¿Es la ética un límite para la IA?

Johnny Godoy,
Camilo Carvajal Reyes y Felipe Urrutia[1]

El uso de técnicas de aprendizaje automático e inteligencia artificial no está exento de reflexión ética. Estas herramientas a menudo ofrecen soluciones aproximadas a problemas críticos. Incluso cuando las soluciones óptimas son posibles, los problemas pueden estar mal planteados debido a sesgos humanos. Ante esta situación, ¿cómo podemos garantizar una responsabilidad algorítmica adecuada al utilizar técnicas de aprendizaje automático e inteligencia artificial? ¿Están los/as estudiantes, investigadores/as y profesionales del área familiarizado/as con la responsabilidad algorítmica?

Son estas cuestiones las que han motivado la fundación de la Asociación de Ética en Datos e Inteligencia Artificial (AEDIA), impulsada como asociación estudiantil en la Universidad de Chile. La presente edición de Cuadernos de Beauchef pretende abordar estas problemáticas. Sin embargo, existen voces que plantean, directa o indirectamente, que las regulaciones y las consideraciones éticas son una traba para el progreso, en particular, de estas nuevas tecnologías. ¿Es la ética un límite para la IA? Para responder esta pregunta, es importante entender un poco qué hace que consideremos que un algoritmo sea considerado una «Inteligencia artificial».

[1] Miembros directivos, Asociación de Ética en Datos e Inteligencia Artificial de la Universidad de Chile. jdgod98@gmail.com, ccarvajal@dim.uchile.cl, furrutia@dim.uchile.cl

En general, una IA es un algoritmo que muestra un desempeño impresionante en una tarea que percibimos como difícil de realizar por una computadora: asistir a humanos a través de conversaciones (ChatGPT), generar imágenes similares a las de un artista (DALL-E), personalizar recomendaciones de música (Spotify), predecir los mejores candidatos para un trabajo (Manatal), encontrar los mejores resultados a una búsqueda (Google), jugar un juego de ajedrez a un nivel sobrehumano (Stockfish).

El proceso de «aprendizaje» es un meta-algoritmo que construye a la «inteligencia» de la IA, de tal forma que se maximice o minimice una métrica de interés, medida a través de simulaciones o del uso de datos históricos. Esta métrica tiene que ser cuantificable y debe estar alineada con la tarea por resolver. Comúnmente, la métrica que se debe maximizar es un tipo de similitud de los valores generados o predichos con respecto a los datos de entrenamiento.

Supongamos que estamos diseñando un algoritmo que sea capaz de jugar el videojuego Tetris de forma automatizada, *mejor* que cualquier humano. Tetris es un juego imposible de «ganar» y se trata de sobrevivir cuanto más se pueda hasta que las fichas caigan más rápido de lo que el jugador pueda eliminarlas. Una métrica que puede parecer razonable es el tiempo que el algoritmo juega, pues un mejor jugador es capaz de jugar más tiempo, pero hay que considerar otro factor: Si permitimos que pueda presionar el botón de pausa, ¡el algoritmo puede maximizar su métrica sin saber jugar Tetris, simplemente pausando y dejando al contador correr!

Si bien existen soluciones prácticas para este caso particular, el ejemplo refleja la importancia de diseñar una métrica de forma cuidadosa, lo que puede ser muy difícil, especialmente en casos en los cuales no entendemos a la perfección qué es lo que debemos optimizar. Por ejemplo, podemos considerar motores de ajedrez que evalúan qué tan buena es una posición en el tablero, combinando para ello valores como la cantidad de piezas de cada jugador, el número de movimientos que tienen disponibles, la seguridad del rey, entre otros; pero ni el mejor de los grandes maestros sabe una fórmula matemática para evaluar qué tan buena es una posición. Si lo supiera, el juego no tendría sentido.

Fuera de los ejemplos lúdicos, la ética no actúa como limitante de la ingeniería, sino que como guía para generar la métrica apropiada que permite resolver un problema. Encontrar los mejores candidatos para un puesto de trabajo puede suponer la búsqueda de aquellos que tengan características parecidas a las de candidatos exitosos. Sin embargo, una métrica cuyo criterio se basa en candidatos similares a los anteriores no está alineada con el objetivo de encontrar a los mejores, independientemente del buen desempeño que parezcan tener: este algoritmo perpetúa sesgos que existen en el proceso de contrataciones, tales como preferencias de género o edad mal guiadas, sin considerar el valor productivo positivo que trae un ambiente de trabajo diverso.

