de decisiones algorítmica y la humana: los algoritmos ofrecen "coherencia y seguimiento de reglas" mientras que los humanos ofrecen "flexibilidad y discreción" además de la coherencia y el seguimiento de reglas. Cuando los formuladores de políticas piden que los humanos supervisen los algoritmos, a menudo no reconocen la "tensión inherente" entre estos aspectos y no brindan orientación suficiente sobre cómo resolver esta tensión.

7 El conflicto entre exactitud e imparcialidad en la IA GEN.

Los datos cuantificables no consisten en hechos neutrales; dichos datos son creados y seleccionados por seres humanos, lo que introduce sesgos en el algoritmo como venimos considerando. Hipotéticamente, proponen los optimistas de la IA Gen, que los datos podrían eliminarse de todo sesgo. Pero incluso si se eliminan los sesgos más obvios de los datos, estos todavía provienen de una sociedad de humanos donde los sesgos se encuentran en todas partes porque forman parte de nuestra forma de razonar y eso sencillamente está más allá del alcance de cualquier tecnología de proceso de datos. Cuando a los sesgos se les añade la falsedad de la información nos encontramos ante el problema de una IA Gen que difunde desinformación tendenciosa, es la base del reciente caso advertido por la UNESCO de la falsificación histórica del holocausto a través de la IA Gen que lo niega mediante narrativas falsas de amplia difusion de escala. Esto es un fenómeno mucho más común de lo que se puede suponer pero con efectos demoledores en los estudios superiores porque hacen uso de estas fuentes no confiables para el aprendizaje. Por esa razón los investigadores (Hicks et al, 2024) denominan mentiras y simples estupideces a una parte considerable de los productos generados por la a IA Gen.

A menudo intentar diseñar un algoritmo para producir un resultado preciso e imparcial implica pedirle que haga dos cosas contradictorias. Si precisión significa reflejar a la sociedad, entonces el resultado exacto debería estar sesgado porque la sociedad está plagada de prejuicios y sesgos al márgen de la información falsa y errónea. La producción algorítmica sin sesgos está creando una imagen falsa de la realidad, como una historia social en la que se han eliminado todos los asesinatos y saqueos. Por ejemplo, si le pedimos a un algoritmo que prediga la reincidencia, es difícil imaginar cómo puede evitar reflejar el sesgo porque la reincidencia se ve afectada por el propio sesgo de forma recursiva pero precisa.

Los algoritmos utilizados para las decisiones de contratación y la calificación crediticia también suelen estar sesgados (O`Neil, 2017) porque estas materias -como casi todas las sociales- están sesgadas y los datos también lo están. Por tanto, un algoritmo a menudo no puede producir una decisión precisa e imparcial; podría ser capaz de producir una u otra. Pero tomar decisiones imparciales es muy difícil porque el sesgo vive en cada grieta y rincón de la sociedad y se correlaciona con muchos tipos diferentes de datos de forma granular y transversal. Una sociedad libre de prejuicios es, sencillamente, una ficción que no se deriva de los datos sesgados en los que se entrenan los algoritmos de IA Gen.

Podríamos valorar las decisiones libres de sesgos en lugar de las que se derivan con precisión de los datos de entrenamiento. Después de todo, uno de los propósitos del uso de algoritmos es mejorar la toma de decisiones humana. Pero esto demuestra que es posible que en realidad no queramos una decisión "exacta"; queremos una decisión ideal. Las clasificaciones se distorsionan de manera intensa. Se puede reducir una obra de Shakespeare a un simple resumen de la trama, pero esto elimina la parte más importante: el arte y la belleza del lenguaje. La historia puede reducirse a fechas y acontecimientos, pero eso no la hace útil ni esclarecedora y, desde luego, en ningún sentido relevante para la educación superior y para la ciencia. En ese sentido (Messeri y Crokett, 2024) se centran en los riesgos epistémicos (es decir, los riesgos asociados con la forma en que se produce el conocimiento) que la IA Gen podría representar para las ciencias y por lo tanto para el sistema que las produce, señalan los autores que si bien la IA produce beneficios en términos de velocidad y eficiencia, si se utiliza sin pensamiento crítico, su uso podría limitar inadvertidamente la diversidad de perspectivas involucradas en la producción de datos, conocimientos y teorías científicas. La ciencia sólida o solvente proviene de un conjunto diverso de conocedores de la misma, no sólo experiencialmente diversos, sino también cognitivamente diverso, disciplinariamente diverso. Si la IA Gen se convierte en esta herramienta por la que se transmite todo, se corre el riesgo de limitar el tipo de preguntas formuladas y el tipo de perspectivas que se aplican a un problema científico.

