



Cuadernos de Beauchef

Ciencia, Tecnología y Cultura

Volumen VIII - Número 1

LA ÉTICA COMO EL CENTRO DE LA IA

ethics

Estudios Transversales en Humanidades
para las Ingenierías y Ciencias

CUADERNOS DE BEAUCHEF

CIENCIA, TECNOLOGÍA y CULTURA

VOLUMEN VIII - NÚMERO 1

LA ÉTICA COMO EL CENTRO DE LA IA

Una publicación del área de Estudios Transversales en Humanidades para las
Ingenierías y Ciencias (ETHICS)
2024

ethics

© Una publicación del área de Estudios Transversales en Humanidades para las Ingenierías y Ciencias (ETHICS)

Escuela de Ingeniería y Ciencias - Facultad de Ciencias Físicas y Matemáticas
UNIVERSIDAD DE CHILE

© De esta edición:

Agosto 2024, Universidad de Chile
Santiago de Chile

ISSN: 2452-493X

<https://revistasdex.uchile.cl/index.php/cdb>

Impreso en Chile - Printed in Chile

Volumen VIII - Número 1

Director:

Andrés Monares

Editores:

María Torres

Editores invitados:

Johnny Godoy, Camilo Carvajal Reyes y Felipe Urrutia

Comité editorial:

Rubén Boroschek, Departamento de Ingeniería Civil, FCFM, Universidad de Chile

Ziomara Gertzen, Departamento de Ingeniería Civil Química, Biotecnología y Materiales,
FCFM, Universidad de Chile

Claudio Gutiérrez, Departamento de Ciencias de la Computación, FCFM, Universidad de Chile

Cecilia Ibarra, Centro de Ciencia del Clima y la Resiliencia (CR2) y Centro de Excelencia en
Geotermia de los Andes (CEGA)

Viviana Meruane, Departamento de Ingeniería Civil Mecánica, FCFM, Universidad de Chile

Pablo Ramírez, ETHICS, FCFM, Universidad de Chile

Claudia Rodríguez, ETHICS, FCFM, Universidad de Chile

Ximena Vergara, ETHICS, FCFM, Universidad de Chile

Diseño:

Marcos Andrés Pérez F.

Ediciones Eutópia Ltda.

www.eutopia.cl

Todos los derechos reservados:

Universidad de Chile

Avda. Beauchef 850, 3er. piso

Santiago de Chile

ÍNDICE

PRESENTACIÓN	7
Johnny Godoy, Camilo Carvajal Reyes y Felipe Urrutia <i>¿Es la ética un límite para la IA?</i>	9
LA ÉTICA COMO EL CENTRO DE LA IA	21
Josefa Cerda Maureira, Camilo Carvajal Reyes, Pablo Ramírez Rivas y Eduardo Hurtado Mila <i>Análisis exploratorio de juicios morales en la discusión de dilemas. Profundizando en la formación ética canalizada mediante la tecnología</i>	23
Bernardo Subercaseaux <i>Complicaciones y complejidades de convivir con decisiones tomadas por modelos de IA</i>	41
Camila Henríquez Beltrán <i>Un mundo nuevo descubierto a través de los datos</i>	63
Pablo Ramírez Rivas, Josefa Cerda Maureira y Jorge Martínez Silva <i>Propuesta de modelo para la formación ética mediante la discusión de dilemas morales</i>	75
Ricardo Baeza-Yates <i>Recomendaciones para una IA responsable</i>	99
Amanda Cercas <i>Lengua, computadoras y emociones: interdisciplinariedad en la era de los Large Language Models (LLMs)</i>	113

Javier Moyano	
<i>Usuarios de IA generativa responsables de obras mal atribuidas a grandes artistas</i>	123
Yuval Abraham	
<i>Lavender: la máquina de IA que dirige los bombardeos de Israel en Gaza</i>	135
Roberto Cristián Urrutia	
<i>Estrategia militar e inteligencia artificial: algunas consideraciones éticas</i>	155
Jorge Maldonado Soto	
<i>Consideraciones tecnoéticas del uso de inteligencia artificial generativa de imágenes en procesos de restitución de identidad de personas desaparecidas</i>	179
Macarena Mesa Maldonado	
<i>Humanidad y tecnología: reflexionando con ChatGPT sobre la ética de la inteligencia artificial en la medicina</i>	197
CLÁSICO	213
José Ortega y Gasset	
<i>Meditaciones de la técnica (fragmentos)</i>	215
Tabla de contenidos. Cuadernos de Beauchef. Volumen VII-2	226
Normas de publicación para Cuadernos de Beauchef	228

PRESENTACIÓN

¿Es la ética un límite para la IA?

Johnny Godoy,
Camilo Carvajal Reyes y Felipe Urrutia¹

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.

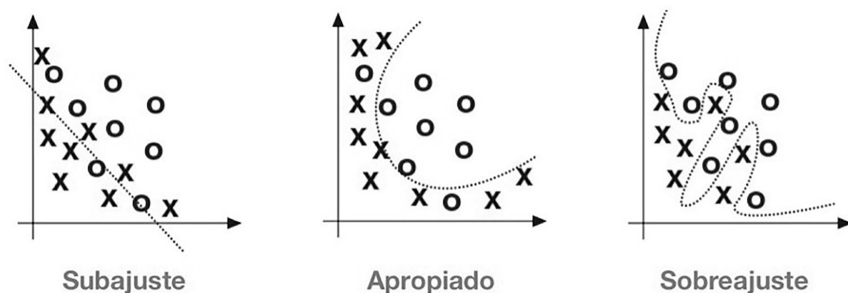


Figura 1: Distintos problemas de ajuste para una IA que busca separar el espacio entre los puntos X y O. En el sobreajuste, la curva es demasiado compleja, y con el propósito de no cometer errores, se adapta a los elementos excepcionales en vez de a los comunes.

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 siendo 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 el 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 algunos de los problemas que surgen al 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 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».

LA ÉTICA COMO EL CENTRO DE LA IA

Análisis exploratorio de juicios morales en la discusión de dilemas. Profundizando en la formación ética canalizada mediante la tecnología¹

Josefa Cerda Maureira, Camilo Carvajal Reyes,
Pablo Ramírez Rivas, Eduardo Hurtado Mila²

La Unidad de Ética (UE) del Área de Estudios Transversales en Humanidades para las Ingenierías y las Ciencias (ETHICS), perteneciente a la Escuela de Ingeniería y Ciencias (EIC) de la Facultad de Ciencias Físicas y Matemáticas (FCFM), tiene como propósito formar en competencias éticas al estudiantado. Una de las actividades que realiza con ese fin es la discusión en torno a dilemas éticos. El caso por estudiar se denomina Caso Adela y para ello se utiliza la aplicación *EthicApp*. El gran volumen de datos dificulta el análisis. Este trabajo aborda el uso de modelos de aprendizaje de máquina supervisados y no-supervisados (*Naive-Bayes* y *BERTopic*, respectivamente) para modelar la estructura textual de las respuestas, y con esto, apoyar el análisis que puedan hacer los equipos docentes. Pese al potencial de mejora, la metodología ofrece una visión general de las respuestas y conceptos utilizados, lo que aportará en la

¹ Agradecemos a Gustavo Zurita (FEN, UChile) y a Claudio Álvarez (UAndes) por facilitarnos el uso de la aplicación web *EthicApp*, además de su constante colaboración y paciencia para la realización de las actividades formativas. Agradecemos también a Claudia Rodríguez, Coordinadora de ETHICS, por acompañar y apoyar el trabajo de la Unidad de Ética. Finalmente, un agradecimiento a Ignacio González, coordinador del Área de Ingeniería e Innovación, por su compromiso y tiempo para desarrollar la competencia ética en los cursos de plan común.

² Josefa Cerda Maureira, Universidad de Chile, jozefacerda@uchile.cl; Camilo Carvajal Reyes, Universidad de Chile, ccarvajal@dim.uchile.cl; Pablo Ramírez Rivas, Universidad de Chile, pabramirez@uchile.cl; Eduardo Hurtado Mila, Universidad de Chile, eduardo.hurtado@ug.uchile.cl.

evaluación de la competencia ética de estudiantes de la FCFM. Los resultados obtenidos permiten concluir que los y las estudiantes no incluyen, explícitamente, los principios éticos en su justificación, pudiendo esto implicar que se encuentran en un nivel básico de la competencia; además, cuando la responsabilidad se enfrenta con el respeto (como ocurre en el Caso Adela), el estudiantado se inclinará por el principio de responsabilidad.