También existen otros casos donde las preocupaciones éticas tienen beneficios prácticos para los ingenieros. Por ejemplo, consideremos la siguiente pregunta de ética referente a los chatbots como ChatGPT:

¿Qué debe responder un asistente virtual si le pides instrucciones para crear una bomba?

Que ChatGPT rechace dar las instrucciones es una limitación para los creadores de bombas novatos, pero como problema de ingeniería, un *requisito* de un asistente virtual útil es ser uno que pueda ayudar a la humanidad, no que cree un ambiente inseguro.

Sin embargo, desde el lanzamiento de ChatGPT, se ha mostrado una gran cantidad de respuestas éticamente dudosas. Volviendo a las métricas, esto es de esperarse: el objetivo de aprendizaje de los modelos de lenguaje es generar un texto similar al que está escrito en los datos de entrenamiento obtenidos de la Internet. Sin embargo, que un chatbot sea capaz de hablar como un texto promedio de la internet, es algo que no coincide para nada con el objetivo de ser un asistente virtual útil. En este sentido, debemos definir una métrica que realmente busque medir la utilidad del asistente virtual, y esta tarea debe considerar la privacidad, seguridad, transparencia, justicia, igualdad, confiabilidad; todos aspectos que se estudian en la ética.

Matemáticamente hablando, imponer restricciones a un problema de optimización implica que el resultado de aquel ajuste será igual o no óptimo que el problema sin restricciones. Esta «penalización» es el argumento para quienes abogan por la llamada libertad al momento de optimizar una variada gama de objetivos, a partir de la premisa de que no es con la regulación que se alcanzará el mejor rendimiento. En particular, esta es la postura tomada hacia una IA sin restricciones para no «entorpecer» su desarrollo. Sin embargo, proponemos dos contraargumentos principales a esta postura:

El objetivo planteado no es siempre el correcto

La clave radica en cómo constantemente se asume la métrica por maximizar como fiel reflejo de la realidad deseada. La inteligencia artificial actual y sus defectos es un ejemplo más de una sociedad donde las premisas apenas alcanzan a correlacionarse con objetivos. Estos, a su vez, son muy probablemente imposibles de modelar con funciones parametrizables con un número razonable de elementos. ¿Acaso es el producto interno bruto de un país un reflejo absoluto de la felicidad de sus habitantes? ¿Son las horas de trabajo sinónimo de mayor productividad en cualquier contexto?

El considerar la ética como elemento a la hora de utilizar o no algoritmos automatizados puede efectivamente ser una restricción a la optimización de los parámetros de un modelo. Sin embargo, esta nos guía hacia el objetivo implícito de cualquier sistema que pretenda ayudar a la humanidad. Este no es un objetivo metrizable, pero, sin duda, comparte la dirección de maximizar aquellos valores que defendemos como sociedad desarrollada.

El objetivo de entrenamiento no es el objetivo productivo

Cuando una IA realiza el aprendizaje, esperamos que tenga el mejor desempeño para resolver el problema *en general*, no solamente *en los datos de entrenamiento*. Por ejemplo, supongamos

que queremos construir un sistema de detección de imágenes que distinga entre perros y gatos, y devuelva la respuesta al usuario. Para esto, el algoritmo ve un conjunto de imágenes que ya fueron etiquetadas por humanos como «perro» o «gato». Pero, en realidad, no queremos un algoritmo que sea bueno en detectar las imágenes que ya vio en entrenamiento (sabemos si tienen perros o gatos), nos interesa qué tan bueno es detectando imágenes que no usó en su entrenamiento, es decir, qué tan bien *generaliza*.

Sin embargo, entrenar una IA implica encontrar los parámetros que maximicen su desempeño en el conjunto de entrenamiento. Esto comúnmente causa un fenómeno conocido como *sobreajuste*, en el cual la IA se ajusta demasiado bien a los datos de entrenamiento, incluyendo los que sean excepciones en vez de reglas.