Por supuesto, las personas quieren el camino fácil, pero la vida no es fácil ni es sencilla. El botón fácil es una ilusión. Desafortunadamente, muchas herramientas algorítmicas se utilizan de esta manera. Funcionan para reducir la riqueza de la vida a ciertos elementos cuantificables. Los

algoritmos se centran en las correlaciones, no en la causalidad y rara vez quienes diseñan o utilizan el algoritmo preguntan por qué existen las correlaciones. A menudo no les importa la causalidad; la correlación hace el trabajo a costa de una simplificación contraproducente que para la educación superior que es inaceptable. La excesiva dependencia humana de la IA Gen puede disuadir facilmente de la reflexión crítica, el estudio sistemático profundo y el pensamiento sobre las relaciones causales. La IA Gen decide de manera diferente a los humanos como podemos observar.

8 Conclusión

La IA Gen como hemos tratado de exponer en este breve estudio, es una herramienta expresamente contextual a la que las multinacionales que impulsan esta tecnología han proyectado a la sociedad con la idea de que sus instrumentos de IA Gen son una especie de navajas suizas digitales, capaces de resolver innumerables problemas o reemplazar el trabajo humano intelectual en todo tipo de contextos. Pero si se aplican en los entornos equivocados, estas herramientas simplemente fallan y no se puede evitar que fallen. Incorporar estas tecnologías a la educación superior supone asumir algunos o todos los riesgos señalados y lo sensato y en base al principio de precaución debería limitarse su uso para evitar un abuso indiscriminado que tiene la capacidad de dañar gravemente los sistemas educativos superiores, la institución Universitaria y, en suma, la calidad de la ciencia.

Este trabajo es resultado del proyecto de investigación «Educar en valores, construir ciudadanías», Ministerio de Ciencia e Innovación. Agencia Estatal de Investigación. Proyectos de Generación de Conocimiento 2021. Referencia: PID2021-127680OB-I00.

9 Bibliografía

Alkaissi, H., y McFarlane, S. I. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2) 2023. De recomendable consulta en: https://pubmed.ncbi.nlm.nih.gov/36811129/

Atkinson-Abutridy J, Grandes modelos de Lenguaje, Marcombo, Madrid, 2023

Ben G, Los defectos de las políticas que requieren supervisión humana de los algoritmos gubernamentales, 45 *Ley y seguridad informática* Rev. 1,2022.

Berzal F, Redes Neuronales & Deep Learning, Granada, 2019

Burke, P, Ignorancia. Una historia global, Alianza Ed, Madrid, 2023

Carr, N, *Atrapados. Cómo las máquinas de apoderan de nuestras vidas*, Taurus, Madrid, 2014

Carr, Nicholas, *"¿Qué está haciendo Internet con nuestras mentes? Superficiales*", Taurus, Barcelona, 2020

Edelson, M, y otros, Following the Crowd: Brain Substrates of Long-Term Memory Conformity, *Science*, 333, 2011

Emsley, R. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophr* 9, 52, 2023. https://doi.org/10.1038/s41537-023-00379-4

Garnier-Brun, J, M. Benzaquen y J.P. Bouchaud, "Unlearnable Games and "Satisficing" Decisions: A Simple Moldeo for a Complex World", arvix.org 2312.12252v2 5 de enero de 2024. https://link.aps.org/doi/10.1103/PhysRevX.14.021039

Guía para el uso de IA generativa en educación e investigación. https://unesdoc.unesco.org/ark:/48223/pf0000389227

Grant, R W, *Los hilos que nos mueven. Desenmarañando la ética de los incentivos*, Avarigani, Madrid, 2021

Glauner, Patrick, Petko Valtchev y Radu State, "Impact of Biases in Big Data", *Actas del 26º Simposio Europeo sobre Redes Neuronales Artificiales, Inteligencia Computacional y Aprendizaje Automático* (ESANN 2018). arXiv: 1803.0089

Haidt, J, *La generación ansiosa. Por qué las redes sociales están causando una epidemia de enfermedades mentales entre nuevos jóvenes*, Deusto, Barcelona, 2024

Hicks, M,T, J Humphries, y J, Slater, ChatGPT is bullshit, *Ethics and Information Technology*, 2024. https://doi.org/10.1007/s10676-024-09775-5

Horwitz, J, *Código roto. Manipulación política, fake news, desinformación y salud pública*, Ariel, Barcelona, 2024.