Introducción

La formación de la ética (Cortina, 2000) plantea una serie de interrogantes y desafíos: ¿cómo puede enseñarse un tipo de saber de naturaleza práctica en contextos masivos como el de la FCFM? ¿Cómo se evalúa y se plantea en indicadores de evaluación un desempeño que tiene lugar en tan diversos contextos? ¿Puede o debe la formación en ética hacerse cargo de todas las actuaciones de quienes son sujetos de enseñanza? ¿Cómo se forma «el carácter» de las personas hoy en la educación superior?, ¿se puede hacer? ¿Cómo pueden abordarse estas dificultades en el diseño, implementación y evaluación de un modelo de competencia ética exitoso en cuanto al aprendizaje de los y las estudiantes? La EIC ha tomado parte en estos desafíos y creó hace dos años la UE al interior de ETHICS, con el propósito de hacer frente a estas interrogantes con personas formadas y especializadas en el tema.

En el XXXIV Congreso de la Sociedad Chilena de Educación en Ingeniería, la Unidad de Ética presentó la ponencia «La formación ética canalizada mediante la tecnología: experiencia y resultados preliminares del uso de la herramienta web *EthicApp*» (cfr. Ramírez, P., Guerrero, S., Cerda, J., Ross, J. y Flores, G., 2022) en la cual se aborda un caso ético denominado «Caso Julieta», en una actividad de formación en una asignatura obligatoria de la carrera de ingeniería. Los principales resultados fueron: (1) el impacto que tiene el diálogo y la confrontación de ideas en el enriquecimiento de los propios juicios, puntos de vista y preferencias morales, (2) luego de la fase grupal del trabajo del caso, los y las estudiantes tienden a tomar posturas más moderadas que extremas, (3) el 60% de los y las

estudiantes mantienen la postura grupal de la etapa previa a la etapa individual final; y (4) la postura de cometer un acto deshonesto está basada en argumentos de juicio moral que integran elementos que van más allá del beneficio personal y que consideran la relación y corresponsabilidad con otros.

Este artículo pretende ser una continuación y profundización de la ponencia ya mencionada por medio de un análisis exploratorio más profundo y acabado de las justificaciones escritas. Debido a la naturaleza masiva de la formación en ingeniería, la cantidad de texto a analizar representa un desafío para la evaluación; por ello se utilizarán herramientas de procesamiento de lenguaje natural, en particular, modelos de aprendizaje de máquina supervisados y no-supervisados para modelar la estructura textual de las respuestas. Esto permitirá comprender con más detalle las elecciones de los y las estudiantes, la identificación de juicios morales y los principios éticos o elementos que consideran para discernir por una determinada postura.

El contexto

La Escuela de Ingeniería y Ciencias recibe cada año más de mil estudiantes en su programa de Plan Común, que dura cuatro semestres y es un paso obligatorio para elegir una especialidad o licenciatura. Durante los primeros tres semestres del programa de Plan Común, los y las estudiantes deben rendir los siguientes cursos en el orden señalado: Desafíos de innovación, Proyecto en innovación en Ingeniería y Ciencias y Módulo interdisciplinario. Estos cursos son parte de la línea de formación impartida por una unidad de la Escuela denominada Área de Ingeniería e Innovación que, entre otros objetivos, busca desarrollar las competencias transversales de innovación, trabajo en equipo, comunicación oral y escrita y compromiso ético.

Cada uno de estos cursos tiene diez secciones de cien estudiantes cada una, cantidad que representa un desafío tanto para los equipos docentes, en tanto formadores en competencias

específicas de las ingenierías y las ciencias, así como para las unidades a cargo de la formación y evaluación de las competencias transversales o genéricas. Por ello, la Unidad de Ética, a cargo del desarrollo de la competencia de compromiso ético, ha buscado distintas estrategias "algunas de ellas serán mencionadas ahora" para cumplir con el propósito de aprendizaje, para lo cual, debe levantar evidencias del desempeño de los y las estudiantes.

La competencia transversal de compromiso ético está dividida en dos momentos de acuerdo con el ciclo de formación del que se trate. En el caso del curso en que se aplicará el caso en cuestión "en adelante Caso Adela", corresponde la competencia transversal en su momento inicial: «Reflexionar sobre el propio actuar y sus consecuencias, en el marco de la honestidad, la responsabilidad y el respeto, buscando la excelencia y rigurosidad en su proceder en contextos académicos, en las relaciones interpersonales y con su entorno». Esta competencia se manifiesta a partir de una serie de indicadores de evaluación que se relacionan con los principios de responsabilidad, respeto e integridad en contextos profesionales.

A continuación, se comparte una síntesis del Caso Adela y sus dilemas morales distribuidos en etapas según diferenciales semánticos³:

En Chile, la deficiencia de vitamina D es un problema serio tanto en adultos mayores como en niños. Para abordar esta preocupante situación, un grupo de profesionales creó una startup para buscar una solución. Investigando, encontraron una fruta ancestral de las comunidades diaguitas con alta concentración de vitamina D y un sabor atractivo para el consumo. Adela, una ingeniera del equipo, diseña el proceso de producción de un nuevo alimento a partir de esta fruta. Sin embargo, se enfrenta a desafíos, ya que el árbol solo crece cerca de los ríos y necesita abundante luz solar, lo que dificulta llevarlo a zonas más australes; por lo tanto, debe cultivarse en el Norte Chico, región afectada por la sequía. Además,

³ Para más información sobre la metodología de dilemas morales ver sección correspondiente en ponencia anterior: <https://revistasdex.uchile.cl/index.php/cdb/article/view/12153/12192>

para conservar la vitamina D durante el transporte a otras zonas, el equipo decide liofilizar la fruta y agregar conservantes. Aunque aún no tienen la obligación legal de integrar a las comunidades diaguitas en el proyecto, Adela escucha sus preocupaciones sobre cómo estos cambios afectarían sus tradiciones. Pese a que los cambios son necesarios para ayudar a quienes sufren deficiencia de vitamina D, las comunidades prefieren mantener sus prácticas tradicionales, ya que estas son parte fundamental de su identidad.

Para el análisis de este caso, se trabaja en torno a dos diferenciales semánticos⁴ presentados a continuación: **DS1** - Considerando el uso de agua en un sector de escasez hídrica para la producción de este nuevo alimento y la potencial población beneficiada, Adela debería priorizar... : (1) Preservar el recurso natural escaso o (6) Beneficiar la salud de ancianos y niños; **DS2** - Considerando la futura población beneficiada por este nuevo alimento y los efectos generados a las tradiciones identitarias del pueblo originario del caso, Adela debería priorizar...: (1) Producir el alimento contra déficit vitamínico o (6) Resguardar las tradiciones identitarias. En el modelo de la Figura 1 se esquematiza lo anterior.

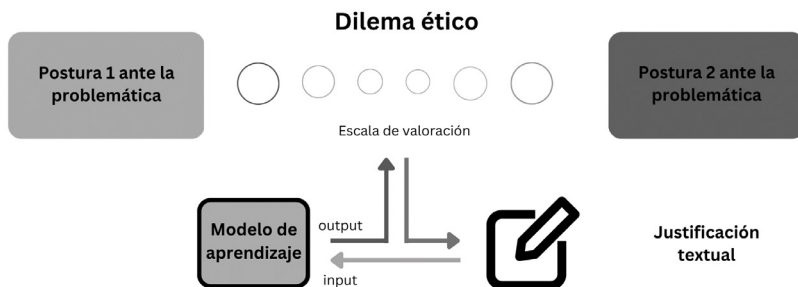


Figura 1. Esquema de modelo para predecir posturas de estudiantes.

⁴ Cada diferencial semántico tiene seis opciones dentro de las cuales los y las estudiantes pueden escoger según el grado de acercamiento que tengan con determinada postura. El número de opciones busca evitar posturas neutras.

