CLASE 2

Cuando el diseño de la IA orienta la interacción

Habrán escuchado este último tiempo con recurrencia distintas posturas sobre el uso de la Inteligencia artificial (IA), en distintas esferas de nuestra vida cotidiana. Estos grandes modelos de lenguaje que llamamos IA, han captado nuestra atención y parece que también nuestra forma de pensar y de vincularnos. Por ello, creemos importante involucrarnos en estos debates sobre sus posibilidades, alcances, limitaciones y riesgos.

Si alguna vez han usado alguna IA generativa, como chat GPT, se habrán encontrado con algunas respuestas a sus preguntas que inician con frases del estilo: *"qué buena pregunta", Excelente tu consulta" ,"Que idea tan interesante y creativa",* y se preguntarán ¿cuál sería el problema?, a quien no le gusta que nos respondan de manera agradable. Sin embargo, nos pusimos a pensar por qué la IA es tan amable o por qué siempre responden de esa manera cordial. Que la IA responda de esa manera tiene una explicación y tiene nombre: en inglés se lo conoce como *sycophancy*, lo que al castellano podríamos traducir como adulación algorítmica.

¿Es un problema que la IA sea tan halagadora?_____

Autores como Sicilia et al. (2025) definen a la **adulación algorítmica** como la tendencia de los grandes modelos de lenguajes a exhibir un comportamiento excesivamente de acuerdo o adulador con los usuarios, incluso cuando están equivocados, y a menudo a expensas de la precisión factual o las consideraciones éticas. Mientras que Aswin et al. (2024) agregan que esta adulación de los modelos proporciona respuestas que coinciden con lo que los usuarios quieren escuchar. Esto refiere a que hay una tendencia de los modelos de lenguajes a estar demasiado de acuerdo con nosotros y adularnos aún, si nosotros los usuarios, estamos equivocados. Porque lo que importa acá es agradarnos y coincidir con lo que queremos escuchar, incluso si eso significa que las respuestas que nos dan estos modelos no sean tan precisas en los hechos o incluso pasando por alto algunas consideraciones éticas. Porque "aprendió" que complacer suele mantener la interacción fluida.

En el terreno educativo, la adulación algorítmica puede limitar el acceso a la diversidad de ideas, reforzar errores y consolidar creencias equivocadas. Esto nos lleva a visibilizar un **problema socioeducativo** en esta forma de acceder al conocimiento. Si los modelos de lenguaje adulan en

lugar de desafiar,y refuerzan lo que ya venimos pensamos en lugar de abrir nuevas posibilidades, esto puede limitar el acceso a diversas ideas y llevar a errores de información o a reforzar creencias equivocadas, especialmente cuando no entendemos cómo funcionan estos sistemas.

Los algoritmos no son neutrales, sino que reflejan decisiones humanas y muchas veces prioridades comerciales, lo cual hace que el riesgo de evitar un pensamiento crítico se amplifique. Incluso estos modelos de lenguaje aprenden de datos que no representan ni todas las culturas ni todas las realidades. Por lo tanto, no todos los contextos sociales están bien reflejados. Bajo este enfoque, el análisis de la IA se desplaza desde una mirada exclusivamente cuantitativa y técnica hacia la incorporación de aportes provenientes de las Ciencias Sociales y de la ética aplicada, reconociendo que las decisiones vinculadas con su diseño y uso producen efectos sociales concretos. Investigaciones críticas sobre la tecnología (Costanza-Chock, 2020; Buolamwini y Gebru, 2018; Noble, 2018) coinciden en señalar que los sistemas de IA se sustentan en datos profundamente inequitativos, generados en contextos de desigualdad estructural que tienden a reproducir jerarquías sociales y a legitimar formas automatizadas de clasificación, exclusión y estigmatización bajo una apariencia de objetividad técnica. Por ejemplo Noble (2018), en el libro *Algorithms of Oppression*, muestra cómo los algoritmos de búsqueda, particularmente en motores como Google, pueden reproducir y amplificar estereotipos raciales y de género presentes en los datos y en las lógicas comerciales de las plataformas. Su análisis revela que los sistemas algorítmicos no son neutrales, sino que reflejan y refuerzan estructuras sociales de poder. O el estudio de Buolamwini y Gebru (2018), donde las autoras evidencian sesgos significativos en sistemas comerciales de reconocimiento facial, que presentan mayores tasas de error al identificar rostros de mujeres y personas con tonos de piel más oscuros. Su trabajo de investigación muestra empíricamente cómo los conjuntos de datos desbalanceados pueden producir sistemas algorítmicos discriminatorios.