Inteligencia artificial y educación: guía para las personas a cargo de formular políticas. https://unesdoc.unesco.org/ark:/48223/pf0000379376

Inteligencia artificial y educación: una visión crítica a través de la lente de los derechos humanos, la democracia y el estado de derecho (2022) https://book.coe.int/en/education-policy/11333-artificial-intelligence-and-education-a-critical-view-through-the-lens-of-human-rights-democracy-and-the-rule-of-law.html

Jonas, H, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Herder, Barcelona, 2015, p.-16.

Kearns, M y A Roth, *El algoritmo ético. La ciencia del diseño de algoritmos socialmente éticos*, La Ley Wolters Kluwer, Madrid, 2020

Keita Kurita y otros, "Measuring Bias in Contextualized Representations", *1st ACL Workshop on Gender Bias for Natural Language Processing* 2019. Puede verse en: Cornell *University https://arxiv.org/abs/1906.07337v1*, p. 31.

Khaneman, D, O Sivony y C.R. Sunstein, Ruido. Un fallo en el juicio humano, Debate, Barcelona, 2021

Kranzberg, M, Kranzberg's Laws, *Technology and Culture*, jul, 1986, Vol. 27, No. 3, p. 547.

Larson, E J. *El mito de la inteligencia artificial. Por qué las máquinas no pueden pensar como nosotros lo hacemos*, Shackleton, Barcelona, 2022.

López de Matarás, Ramón, El traje nuevo de la inteligencia artificial, *Investigación y Ciencia*, nº 526 Julio 2020.

Matute, H, "Ilusiones y sesgos cognitivos", *Investigación y Ciencia*, noviembre, 2019.

Maher C, *Mentes vegetales, una defensa filosófica*, Bauplan, Madrid, 2022

Messeri L, Crockett M J, Artificial intelligence and illusions of understanding in scientific reserarch, *Nature*, 627, 2024

McLuhan, M, *Comprender los medios de comunicación. Las extensiones del ser humano*. Paidós, Barcelona, 2009

Nguyen A, Ngo HN, Hong Y, Dang B, Nguyen BT. Ethical principles for artificial intelligence in education. *Educ Inf Technol* (Dordr). 2023;28(4):4221-4241. doi: 10.1007/s10639-022-11316-w. Epub 2022 Oct 13. PMID: 36254344; PMCID: PMC9558020.

Nussbaum, M. *Paisajes del pensamiento. La Inteligencia de las emociones*. Paidós, Barcelona, 2008.

O`Neil C, *Armas de destrucción matemática. Como el Big Data amenaza la desigualdad y amenaza la democracia*, Capitán Swing, Madrid, 2017, p. 175.

Pohl, Rüdiger F, *Cognitive Illusions. A Handbook of Fallacies and Biases in Thinking, Judgment and Memory*, Psychology Press, Taylor & Francis Group, New York, 2004, p. 40.

Recomendación sobre la ética de la inteligencia artificial. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa?posInSet=2&queryId=a2836fbc-3a67-4c9c-b414-3b1bd9998da0

Roselli, D.; Matthews, J.; Talagala, N. Managing Bias in AI. *In Proceedings of the Companion Proceedings of the 2019 World Wide Web Conference*, San Francisco, CA, USA, 13–17 May 2019, pp. 539-544.

Solove, D J. y Matsumi, H, AI, Algorithms, and Awful Humans. 96 *Fordham Law Review*, 16 de octubre de 2023.

Sunstein C. R y R H. Thaler, *Un pequeño empujón*, Taurus, Madrid, 2009, pp. 255-272.

Sharot, Tali, Christof W Korn y Raymond J Dolan, How unrealistic optimism is maintained in the face of reality, *Nat Neurosci*, 14 (11), May 1, 2012, p. 6.

Vicente, L., Matute, H. Humans inherit artificial intelligence biases. *Sci Rep* **13**, 15737, 2023.

Vicente, L, Matute, H, Humans inherit artificial intelligence biases. *Sci Rep* **13**, 15737, 2023.

Xu Z, Jain S, M Kankanhalli, Hallucination in Inevitable: An Innate Limitation of Large Language Models, *arXiv:2401.11817v1* [cs.CL]