El caso en cuestión ha tenido tres aplicaciones supervisadas por la UE de ETHICS. En el semestre de primavera del año 2021 se aplica en el tercer curso (CD2201) de la línea de cursos ya mencionados; en el semestre de primavera del año 2022, se aplica en el segundo curso (CD1201) y, por último, en el semestre de otoño del año 2023 se aplica en el primer curso (CD1100). En la Tabla 1 se muestran las realizaciones del caso, las etapas de cada una, el número de personas analizadas y el porcentaje de realización de la actividad. La modalidad de aplicación del caso se ejecuta en cuatro etapas: Etapa 0 (E0), cuya realización es previa al día de la actividad sincrónica que tiene como objetivo que los y las estudiantes creen sus cuentas de usuario/a en *EthicApp*; Etapa Individual 1 (EI-1) en la cual se lee el caso en clases y se responden los dos diferenciales semánticos; Etapa Grupal en la cual los y las estudiantes son asignados a grupos con criterio de heterogeneidad en sus respuestas y discuten grupalmente en torno a los dos diferenciales; por último, la Etapa Individual 2 (EI-2) en la que se vuelve a responder individualmente luego de la etapa grupal.

Tabla 1
Resumen de las realizaciones del caso Adela.

Año	Curso	Etapas	Número de personas analizadas	Porcentaje realización
2023-1	CD1100	E0, EI-1, EG, EI-2	811 de 866	94%
2022-2	CD1201	E0, EI-1, EG, EI-2	591 de 716	83%
2021-2	CD2201	EI-1, EG, EI-2	160 de 213	75%

Debido a los límites de extensión del presente artículo, en adelante se analizará únicamente el diferencial semántico 2 y su comportamiento en la EI-1, EG y EI-2.

Metodología

Para remediar la limitación de la gran magnitud de datos para el análisis de las justificaciones, se propone el uso de modelos para

procesamiento de texto que se usan en el contexto de la ciencia de datos y la inteligencia artificial. Estos modelos cuentan con una base estadística y permiten realizar predicciones y/o separaciones de los datos que se desean modelar, siendo estos los textos de estudiantes en este caso. El método que se propone consiste en utilizar estos modelos y ajustarlos a la tarea de predecir la valoración frente al dilema ético. Esto permitirá, con un análisis a posteriori, investigar qué elementos semánticos son los más relevantes a la hora de tomar una u otra postura, que es el objetivo final.

En el contexto del procesamiento de lenguaje natural se han estudiado ya las respuestas a preguntas o cuestiones morales. Muchos de estos trabajos toman como base la teoría de fundamentos morales (TMF) propuesta por Graham et al. (2013), siendo Garten, Boghrati, Hoover, Johnson y Dehghani (2016) y Xie, Hirst y Xu (2020), dos ejemplos. En el primer trabajo se busca la predicción/identificación de retóricas morales y, en el segundo, se evalúan modelos que también predicen e identifican retóricas morales. Por otro lado, Kennedy et al. (2021) buscan la predicción de preocupaciones morales propias de un individuo, usando evidencias de lenguaje moral que escriben usuarios y usuarias en redes sociales. Se utilizaron distintas técnicas de procesamiento de lenguaje para predecir los puntajes obtenidos por usuarios y usuarias, para cada una de las dimensiones morales planteadas en la TMF. Se destaca la variedad de métodos testeados para vectorizar texto, incluyendo *latent dirichlet allocation (LDA)* (Blei, Ng y Jordan, 2003), representaciones de palabras distribucionales (*word embeddings*), conteo de ocurrencias de conceptos relacionados a TMF y *BERTopic* (Devlin, Chang, Lee y Toutanova, 2019), que es un modelo de lenguaje profundo. Es este último el que obtiene mejores resultados. Finalmente, tanto conteos de ocurrencias de conceptos, como *LDA*, se usaron para interpretar qué elementos lingüísticos específicos explicaban cada fundamento por separado.

Los trabajos anteriores muestran la alta variedad de formatos en los cuales se ha evaluado la presencia y grado de categorías morales, así como también, la capacidad de distintos modelos de procesamiento de lenguaje natural para modelarlas. No obstante, ninguno de los artículos estudiados enfrenta un desafío tan específico

como el de este trabajo. Por ello, la modalidad consistirá en una exploración de características básicas de las justificaciones textuales "que se considerará el único elemento a utilizar", para luego usar modelos predictivos que modelen el texto. Por su parte, el valor por predecir corresponderá a la opción ante la problemática en cuestión. Este proceder busca obtener una fotografía global de las justificaciones de estudiantes, condicionadas a la preferencia escogida.

Cabe señalar, que tener un modelo que prediga la postura (concretizada por la opción escogida) dada en el texto, no es de interés *per se*. La ventaja es que, a través de tal herramienta, se identifiquen los elementos semánticos que justifiquen los argumentos dados por los y las estudiantes. Es por esto que se preferirán modelos que tengan capacidad interpretativa, porque de esta manera se sabe qué elementos del *input* (secuencias de palabras en este caso) hacen que el resultado sea uno y no otro. Estos elementos (dado un modelo bien ajustado a los datos) mostrarán, entonces, el comportamiento general de las justificaciones, facilitando así el análisis de la gran cantidad de textos que se generan en cada actividad.

El modelo supervisado *Naive-Bayes*, que naturalmente asigna probabilidades por postura a cada palabra, es un modelo de clasificación que será testeado separando, tanto las seis posturas y casos independientes, así como agrupando las opciones en dos. Al analizar las probabilidades se pueden identificar aquellos elementos (palabras o tópicos, según la vectorización) que contribuyan más al resultado final del algoritmo, obteniendo la capacidad interpretativa que se señalaba. Se concluye con la familia de modelos no-supervisados llamada Modelamiento de Tópicos, para el cual se probaron, tanto el modelo *LDA* como *BERTopic* (Grootendorst, 2022), basado en redes neuronales. Estos modelos también permiten asociar palabras y textos completos a probabilidades de clase, pero esta vez las clases corresponderán a los tópicos encontrados.

Resultados

En la Figura 2, el gráfico agregado de las tres realizaciones del caso Adela muestra que las opciones 1, 2 y 3 son predominantes a lo

largo de la actividad, siendo en la EG donde se agudiza la diferencia, por tanto, hay una clara concentración de la muestra de estudiantes en una postura, oscilando entre el 64% y el 71% del total. En particular, la mayor cantidad de estudiantes se concentra en el valor 2 en la EI-1 (29%) y en el valor 3 para la etapa grupal e individual 2 (42% y 30%, respectivamente). Destaca que en la EI-1 el 21% escoge la opción 3, mientras que en la EG esta preferencia aumenta al 42%, para luego, en la etapa 3, disminuir a 30%. Este comportamiento señala que en la etapa grupal las posturas se moderan, posiblemente por el contraste de ideas que se generan entre quienes conforman el grupo.

El examen exploratorio básico incluyó un análisis de la extensión de las respuestas (número de palabras). El promedio de palabras es inferior a 50 para el caso estudiado. Las opciones (1) y (6) son menos extensas que el promedio en comparación con las opciones más moderadas (2, 3, 4 y 5). Este fenómeno puede responder al hecho de que una persona absolutamente convencida de su postura vea menos necesidad de explicarla. De cualquier modo, la diferencia es sutil.

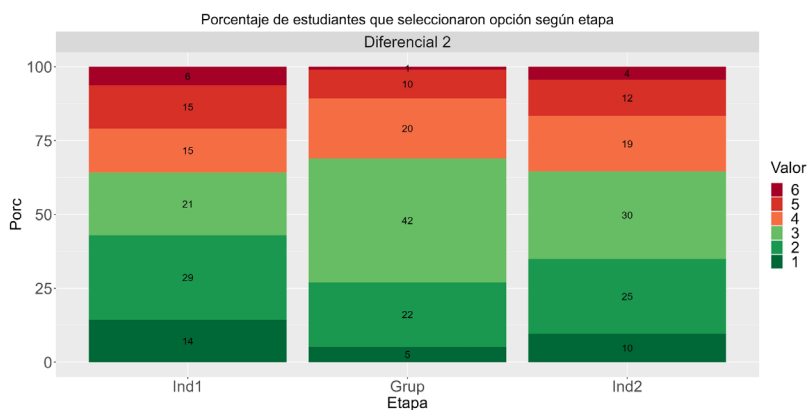


Figura 2. Porcentaje de estudiantes que seleccionaron opción según etapa (EI-1, EG, EI-2), Caso Adela en las tres realizaciones.