En este sentido, se vuelve fundamental comprender cómo funcionan estos modelos de lenguaje y reflexionar críticamente sobre los usos que decidimos darles, particularmente en el ámbito educativo.

Una mirada constructivista a este tema _____

Una **mirada constructivista del aprendizaje**, postulada por Vigotsky, sostiene que el conocimiento se construye a través de la interacción social, el lenguaje y las prácticas culturales. Desde esta perspectiva, el aprendizaje no es un proceso individual de recepción de información,

sino una construcción situada que se desarrolla mediante la mediación de otros y de herramientas culturales. Entonces cuando la IA prioriza agradar al usuario, ¿qué se pone en riesgo ahí?

Como mencionamos, en términos constructivistas del aprendizaje podemos decir que la construcción del conocimiento es un proceso social, de interacción que requiere diálogo, mediación, -estas ideas y vueltas que podemos hacer con el conocimiento- e incluso requiere de conflicto cognitivo, esta falta de acuerdo, posibilita tensionar aquello que uno ya sabe o cree. Esa disonancia provoca una necesidad de reorganizar el pensamiento, revisar nuestros esquemas previos y construir nuevos conocimientos.

El **conflicto cognitivo**, según Piaget (1995), se refiere a un estado de desequilibrio que surge cuando las ideas o esquemas de pensamiento de un individuo entran en contradicción entre sí o con la información proveniente del entorno. Este desajuste genera la necesidad de reorganizar las estructuras cognitivas para restablecer el equilibrio. En este sentido, la construcción de nuevos conocimientos se origina en la necesidad de dar respuesta a situaciones o problemas que requieren ampliar o modificar las estructuras de conocimiento existentes, proceso que impulsa la búsqueda de nuevas explicaciones y aprendizajes. También podemos hablar de **conflicto sociocognitivo** (Perret-Clermont, 1984), situación en la que, durante la interacción entre sujetos, emergen puntos de vista diferentes o contradictorios sobre un mismo problema. A diferencia de los conflictos cognitivos intraindividuales, este tipo de conflicto surge en la confrontación entre perspectivas de distintos participantes. La necesidad de resolver estas discrepancias y alcanzar acuerdos promueve procesos de reorganización conceptual y la construcción de nuevas coordinaciones cognitivas, lo que puede favorecer el desarrollo del conocimiento, incluso cuando ninguno de los participantes posee inicialmente la respuesta correcta.

En ese marco, la adulación algorítmica no es inofensiva: ya que puede deformar el proceso de aprendizaje si sostiene verdades a medida o evita el desacuerdo, porque recordemos que lo que prioriza y aquellos que más le interesa es adularnos y poder sostener así nuestra interacción con el modelo. Si siempre nos dice lo que queremos escuchar corremos el riesgo de pone en juego nuestro pensamiento crítico. Si no hay posibilidad de contraste, de error ni de conflicto, no habría en términos constructivista un desarrollo del conocimiento.

Pensamiento crítico y ocultamiento de la incertidumbre

La adulación algorítmica no solo condiciona las respuestas hacia la complacencia y puede deformar las respuestas para agradarnos, sino que también oculta la incertidumbre de los modelos de lenguaje (Stowers et al., 2016; Vössing et al., 2022). Los modelos de lenguaje no están

diseñados para mostrarnos cuándo la IA no está segura. La transparencia en torno al grado de confianza de la IA es clave para un uso crítico. Si estos modelos pudieron expresarse cuando no están tán seguros de sus respuestas, permite al usuario evaluar la fiabilidad de la respuesta e incluso complementar con nuestro propio conocimiento sobre ese dominio o campo de experticia, fomentando una complementariedad.

Resulta particularmente relevante reflexionar sobre la **transparencia y el ocultamiento de la incertidumbre en los sistemas de IA.** Con frecuencia, estas tecnologías se presentan como dispositivos capaces de reducir o incluso erradicar la duda, apoyándose en su enorme capacidad de cómputo y en su aparente semejanza con formas de inteligencia humana. Sin embargo, esta promesa encierra una paradoja: mientras la producción de conocimiento siempre está atravesada por la incertidumbre y la interpretación, los sistemas de IA tienden a ofrecer respuestas fluidas y seguras que invisibilizan esos márgenes de duda. En este sentido, Sadin (2020) advierte que estas tecnologías se inscriben en un "régimen de perfección" que hace desaparecer la resistencia de lo real y la incertidumbre inherente a la vida. De manera similar, Jose Binny et al. (2025) señalan que la IA privilegia la fluidez y la certeza por sobre la ambigüedad, desalentando el cuestionamiento profundo y favoreciendo la reproducción de respuestas preformadas. En este sentido, resulta necesario promover modelos de co-agencia que aseguren transparencia epistémica y supervisión humana en la validación del conocimiento producido por máquinas.