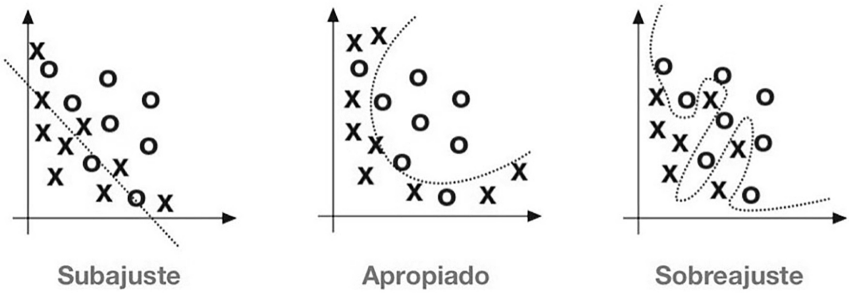

Figura 1: Distintos problemas de ajuste para una IA que busca separar el espacio entre los puntos X y O. En el sobreajuste, la curva es demasiado compleja y con el propósito de no cometer errores, se adapta a los elementos excepcionales en vez de a los comunes.

En nuestro ejemplo, un modelo sobreajustado puede ser perfecto para distinguir perros y gatos del conjunto de entrenamiento, sin realmente haber capturado los patrones importantes para distinguirlos en general: solamente «memoriza» lo que distinguía a cada imagen en particular.

Nuestra manera de combatir el sobreajuste es sancionar de alguna forma que el modelo se ajuste *tanto* a los datos de entrenamiento, utilizando para ello distintas penalizaciones de forma intencional, tales como recompensar a modelos que se alejen un poco de los datos de entrenamiento, pero que compensen este déficit siento más pequeños o que incorporen de mejor manera creencias previas que uno tenga sobre el problema.

En este sentido, usar criterios éticos para penalizar un modelo puede tener efectos positivos, aun si es que el modelo ya no es «el mejor» en los datos de entrenamiento. Como ejemplos de penalizaciones usadas en la realidad:

1. Si queremos un modelo interpretable, es importante que la cantidad de parámetros no sea demasiado masiva como para ser entendido. Penalizaciones de tamaño ocurren de forma explícita (como en regresión LASSO) o implícita (como en árboles de decisión que ejecutan hasta un límite de tamaño), o hasta se realiza postprocesamiento de redes neuronales para anular sus pesos; todos estas son prácticas comunes para reducir el sobreajuste.

2. En un trabajo reciente del Centro de Modelamiento Matemático, se utilizaron modelos de lenguaje para clasificar texto clínico en Chile, y verificaron que anonimizar los datos de los pacientes mejoraba el desempeño del modelo fuera de los datos de entrenamiento, pues en vez de «distraerse» por los datos personales del paciente, capturaba patrones que eran más relevantes para la tarea. Entonces, el modelo entrenado en datos anonimizados era capaz de generalizar mejor, a pesar de haber sacrificado desempeño en los datos originales.

Los textos presentes en este número son evidencia de cómo la ética debe ser abordada con perspectiva crítica y con herramientas que toquen tanto las disciplinas matemáticas como las ciencias naturales y, desde luego, las humanidades. *Análisis exploratorio de juicios morales en la discusión de dilemas* aborda el uso de modelos de aprendizaje de máquinas para examinar automáticamente las respuestas de estudiantes de la Facultad de Ciencias Físicas y Matemáticas (FCFM) ante dilemas éticos, para apoyar a los equipos

docentes en analizar las respuestas a nivel general. Esta metodología proporciona una visión amplia de las respuestas y sus conceptos empleados, contribuyendo a la evaluación de la competencia ética de estudiantes. De hecho, este artículo demuestra que los y las estudiantes no incluyen explícitamente los principios éticos en sus justificaciones, lo que levanta alertas sobre las medidas que pedagógicamente es necesario tomar para mejorar el desarrollo de la competencia.