Por otro lado, se analizan las palabras más comunes condicionado cada postura (ver Figura 3). Se puede notar que algunas palabras utilizadas por los y las estudiantes para justificar son las mismas que las palabras de cada postura para el diferencial. Por ejemplo, se aprecia que para apoyar la producción del alimento se evocan conceptos como «salud» y «personas», así como también, acciones como «producir» y «priorizar», este último, reflejando que estudiantes que asumen esa postura probablemente priorizan el bien común y/o salud de niños y ancianos en su decisión. En contraste, sustantivos como «cultura» y «pueblos» (en plural) aparecen al escoger resguardar las tradiciones. Asimismo, expresiones como «resguardar», «respetar» y «debería» son usadas con el mismo propósito. Este análisis también permite identificar palabras ampliamente repetidas sin importar la postura del estudiante, como es el caso de «tradiciones». Adicionalmente, el concepto «acuerdo» aparece solo en las opciones (3) y (4), que son las más moderadas del espectro de posturas.

En la nube de palabras (Figura 4) se muestra una visualización de la frecuencia de palabras por postura (la postura de la izquierda corresponde a las opciones 1, 2 y 3; y la de la derecha, a las opciones 4, 5 y 6), y se agregan, además, pares de palabras consecutivas. Una mayor frecuencia se traduce en un mayor tamaño de la palabra en la nube. Se distingue que en la nube de la izquierda, que corresponde a la postura de *producir alimento contra déficit vitamínico*, aparece la palabra «beneficio», posiblemente para aportar mayor comprensión a la elección. Además, tanto en la nube de palabras de la izquierda como en la de la derecha, que corresponde a la postura de *resguardar tradiciones identitarias*, se repite la palabra «acuerdo». Las otras palabras y frases no aportan información adicional a la comprensión de las justificaciones porque aparecen en la descripción del caso, o bien, en la construcción del diferencial semántico.

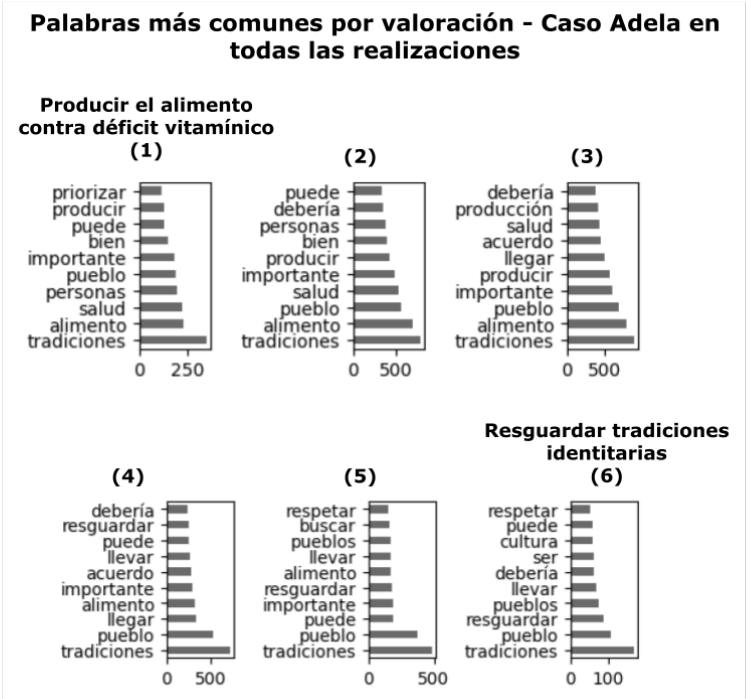


Figura 3. Frecuencias de posturas por etapa en caso Adela.

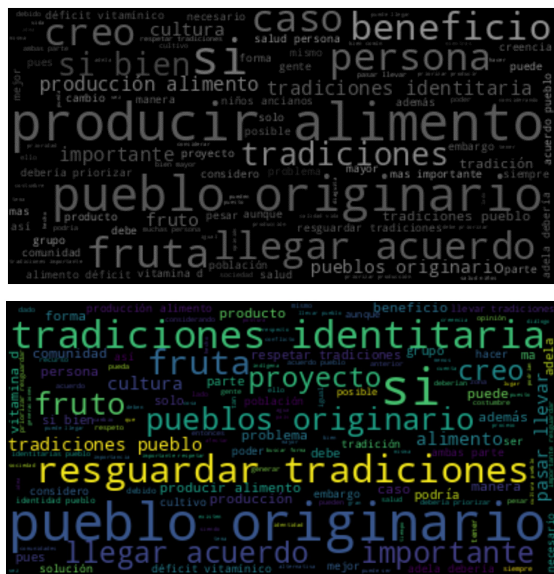


Figura 4. Nube de palabras por postura para el caso Adela.

Resultados de modelos estadísticos

Dado que predecir de manera exacta la opción (de 1 hasta 6) es una tarea difícil, se considera predecir la postura binaria (agrupando las opciones (1), (2) y (3), por un lado, y las opciones (4), (5) y (6), por el otro). Con esta estrategia la exactitud del modelo sube de 0.43 a 0.79 (mientras más alta, mejor es la predicción), pero además permite observar la importancia de palabras para cada postura.

Se incluyen a continuación aquellas palabras que tienen más probabilidad para cada postura, como se ha explicado en el párrafo anterior. La Tabla 2 muestra conceptos para los cuales se identificó una probabilidad alta para cada una de las posturas tomando como referencia el modelo *Naive-Bayes*. La presencia de estos conceptos

en el texto se relaciona con las posturas en cuestión, lo cual pone en evidencia los elementos semánticos más recurrentes a la hora de escribir las justificaciones. Es posible llevar a cabo un análisis similar con los modelos de tipo lineal. De la tabla destaca que para la postura de producir el alimento hay más probabilidad de que en la justificación se usen palabras como «salvar», «juego», «vidas», «tribu», «religiosas» y «riesgo»; estas palabras podrían mostrar que quienes escogen esta postura consideran que salvar vidas o las vidas que están en juego son más importantes. Igualmente, señalan a los pueblos originarios como «tribu» y para referirse a sus tradiciones las consideran de tipo «religiosas». Por otro lado, las palabras que destacan de aquellas que más probabilidad tienen de ser usadas para la postura de resguardar tradiciones identitarias son «suplementos», «sol» y «obtener»; esto podría relacionarse con que quienes escogen esta opción aluden a alternativas para abordar el déficit vitamínico vía suplementos u obtenerla a través del sol. El mismo análisis se obtiene al considerar un modelo que prediga el valor seleccionado entre 1 y 6 y puede también generalizarse a otro tipo de modelos.