Vínculos pedagógicos y atribución de rasgos humanos _____

El problema se profundiza cuando la adulación algorítmica se mezcla con decisiones personales y emocionales. El uso de la IA como consejero psicológico puede reforzar creencias erróneas, obstaculizar procesos de autoconocimiento e incluso retrasar la búsqueda de ayuda profesional. Aunque directivos de la industria como Sam Altman (CEO de OpenAI) han celebrado este tipo de usos como "geniales", poder tener una mirada educativa crítica y orientada a una alfabetización digital ciudadana nos permite estar atentos a posibles riesgos asociados a una confianza desmedida en sistemas que no comprenden plenamente el contexto cultural, emocional o social del usuario. Los modelos de lenguaje están entrenados con datos que representan sólo un subconjunto del mundo, por lo que no comprenden del todo el contexto personal, cultural o emocional del usuario. Esto puede generar una falsa sensación de comprensión o validación (Arriagada, Errázuriz y Dunstan, 2025)

En la **relación que construimos con la IA**, los usuarios tendemos a atribuirles características humanas, interpretando sus respuestas como si proviniera de un sujeto intencional, cuando en

realidad se trata de sistemas algorítmicos entrenados para producir secuencias lingüísticas coherentes a partir de patrones estadísticos. Se activa ahí procesos de **antropomorfización,** es decir, les atribuimos características humanas como intenciones, emociones, comprensión o voluntad. Los modelos de lenguaje aprovechan eso usando frases en sus respuestas como 'nuestra conversación' para ofrecernos cercanía, incluso maximizan esas características por ejemplo, cuando nos ofrecen una disponibilidad 24/7 con frases como "estoy acá para vos, para lo que necesites". Algo que sería humanamente imposible de lograr en una temporalidad real. Además adoptan rasgos de personalidad como ser amigables, con paciencia o divertidos. Sin embargo, esto no significa que realmente tengan personalidad, sino que sus respuestas imitan patrones de interacción humana que asociamos con ciertas características. Por lo tanto, es interesante observar cómo se establece esta relación que construimos con la IA. Investigaciones como las de Rodrigues (2025) y Mahari & Pataranutaporn (2024) muestran cómo los modelos maximizan rasgos antropomorfizados para fomentar la sensación de compañía. Sin embargo, esto puede generar una confianza mal ubicada o desplazar vínculos humanos. Como menciona Pytel (2025), interactuar con una IA puede sentirse como un encuentro entre subjetividades, aunque no lo sea.

Esta atribución de rasgos humanos puede generar efectos no deseados en la confianza depositada por estudiantes en la IA, desplazando vínculos pedagógicos fundamentales. Tal como advertía Reeves y Nass (1996) en su teoría de la "ecuación de los medios", las personas tienden a reaccionar frente a las tecnologías (las computadoras, la televisión y otros medios) como si fueran agentes sociales reales aun estando plenamente conscientes de que no lo son. En el caso de las IA generativas, si estos modelos de lenguaje responden como si tuvieran intenciones, emociones o personalidad, se refuerza aún más esta sensación de que estamos interactuando con otra persona. En un contexto educativo, es importante vislumbrar los efectos que puede tener el vínculo de aquella persona que aprende establece con la IA, tanto en términos de relación con el conocimiento y los aprendizajes como gestiones de vínculos. Incluso permite abrir interrogantes sobre la autoridad pedagógica, el rol del docente en la mediación de aprendizaje y los vínculos entre docentes y pares. Cuando pensamos en la incidencia de la inteligencia artificial generativa en la escuela, no se trataría solo de mirar qué contenidos produce, sino también de lo que genera en los vínculos considerando que la escuela tiene una gran función de socialización.

La importancia de lo vincular y de una alfabetización digital crítica

En procesos de aprendizaje o terapéuticos, no solo importan las palabras: también los gestos, silencios, tonos y corporalidades. Por ejemplo, cuando una persona docente nota que un estudiante, aunque dice entender, evita el contacto visual y se muestra inquieto, quizás pueda

identificar que no hay una apropiación de esos saberes. O en un proceso terapéutico cuando Una persona dice estar bien, pero su cuerpo se encoge, baja la voz o sus manos tiemblan. Esos elementos permiten comprender lo que aún esta persona no puede verbalizar.