Complicaciones y complejidades de convivir con decisiones tomadas por modelos de IA presenta algunas de los problemas que surgen automatizar decisiones con modelos de IA que no sean interpretables, es decir, cuyas decisiones no pueden ser fácilmente entendidas por un humano. El trabajo propone una definición de la interpretabilidad de un modelo a través de la teoría de la complejidad computacional y exploran resultados que reflejan un dilema importante: parece existir una oposición inherente entre interpretabilidad y precisión. El trabajo concluye argumentando por qué es importante lidiar con este dilema, advirtiendo, asimismo, acerca de la necesidad de regulaciones legales y cuestionamiento de parte de las ingenieras e ingenieros al implementar estos sistemas para entender el grado de interpretabilidad y precisión necesarios teniendo a la vista el impacto de las decisiones tomadas por los modelos.

Un mundo nuevo descubierto a través de los datos explora la creciente relevancia de la inteligencia artificial (IA) y la ciencia de datos en diversas disciplinas, y examina cómo esto llevó a emplear a astrofísicos para el análisis de datos en el fútbol. La autora describe su transición de la astronomía a la industria de datos, señalando las diferencias metodológicas y la necesidad de adaptarse continuamente a los avances tecnológicos. Además, advierte que, debido a la necesidad de resultados rápidos, a menudo se pasan por alto los sesgos y las cuestiones éticas en las soluciones de IA. Muy interesante es la reflexión que realiza sobre la formación ética recibida en su licenciatura y la importancia de considerar estos aspectos en el trabajo con IA, ejemplificando sesgos observacionales y de género que impactan en la interpretación de datos y la toma de

decisiones. En este mismo sentido, enfatiza la responsabilidad de los profesionales de IA de evaluar el impacto de sus desarrollos en la sociedad, instándoles a considerar las consecuencias reales de sus contribuciones.

Propuesta de modelo para la formación ética mediante la discusión de dilemas morales propone un modelo de formación ética basado en el estudio de dilemas, que basado en la reflexión crítica y en el diálogo en equipos, contribuye a la construcción del juicio moral. El modelo ofrece la estructura de un dilema posible, así como también orientaciones para el trabajo con estos, enfatizando la necesidad de ponerse en el lugar de la persona que toma una decisión. El ejemplo propuesto se refiere a la aplicación de algoritmos producidos por inteligencia artificial, aplicación que, al estar situada en contextos sensibles, demanda la reflexión moral. Se trata de un ejercicio valioso que enfrenta a los sujetos en formación, por medio del dilema, a pensar sobre las propias decisiones y a integrar nuevos criterios que enriquezcan sus juicios.

Recomendaciones para una IA responsable aborda las propiedades esenciales de los sistemas de inteligencia artificial (IA) y ofrece recomendaciones para su análisis, enfatizando la necesidad de una IA responsable y ética. Además de poner en cuestión términos como "IA confiable" e "IA ética", el texto propone un marco de análisis para algoritmos de la familia de aprendizaje de máquinas. Este aborda 30 propiedades (incluyendo conceptos como responsabilidad y transparencia) desde el punto de vista de la aplicación e impacto. El texto incluye, además, criterios para un adecuado control de calidad de las distintas etapas que componen el desarrollo de un algoritmo de IA.

Lengua, computadoras y emociones: interdisciplinariedad en la era de los Large Language Models (LLMs) se adentra en el mundo de las emociones como materia prima, en la actualidad, de modelos de IA, para los cuales son fundamentales pues permiten un correcto modelamiento de las personas. No obstante, el texto se pregunta ¿cómo afecta el paradigma que se tenga de las emociones? ¿Qué refleja la inteligencia artificial (IA) al analizar las emociones según

género? ¿Qué consecuencias éticas negativas puede tener una IA empática?

Usuarios de IA generativa responsables de obras mal atribuidas a grandes artistas se aproxima al impacto de la inteligencia artificial generativa en la creación artística, destacando cómo esta tecnología permite a cualquier persona producir obras que imitan el estilo de grandes artistas, lo que plantea serias preocupaciones sobre la atribución y la reputación de los creadores originales. El texto discute las posibilidades de replicar elementos estéticos de los artistas tanto en pinturas como en música a través de ejemplos. Además, aborda la necesidad de que los usuarios de IA sean responsables al publicar sus creaciones, en particular, poniendo sobre la palestra las limitaciones de los resguardos que se observan actualmente. Además, el texto se refiere al potencial de la IA para transformar la producción audiovisual, ámbito en el cual la creación de contenido casi indistinguible del de realizadores humanos hace que sea esencial un equilibrio entre la capacidad de generación y el uso responsable de estas nuevas tecnologías.