Tabla 2
Palabras con alta probabilidad para ambas posturas con modelo Naive-Bayes

Producir el alimento contra déficit vitamínico				Resguardar tradiciones identitarias			
Palabra	Prob	Palabra	Prob	Palabra	Prob	Palabra	Prob
1. <i>salvar</i>	0.985	6. <i>poniendo</i>	0.942	1. <i>siglos</i>	0.880	6. <i>usuario</i>	0.761
2. <i>juego</i>	0.958	7. <i>riesgo</i>	0.938	2. <i>suplementos</i>	0.863	7. <i>consentimiento</i>	0.761
3. <i>vidas</i>	0.952	8. <i>tiempos</i>	0.934	3. <i>sol</i>	0.825	8. <i>integridad</i>	0.730
4. <i>tribu</i>	0.950	9. <i>cambiarían</i>	0.934	4. <i>existen</i>	0.822	9. <i>sacar</i>	0.724
5. <i>religiosas</i>	0.948	10. <i>ayudarían</i>	0.934	5. <i>única</i>	0.770	10. <i>obtener</i>	0.723

En la Tabla 3 se muestra el resultado del modelamiento de tópicos con el método *BERTopic*. Respecto a los modelos, gracias a una inspección de los tópicos obtenidos, se concluye que el modelo *LDA* identifica temas menos distintos unos de otros con respecto a *BERTopic*, por ende, se restringe el análisis a este último. En general, no se observa una relación significativa con la postura tomada, sin embargo, esto es normal, considerando que se pueden tomar posturas contrapuestas al usar elementos semánticos similares para justificarlas. Para ver ejemplos de justificaciones se incluyen algunas muestras en la Tabla 3. Se observa que el tópico 1 tiene una alta probabilidad de aparición (70,99%) en contraste con los otros. Como indica el ejemplo, la justificación esgrimida en este tópico no muestra necesariamente una complejidad en el desarrollo moral, en ese sentido, destaca la ausencia de principios éticos en las palabras clave para cada tópico. Si bien, este modelo no permite un estudio exhaustivo de todos los casos para apoyar la hipótesis acerca de la complejidad en el desarrollo moral, sí conviene tenerla en consideración para futuras investigaciones. Los otros tópicos muestran una baja probabilidad de aparición, pero se pueden notar perspectivas éticas de tipo utilitaristas para escoger la opción de priorizar la salud (porque impactaría a más personas, como muestran los ejemplos).

Tabla 3
Caracterización de Tópicos

Tópico	Palabras clave	Prob	Ejemplo
1	alimento, tradiciones, pueblo	70,99%	Adela debería producir el alimento. Creo que es más importante beneficiar a la gente que posee un déficit vitamínico, ya que es importante resguardar las tradiciones identitarias. Pero eso no está sobre la salud de la población.
2	tradiciones, pueblo, acuerdo	16,8%	Primero sería intentar llegar a un acuerdo con el pueblo, pero si no se puede, si puede salvar personas este fruto entonces es más importante la vida de las personas que una tradición.
3	salud, importante, personas	5,27%	Considero mucho más importante priorizar a la mayor parte de la población por el bien de la salud.
4	proyecto, innovación, si	2,14%	Si bien Adela no tiene limitaciones legales para su proyecto de innovación, sí debería tomar en cuenta la opinión de los pueblos originarios de la zona para tratar de llegar a un acuerdo con ellos, con el fin de respetar las tradiciones y la cultura.
5	Chile, relleno, chilena	1,83%	Ojalá en lo posible se pueda llegar a un acuerdo con los pueblos originarios para respetar aunque sea una parte. Pero definitivamente producir el alimento genera un bien mayor ya que puede ayudar a más del 50% de los niños y adultos mayores en Chile.
6	opinión, mantengo, no	0,97%	Mantengo mi postura

Conclusiones

Respecto a los modelos usados en este trabajo, si bien fueron modelos simples basados en conteos de palabras, la formulación probabilística los dota de la capacidad de clasificar nuevos textos, tanto de manera supervisada como no supervisada. Estas técnicas son además interpretables, característica que es de especial importancia en el contexto del ámbito disciplinar de la ética y su formación. En

el caso específico de *Naive-Bayes*, se tiene tanto una distribución de palabras por clase como una distribución de clase por palabra, lo cual es particularmente útil para analizar el lenguaje empleado en las decisiones de los y las estudiantes.

El desarrollo de otros métodos de aprendizaje profundo capaces de declarar variables (deseablemente, conceptos) que sean destacables a la hora de reflejar la competencia ética, es un trabajo futuro por considerar. En cualquier caso, el presente trabajo muestra una metodología que tiene la capacidad fotografiar la naturaleza de las respuestas y los conceptos empleados para justificarlas. Se espera que esta herramienta ayude a graficar la coherencia de las respuestas cada año, de tal manera que con esta información se puedan tomar decisiones justificadas respecto a la metodología para la enseñanza de la ética en la Facultad de Ciencias Físicas y Matemáticas.

En cuanto a los principales resultados obtenidos a través de los modelos *-Naive-Bayes* y *BERTopic-*, al vincularlos con los principios que subyacen al diferencial estudiado (responsabilidad y respeto), se muestra que, si bien el respeto se tendría en consideración por utilizar la palabra «acuerdo» en sus justificaciones, los y las estudiantes se inclinan hacia razonamientos que priorizan el principio de la responsabilidad (asociado a *producir el alimento contra déficit vitamínico*), en cuanto a la perspectiva de salvar vidas que estén en riesgo en contraste con anteponer un principio de respeto (asociado a *resguardar las tradiciones identitarias*), buscando alternativas a la vitamina vía suplementos o a su obtención a través de la energía del sol. Este hallazgo puede ser representativo del foco de la docencia que está entregando el curso, lo que se relaciona directamente con formar parte de una asignatura de la línea de innovación. Sin embargo, una hipótesis que se plantea es que los y las estudiantes, en términos generales, no utilizan principios éticos subyacentes a su justificación, lo que podría indicar el nivel básico en que se encuentran en el desarrollo del discernimiento ético y, en consecuencia, en la competencia ética.

Este trabajo corresponde a una primera aproximación a las justificaciones de los y las estudiantes, mostrando que se hace necesaria una exploración permanente y diversificada en cuanto a los modelos de procesamiento de lenguaje natural a utilizar.

Referencias bibliográficas

- Blei, D. M., Ng, A. Y. y Jordan, M. I. (2003). Latent dirichlet allocation. En *Journal of Machine Learning Research*, 3(Jan), 993-1022.
- Cortina, A. (2000). Presentación. El sentido de las profesiones. En A. Cortina y J. Conill, *10 palabras clave en ética de las profesiones*. Ed. Verbo Divino: Navarra.
- Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–4186.
- Garten, J., Boghrati, R., Hoover, J., Johnson, K. M. y Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. En *Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes*.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P. y Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. En *Advances in experimental social psychology* (Vol. 47). Academic Press. 55, 130.
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. En *arXiv preprint arXiv:2203.05794*.
- Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J. y
- Dehghani, M. (2021). Moral concerns are differentially observable in language. En *Cognition*, 212, 104696. <https://doi.org/10.1016/j.cognition.2021.104696>
- Ramirez, P., Guerrero, S., Cerda, J., Ross, J., Flores, G. (2022). «La formación ética canalizada mediante la tecnología. Experiencia y resultados preliminares del uso de la herramienta web Ethicapp». *Cuadernos de Beauchef*, 6(2), 73-87.

Xie, J. Y., Hirst, G. y Xu, Y. (2020). Contextualized moral inference. En *arXiv preprint arXiv:2008.10762*. <https://doi.org/10.48550/arXiv.2008.10762>

Complicaciones y complejidades de convivir con decisiones tomadas por modelos de IA

Bernardo Subercaseaux¹

Desde los albores de nuestra especie, la vida de cada ser humano es continuamente afectada por las decisiones tomadas por otros. El siglo XXI, no obstante, parece traer consigo un desafío sin precedentes: aquellos otros pueden ser ahora modelos de inteligencia artificial (IA). En este ensayo discuto el desafío de convivir con decisiones tomadas por estos nuevos actores, los modelos de IA, tanto desde un punto de vista matemático como de uno antropológico. Matemáticamente hablando, la ya madura teoría de la complejidad computacional permite demostrar rigurosamente la alta dificultad de obtener explicaciones, incluso, sobre los modelos que comúnmente se consideran «transparentes» o «interpretables», como son los árboles de decisión. Basándome en resultados previos (Arenas et al., 2021, 2022; Barceló et al., 2020a,b), argumento que de momento parece existir una oposición inherente entre (i) interpretabilidad, es decir, nuestra capacidad de entender y explicar las decisiones tomadas por un modelo de IA, y (ii) precisión, es decir, el porcentaje de decisiones que los modelos utilizados toman correctamente.