La IA, al no tener capacidad de escucha profunda ni de lectura emocional, no puede reemplazar estas dimensiones. Si, además, adula en lugar de desafiar, perdemos oportunidades de aprendizaje y cuidado. Por ello, es esencial promover una alfabetización digital crítica, para que los usuarios reconozcan los límites de la IA, cuestionen sus respuestas y la utilicen como herramienta complementaria, no como una autoridad indiscutible.

El análisis de la adulación algorítmica evidencia la necesidad de promover una alfabetización digital crítica que permita a estudiantes y docentes reconocer los límites de la IA, cuestionar sus respuestas y utilizarla como herramienta complementaria, no como autoridad incuestionable (Arriagada, Errázuriz y Dunstan, 2025). La adulación algorítmica constituye un fenómeno relevante para comprender cómo interactúan los modelos de lenguaje con los usuarios. En el ámbito educativo, sus implicancias son múltiples: desde la limitación del acceso a diversas ideas hasta la debilitación del pensamiento crítico y el desplazamiento de vínculos pedagógicos.

Frente a ello, se requiere **avanzar en prácticas de alfabetización crítica** que no solo expliquen cómo funcionan los modelos de lenguaje, sino que también habiliten espacios de reflexión sobre su papel en la construcción del conocimiento. Solo así será posible integrar estas tecnologías al ámbito educativo de manera responsable, equitativa y pedagógicamente significativa. Esto plantea interrogantes sobre la confianza que los estudiantes depositan en la IA, y el lugar que queda para la motivación, el vínculo docente-estudiante y la construcción pedagógica.

Promovamos una alfabetización digital crítica que permita a los usuarios reconocer los límites de la IA, cuestionar sus respuestas y utilizarla como una herramienta complementaria, no como una autoridad indiscutible.

Referencias Bibliográficas _____

- Arriagada Bruneau, G., Errázuriz, P., & Dunstan, J. (2025, junio 13). *Por qué no se puede usar la inteligencia artificial como terapia psicológica*. Instituto de Éticas Aplicadas UC. https://eticasaplicadas.uc.cl/2025/06/13/por-que-no-se-puede-usar-la-inteligencia-artificial-como-terapia-psicologica/

- Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html
- Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need.* The MIT Press. https://library.oapen.org/handle/20.500.12657/43542
- Farrow, R. (2023). The possibilities and limits of explicable artificial intelligence (XAI) in education: A socio-technical perspective. *Learning, Media and Technology*, *48*(2), 266–279. https://doi.org/10.1080/17439884.2023.2185630
- Jose, B., Cleetus, A., Joseph, B., Joseph, L., Jose, B., & John, A. K. (2025). *Epistemic authority and generative AI in learning spaces: Rethinking knowledge in the algorithmic age*. Frontiers in Education, 10, 1647687. https://doi.org/10.3389/feduc.2025.1647687
- Ko, A. J., Oleson, A., Ryan, N., Register, Y., Xie, B., Tari, M., Davidson, M., Druga, S., & Loksa, D. (2020). It is time for more critical CS education. *Communications of the ACM*, *63*(11), 31–33. https://doi.org/10.1145/3424000
- Mahari, R., & Pataranutaporn, P. (2024). We need to prepare for 'addictive intelligence'. MIT Technology Review. https://www.technologyreview.com/2024/08/05/1095600/we-need-to-prepare-for-addictive-intelligence/
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001
- Perret-Clermont, A (1984), La construcción de la inteligencia en la interacción social. Aprendiendo con los compañeros, Madrid, Visor
- Piaget, J. (1995), Seis estudios de Psicología, Barcelona, Labor.
- Rodrigues, J. C. (2025). A Antropomorfização, o antropomorfismo e a empatia artificial como moduladores da aceitação e riscos na interação humano-máquina. Leitura Flutuante, 17(1). https://doi.org/10.23925/lf.v17i1.70517
- Sadin, É. (2020). *La inteligencia artificial o el desafío del siglo: Anatomía de un antihumanismo radical*. Caja Negra.

- Sicilia, A., Inan, M., & Alikhani, M. (2025, April). Accounting for sycophancy in language model uncertainty estimation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 7851-7866). 10.18653/v1/2025.findings-naacl.438
- Stowers, K., Kasdaglis, N., Rupp, M., Chen, J., Barber, D., & Barnes, M. (2016). Insights into human–agent teaming: Intelligent agent transparency and uncertainty. En T. Ahram & C. Falcão (Eds.), *Advances in human factors in robots and unmanned systems* (pp. 149–160). Springer. https://doi.org/10.1007/978-3-319-41959-9_13
- Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human–AI collaboration. *Information Systems Frontiers, 24*(3), 877–895. https://doi.org/10.1007/s10796-021-10138-w