Lavender: la máquina de IA que dirige los bombardeos de Israel en Gaza explora el uso del sistema automatizado *Lavender* en operaciones militares, destacando cómo la inteligencia artificial ha transformado la identificación de objetivos para ataques aéreos, reduciendo drásticamente el tiempo de verificación humana a solo segundos. Debido a que la supervisión humana es mínima y a menudo formal, sin mecanismos efectivos para corregir errores, el uso de este sistema lleva a frecuentes identificaciones erróneas y bajas civiles. La presión militar por generar más objetivos y una definición amplia de lo que constituye un militante en el contexto del actual conflicto en Oriente Próximo agravan esta situación. Aunque se utilizan herramientas automatizadas para estimar daños colaterales y datos en tiempo real de móviles en Gaza, estos métodos no son infalibles, resultando en numerosos ataques y víctimas civiles. Precisamente por las características del fenómeno que describe, este texto es una invitación a reflexionar sobre el impacto real del uso de sistemas automatizados que deciden sobre las vidas de las personas.

Estrategia militar e inteligencia artificial: algunas consideraciones éticas, examina el creciente rol de la inteligencia artificial en la formulación de estrategias militares, subrayando sus implicaciones y la relación con la ética en ingeniería. Este texto destaca cómo la IA puede mejorar los procesos de decisión militar y la gestión de sistemas complejos, pero también plantea preocupaciones sobre la responsabilidad, la transparencia y el sesgo en los sistemas de IA. Ofreciendo una perspectiva histórica y aportes internacionales, esta propuesta especula sobre el futuro de la estrategia militar en la era de la IA, señalando marcos para prácticas éticas y destacando la necesidad de la investigación continua y enfoques interdisciplinarios, que permitan explorar en particular los aspectos éticos de la inteligencia artificial en contextos estratégicos militares.

Consideraciones tecnoéticas del uso de inteligencia artificial generativa de imágenes en procesos de restitución de identidad de personas desaparecidas aborda los riesgos sociales y éticos del uso de IA generativa en la búsqueda de personas perdidas. El texto destaca problemas de discriminación, uso indebido y desinformación, dividiendo los riesgos en categorías como exclusión, creación de contenido perjudicial y generación de imágenes engañosas. Subraya la necesidad de un uso ético y responsable de la tecnología para evitar impactos negativos y garantizar el bienestar de todos. Además, destaca los desafíos y posibles consecuencias dañinas de utilizar IA generativa sin las debidas precauciones en la restauración de identidad. Desde esta perspectiva, rescata la importancia de la transparencia en los métodos y la educación de las familias sobre las limitaciones de estas tecnologías, así como del marco ético internacional que protege los derechos y la dignidad de las personas.

Finalmente, *Humanidad y tecnología: reflexionando con ChatGPT sobre la ética de la inteligencia artificial en la medicina* nos propone una conversación con ChatGPT en la que, al ritmo del diálogo, el chat y la autora exploran la intersección entre la tecnología y la atención médica, destacando la importancia del trato humano en la medicina. Autora y máquina analizan el papel de la IA en el campo médico, sus beneficios y limitaciones, y conversan sobre las consideraciones éticas relacionadas con su uso, enfatizando la

necesidad de un enfoque responsable. La conversación subraya que la atención médica debe incluir empatía y apoyo emocional hacia los pacientes, no solo diagnósticos y tratamientos. También menciona la responsabilidad de los desarrolladores de crear interfaces claras y proporcionar información sobre riesgos y limitaciones, además de la necesidad de capacitar a las y los usuarios en el uso responsable de las recomendaciones de salud.

Como se puede apreciar, la variedad de asuntos es reflejo de cómo las reflexiones pueden y deben nacer desde variados ángulos y de que estas son fundamentales para un crecimiento responsable de las tecnologías que hoy se inscriben en la categoría de «Inteligencia Artificial».