En cuanto a lo antropológico, en este ensayo propongo que el primer paso en pos de una convivencia saludable con modelos de IA, es aceptar la oposición anteriormente mencionada y, basándonos en ella, responder a consciencia la pregunta «¿cuánta interpretabilidad estamos dispuestos a sacrificar en pos de una mejor

¹ Ingeniero Civil en Computación, Universidad de Chile (2020). Estudiante 4to año de doctorado en Ciencias de la Computación, Carnegie Mellon University. bersub@cmu.edu

precisión?». Naturalmente, la respuesta a esta pregunta dependerá de la aplicación y contexto específico en que se utilicen los modelos; pero, en cualquier caso, propongo que plantearse esta pregunta deliberadamente constituye un deber ético para quienes utilizan tales modelos afectando a otros.

1. Ya están entre nosotros

Nuestras vidas son continuamente afectadas por decisiones que toman otros; si somos o no admitidos a una cierta institución, las notas que reciben nuestros trabajos, quién es despedido primero ante una crisis económica, si el banco decide aprobarnos un crédito o, incluso, si una potencial pareja romántica prevé un futuro con nosotros. En el siglo XXI, sin embargo, un nuevo actor ha entrado en escena: los modelos de inteligencia artificial (IA). Permítanme dar algunos ejemplos concretos:

- Desde 2013 a 2020, la Universidad de Texas, en Austin, utilizó un modelo de IA para evaluar postulantes a sus programas de posgrado (Burke, 2020).
- Una variedad de empresas ofrece servicios de reclutamiento y entrevistas laborales con modelos de IA (Kelly, 2024; Sapia.ai, 2024; Talently.ai, 2024).
- Recientemente, Whitney Wolfe Herd, fundadora de la popular aplicación de citas Bumble, comentó que en el futuro serán modelos de IA los que conversarán entre sí para decidir sobre la compatibilidad de los usuarios humanos (Pringle, 2024).
- En Estados Unidos, el sistema judicial utiliza modelos de IA para predecir, entre otras cosas, la probabilidad de que un acusado reincida en un crimen (Jeff Larson and Mattu, 2016; Malmon, 2023). Vale la pena notar, inmediatamente, que uno los principales sistemas de predicción utilizados, COMPAS, ha sido ampliamente criticado por su sesgo racial en contra de personas de color (Julia Angwin and Mattu, 2016).

En resumen, modelos de IA están entre nosotros, no solo en calidad de herramientas, sino también, en calidad de tomadores de decisiones, un rol que hasta hace poco estaba reservado para los seres humanos. Esta tendencia, además, parece aumentar mes a mes; desde una perspectiva financiera los modelos de inteligencia artificial son cada vez más rentables en tanto que tomadores de decisiones, y pueden utilizarse a escalas masivas que serían imposibles para seres humanos. Por ejemplo, cuando la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile ofrece el concurso literario Beauchef en 100 palabras, un jurado humano debe leer y evaluar cada uno de los cientos de relatos presentados, por lo que logísticamente sería complejo ofrecer un concurso como aquel más de una vez al año. En contraste, un modelo de IA podría evaluar la totalidad de los relatos recibidos en cuestión de segundos, permitiendo ofrecer semejantes concursos a diario. Es de esperarse que, en ausencia de regulaciones que garanticen a las y los ciudadanos un derecho a explicaciones para las decisiones tomadas por IA, como, por ejemplo, la GDPR en la Unión Europea², las empresas, en cuanto agentes que usualmente maximizan sus finanzas, tenderán a utilizar cualquier modelo que les otorguen beneficios financieros, sin importar si tales modelos son interpretables o no. Peor aún, una variedad de modelos de IA utilizados por empresas y gobiernos afecta desproporcionadamente a grupos tradicionalmente marginados, ya sea desde un punto de vista racial, de género, o socioeconómico³. Si bien, los avances en IA ofrecen un sinnúmero de oportunidades tecnológicas nuevas, el foco de este artículo es la cara opuesta de tan brillante moneda: la convivencia con estos nuevos actores, pues los modelos de IA presentan desafíos éticos, sociales y matemáticos que no pueden ser ignorados.

² Véase Goodman and Flaxman (2017). Considero la GDPR como un ejemplo que ilustra la posibilidad de regular sobre el derecho a explicaciones, más no emito un juicio de valor sobre la calidad de esa regulación específica. Se trata de una materia legal compleja que requiere una aproximación multidisciplinaria.

³ Véase los trabajos de Buolamwini and Gebru, 2018; Dubber et al., 2020; O'Neil, 2016. Una nueva referencia en esta línea, en español, es el libro de Bruneau (2024).

2. Explicabilidad e interpretabilidad

Si hemos de convivir con decisiones tomadas por modelos de inteligencia artificial, una pregunta se vuelve ineludible: ¿podemos entender el por qué detrás de las decisiones tomadas por estos modelos? En caso de no poder hacerlo, nos encontramos en un embrollo significativo: ¿cómo podemos confiar en decisiones que no entendemos? ¿Cómo podemos corregir errores, o sesgos, sin entender su causa?

Una alternativa posible es progresivamente ceder control a los modelos de IA y aceptar sus decisiones sin cuestionarlas. Una ilustración provocadora se puede ver en el episodio número 5 de la serie *Love, Death & Robots* de Netflix, titulado «When The Yogurt Took Over»; este muestra un futuro distópico en que la humanidad ha cedido el control de la economía a una nueva forma de inteligencia cuyas ecuaciones no es capaz de comprender. Cuando la humanidad se desvía de las recomendaciones de esta nueva forma de inteligencia una crisis financiera se desata, mientras que, al volver a seguir tales recomendaciones, la prosperidad económica retorna. Al menos por ahora esta opción parece inadmisiblemente sombría: incluso si en ciertos dominios restringidos los modelos de IA superan a los humanos, el poder entender sus decisiones sigue apareciendo como un requisito fundamental para poder confiar en ellos. En esta línea, el científico británico Geoffrey Hinton, galardonado con el premio Turing 2018 y considerado uno de los padres de la inteligencia artificial moderna, preguntaba en Twitter (Geoffrey Hinton, 2020):

Supón que tienes un tipo de cáncer estadísticamente raro, y debes elegir entre dos opciones. Por un lado, puedes someterte a una cirugía con un modelo de IA que cuenta con un 90% de éxitos en casos similares, pero cuyas decisiones escapan al entendimiento de expertos y expertas. La segunda opción es proceder bajo la mano de una cirujana, humana, que ha tenido un 80% de éxito en casos similares. ¿Cuál eliges?⁴

⁴ Además de traducir la pregunta original de Hinton, he modificado ligeramente su enunciado incluyendo más contexto.

Esta provocadora pregunta conduce a pensar en una dicotomía entre interpretabilidad y precisión. A continuación, presentaré una descripción matemática de estos conceptos y su potencial oposición. Un modelo preliminar de la toma de decisiones consiste en la evaluación de una función, donde una decisión binaria es tomada a partir de piezas de información, también binarias. Por ejemplo, supongamos un banco que ha de modelar la decisión de otorgar un préstamo a una solicitante a partir de los siguientes atributos:

1. ¿Tiene la persona solicitante un trabajo estable? (Sí/No)
2. ¿Tiene la persona solicitante un sueldo anual mayor o igual al 30 % del monto solicitado? (Sí/No)
3. ¿Tiene la persona solicitante un historial crediticio limpio? (Sí/No)
4. ¿Tiene la persona solicitante más de 30 años? (Sí/No)
5. ¿Tiene la persona solicitante un título universitario? (Sí/No)
6. ¿Tiene la persona solicitante un aval? (Sí/No)
7. ¿Tiene la persona solicitante una casa propia? (Sí/No)

En este caso, la dimensión asociada es $d = 7$, y la evaluación $f((1,0,1,1,0,1,0)) = 0$ corresponde al rechazo de la solicitud de una solicitante con las características descritas. Un modelo de IA ha de aprender esta función a partir de un conjunto de ejemplos, cada uno de los cuales se compone de la lista de atributos de una solicitud y de la decisión correcta o esperada. Idealmente, el modelo de IA será capaz de generalizar la información que yace implícita en los ejemplos sobre los cuales ha sido entrenado, y tomará decisiones razonables frente a ejemplos nuevos. Naturalmente, el uso de un tal modelo puede representar un ahorro significativo en tiempo y dinero para un banco que recibe un gran volumen de solicitudes de créditos, y que habitualmente utiliza humanos asalariados para evaluar cada una de ellas en ausencia de modelos. La pregunta al corazón de la explicabilidad, o interpretabilidad⁵ es:

¿Por qué un modelo M ha decidido que $M((1,0,1,1,0,1,0)) = 0$?

⁵ Para propósitos de este ensayo no distinguiremos entre estos conceptos. El libro de Molnar (2022) ofrece una breve discusión al respecto.

La interpretabilidad de un modelo, según Miller (2019), corresponde al grado en el cual un humano determina la causa de sus decisiones. Esta definición, sin embargo, no aclara el sentido en el cual las decisiones son «causadas»; más aún, Miller utiliza la palabra «causa» en singular, asumiendo implícitamente que para la noción de causalidad en juego existirá una única causa para cada decisión⁶.

Una distinción que considero particularmente importante a la hora de explicar una decisión $M(\vec{x})=b$, es la distinción entre una explicación *post hoc* y una explicación basada en los datos de entrenamiento. Una forma posible de explicar la decisión $M(\vec{x})=b$, conociendo los datos de entrenamiento del modelo M , consiste en identificar ejemplos en los datos de entrenamiento que sean similares a \vec{x} , cuya respuesta sea también b , y que han contribuido de alguna manera a que el modelo aprendiese a decidir b para casos similares. Otro paradigma, denominado *post hoc* (Molnar, 2022), consiste en identificar partes de la entrada \vec{x} que son relevantes para la decisión $M(\vec{x})=b$ sin, necesariamente, conocer los datos de entrenamiento. Por ejemplo, si descubrimos que el modelo M siempre decide b cuando el atributo 3 tiene valor 1 (i.e., «Sí») entonces podríamos decir que el hecho $\vec{x}[3]=1$ explica, de alguna manera, la decisión $M(\vec{x})=b$. A continuación, presentaré dos tipos de explicaciones *post hoc* que considero particularmente simples y relevantes en el contexto de la interpretabilidad de modelos de IA.

Notación. Llamaremos *instancias* a los elementos del conjunto $\{0,1\}^d$, e instancias parciales a los elementos del conjunto $\{0,1,\perp\}^d$, donde \perp representa un valor *indeterminado*. En el conjunto de las instancias parciales definiremos una relación de *contención*, denotada por \subseteq , que intuitivamente corresponde a que $\vec{y} \subseteq \vec{x}$ si \vec{x} coincide con \vec{y} en los atributos determinados de \vec{y} , pero potencialmente agrega información en los atributos indeterminados de \vec{y} . Por ejemplo, $(1,0,\perp,1) \subseteq (1,0,0,1)$, pero $(1,0,\perp,1) \not\subseteq (1,0,1,0)$. Formalmente, diremos que $\vec{y} \subseteq \vec{x}$ si para todo i tal que $\vec{y}[i] \neq \perp$ se cumple que $\vec{y}[i] = \vec{x}[i]$. Si dada una instancia parcial \vec{y} , y una instancia \vec{x} se cumple $\vec{y} \subseteq \vec{x}$ que diremos que \vec{x} es una completación de \vec{y} .

⁶ Para un tratamiento matemático general de la idea de causalidad, recomiendo al lector dirigirse al trabajo de Pearl and Mackenzie (2018).

Definición 1 (*Razón suficiente* (Darwiche and Hirth, 2020)). Dado un modelo $M:\{0,1\}^d \rightarrow \{0,1\}$, y una instancia \vec{x} tal que $M(\vec{x})=b$, diremos que una instancia parcial $\vec{y} \subseteq \vec{x}$ es una «razón suficiente» para la decisión $M(\vec{x})=b$ si $M(\vec{z})=b$ para cualquier completación \vec{z} de \vec{y} .

Las «razones suficientes» corresponden a un tipo de explicación «abductiva» (Marques-Silva and Ignatiev, 2022), en la cual un subconjunto de la información contenida en \vec{x} justifica el veredicto $M(\vec{x})=b$. Otro tipo de explicación *post hoc* corresponde a las explicaciones contrafactuales, centradas en entender una decisión $M(\vec{x})=0$ a partir de la pregunta «¿qué tendría que haber sido diferente en \vec{x} para que $M(\vec{x})=1$?». Concretamente, consideremos la siguiente definición:

Definición 2 (*Explicación contrafactual* (Barceló et al., 2020a; Marques-Silva and Ignatiev, 2022)). Dado un modelo $M:\{0,1\}^d \rightarrow \{0,1\}$, y una instancia \vec{x} diremos que una instancia \vec{z} es una explicación contrafactual para $M(\vec{x})=b$ si $M(\vec{x}) \neq M(\vec{z})$ y la cantidad de atributos i en los cuales $\vec{x}[i] \neq \vec{z}[i]$ es mínima.

El trabajo de Miller (1956) sugiere que, para que una explicación sea efectiva en cuanto a su comprensión por humanos, esta debe ser pequeña. En otras palabras, una razón suficiente que requiere examinar cientos o miles de atributos no será efectiva en la práctica. Es decir, si queremos encontrar explicaciones abductivas efectivas, estas deben ser «pequeñas». La noción de una razón suficiente pequeña puede ser formalizada de distintas maneras; por ejemplo, como una que contiene la mínima cantidad de atributos definidos, o que no es estrictamente subsumida (en el sentido que induce el orden de contención) por otra razón suficiente. Encontrar razones suficientes con un número mínimo de atributos definidos es un problema computacionalmente difícil, incluso para modelos de IA que comúnmente se consideran interpretables, como son los árboles de decisión (Barceló et al., 2020a). Más aún, esta dificultad se mantiene al relajar la noción de razón suficiente de manera probabilista (Arenas et al., 2022), y también, al buscar razones suficientes que, sin ser necesariamente mínimas, «aproximan» una

razón suficiente mínima (Kozachinskiy, 2023). A nivel más general, los últimos cinco años de investigación en interpretabilidad desde una perspectiva matemática formal (véase Marques-Silva (2023)), han estudiado la complejidad computacional de responder este tipo de preguntas de interpretabilidad para diferentes tipos de modelos. La siguiente sección se enfocará en la relación entre interpretabilidad y complejidad computacional para diferentes tipos modelos de IA.

3. Interpretabilidad desde la complejidad computacional

Esta sección asumirá una familiaridad básica con la teoría de la complejidad computacional⁷. La hipótesis fundamental de esta sección es que la interpretabilidad de una clase de modelos de IA está relacionada con la complejidad computacional de encontrar explicaciones para modelos de esta clase. Para aceptar esta hipótesis pareciera ser necesario aceptar al menos un par de premisas que ilustraré a continuación.

Premisa 1: buscamos explicaciones formales. En la sección 2 describimos algunas formalizaciones de explicaciones sobre decisiones tomadas por modelos de clasificación. Esta premisa consiste en aceptar que una clase de modelos «interpretables» es una en la que seremos capaces de encontrar explicaciones formales en la práctica. En otras palabras, si un modelo de IA es interpretable, entonces, esperamos ser capaces de encontrar razones suficientes para sus decisiones, o explicaciones contrafactuales para estas, en un tiempo razonable.

Premisa 2: la teoría de la complejidad computacional es predictiva. La teoría de la complejidad computacional busca diferenciar entre aquellos problemas que seremos capaces de resolver en la práctica y aquellos que no. Una hipótesis tradicional en este sentido es que problemas para los cuales contamos con algoritmos que requieren un tiempo exponencial en el tamaño de la entrada para ser resueltos no serán fácilmente resueltos en la práctica. En cambio, aquellos

⁷ Como referencia a las definiciones e ideas fundamentales de esta teoría sugiero el libro de Arora and Barak (2006).

problemas para los cuales contamos con algoritmos que requieren un número polinomial de pasos en el tamaño de la entrada serán resueltos en la práctica⁸.

Si aceptamos estas premisas, entonces, la interpretabilidad de una clase de modelos de IA está relacionada con la complejidad computacional de encontrar explicaciones: si una clase de modelos es interpretable, entonces, el problema de computar explicaciones para ella podrá ser resuelto en tiempo polinomial, mientras que en clases de modelos no interpretables, este problema será computacionalmente intratable (i.e., *NP-hard*). En mi trabajo junto a Barceló et al. (2020a), exploramos esta hipótesis demostrando que encontrar explicaciones para las decisiones tomadas por árboles de decisión o modelos lineales es computacionalmente más sencillo que para redes neuronales. Sin embargo, una serie de explicaciones son intratables incluso sobre árboles de decisión. La siguiente sección plantea que estos resultados reflejan una oposición inherente entre interpretabilidad y precisión.

4. Interpretabilidad y precisión

Dado un conjunto $D = \{(\vec{x}_1, b_1), \dots, (\vec{x}_n, b_n), \dots\}$ de datos \vec{x}_n etiquetados con su clasificación deseada b_n , la manera tradicional de evaluar la precisión de los modelos que entrenamos sobre ellos está basada en particionar D en dos: un subconjunto $T \subseteq D$ de datos se utilizará para entrenar el modelo, y un subconjunto $E = D \setminus T$ se utilizará para evaluar la precisión del modelo ya entrenado (Hastie et al., 2001). De este modo, si un modelo M_T resulta de entrenar sobre T , definiremos su «precisión» sobre el conjunto E según:

$$p_E = \frac{|\{\vec{x}_m \in E \text{ tal que } M(\vec{x}_m) = b_m\}|}{|E|},$$

⁸ Esta hipótesis es comúnmente atribuida a Cobham (1965) y Edmonds (1965). Para una discusión de las limitaciones de esta aproximación, véase el trabajo de Roughgarden (2021).

o, en otras palabras, como la fracción de datos en E que son clasificados correctamente por M_T . Es sabido que, para obtener un cierto grado de precisión en ciertos problemas, el tamaño de los modelos de IA que se requieren, dependerá crucialmente de la clase de modelos a utilizar. Consideremos un problema sencillo: determinar si una entrada en $\{0,1\}^d$ tiene un número par de 1s o no. Llamemos a este problema *Paridad*.

Teorema 1 (Folklore, véase (Wegener, 2000)). Existen redes neuronales de tamaño polinomial en d que resuelven *Paridad* con precisión 1 para cualquier conjunto de datos de evaluación E . Por otra parte, cualquier árbol de decisión que obtiene precisión 1 para cualquier conjunto de datos de evaluación debe tener tamaño exponencial en d .

Más aún, si consideramos una función booleana F , escrita en «forma normal conjuntiva» (CNF (Arora and Barak, 2006)), es fácil construir una red neuronal M que computa la misma función que F y cuyo tamaño es polinomial en el tamaño de F . En contraste, un árbol de decisión que computa la función F tendrá típicamente tamaño exponencial en F (Wegener, 2000). Los resultados de complejidad mencionados en la sección 2 para redes neuronales aplican directamente a cualquier clase de modelos capaces de representar una función booleana F (en CNF) con tamaño polinomial en F . Esto sugiere, entonces, que la complejidad de interpretación es una propiedad necesaria para cualquier clase de modelos suficientemente poderosa, para representar sucintamente funciones booleanas en CNF. No es difícil demostrar que, para clases de modelos más débiles como son los árboles de decisión, cualquier modelo de esa clase con un tamaño razonable (digamos polinomial en $|F|$) tendrá baja precisión en ciertos conjuntos de evaluación E . Peor aún es el caso de modelos lineales, que por definición son incapaces de representar ciertas funciones booleanas (como es el caso para *Paridad*), y que, por lo tanto, tendrán baja precisión en prácticamente cualquier conjunto de evaluación E . En otras palabras, pareciera que la capacidad de una clase de modelos de aprender funciones complejas manteniendo un tamaño razonable está en oposición directa con nuestra capacidad de obtener explicaciones eficientes para sus decisiones.

4.1. La tesis de Rudin

La científica norteamericana Cynthia Rudin plantea en su célebre trabajo *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* (Rudin, 2019) que, en lugar de intentar explicar modelos complejos de IA como son las redes neuronales, deberíamos, en su lugar, utilizar modelos más interpretables como son los árboles de decisión. A mi juicio hay dos problemas con esta tesis. En primer lugar, como mencioné en la sección 2⁹, no es claro que modelos como los árboles de decisión sean efectivamente interpretables en la práctica. Más estudios experimentales en la línea de Piltaver et al. (2016) serán necesarios para esclarecer este problema. En segundo lugar, y posiblemente más importante, la tesis de Rudin se basa en la idea de que es posible obtener modelos interpretables con la misma precisión que aquellos menos interpretables¹⁰. Dicho en breve, el argumento de Rudin es que el espacio de modelos (entendidos como funciones) cuya precisión es cercana al óptimo (i.e., Conjuntos de *Rashomon*) es suficientemente grande como para esperar que contenga modelos interpretables. En mi opinión, hay dos problemas cruciales con esta hipótesis que planteo a continuación.

Problema #1: modelos versus funciones. La interpretabilidad de un clasificador no es una propiedad de la función que implementa sino de la implementación concreta de esa función. En otras palabras, creo que para una función fija $f: D \rightarrow R$, sus diferentes «implementaciones» tendrán distintos grados de interpretabilidad. Por ejemplo, consideremos la función $f: \{0,1\}^d \rightarrow \{0,1\}$ definida según:

$$f(\vec{x}) = \{1 \text{ si } \vec{x} \text{ tiene un número par de 1s o 0 si no.}\}$$

Esta función f puede ser implementada de muchas maneras, por ejemplo, una posible implementación I_1 es iterar sobre los elementos de la entrada \vec{x} contando el número de 1s, y retornar

⁹ Para un seguimiento de esta discusión ver Izza et al., 2020; Lipton, 2018; Marques-Silva and Ignatiev, 2023.

¹⁰ Véase el trabajo de Semenova et al. (2022) para una discusión más extensa de esta hipótesis.

de acuerdo con la condición de paridad. Otra implementación, I_2 podría ser computar primero la expresión $g(\vec{x}) := \zeta(-2 - \|\vec{x}\|_{l_1})$ donde ζ es la función zeta de Riemann, computada según el método de Karatsuba (1995), y retornar 1 si es que $g(\vec{x})$ es 0, y 0 si es que $g(\vec{x}) \neq 0$. Si bien, ambas implementaciones I_1 e I_2 computan la función f , debiese ser claro que la primera es mucho más *interpretable* que la segunda. En general, en el contexto de clasificación binaria $\{0,1\}^d \rightarrow \{0,1\}$ toda función puede ser implementada por un número exponencial de árboles de decisión distintos, y un número infinito de redes neuronales distintas; sería impensable que cada una de estas implementaciones tuviese el mismo grado de interpretabilidad. En resumen, planteo que el espacio por estudiar no debiese ser el de las funciones $f: D \rightarrow R$, sino el de las implementaciones de estas funciones. Este espacio, por supuesto, tiende a ser mucho más complejo de analizar.

Problema #2: quizás no hay aguja en el pajar. Refinando la hipótesis de Rudin según la discusión del párrafo anterior, obtenemos una hipótesis de la forma «*el conjunto de modelos (i.e., implementaciones) con una precisión aceptable es suficientemente grande como para esperar que contenga modelos interpretables*».

Esta hipótesis se parece bastante a la intuición de Knuth (2014) sobre la pregunta $P \stackrel{?}{=} NP$, la que plantea que el espacio de los algoritmos que solucionan SAT ((Arora and Barak, 2006)) es tan amplio e inhumano, que probablemente algún algoritmo en ese espacio corre en tiempo polinomial. Si bien esta idea es atractiva y razonable, pareciera que la gran mayoría de las y los científicos de la computación cree que $P \neq NP$ y que, por lo tanto, incluso si el espacio del pajar es inimaginablemente grande, es posible sospechar que no contiene la aguja que buscamos. De momento, la experiencia empírica sugiere que existe una oposición inherente entre interpretabilidad y precisión, de forma similar en que sugiere que $P \neq NP$. De momento nadie ha sido capaz de obtener precisión comparable a las redes neuronales profundas con árboles de decisión

en problemas complejos¹¹, de la misma forma en que nadie ha obtenido un programa que resuelva instancias de SAT en tiempo polinomial. Por supuesto, «la ausencia de evidencia no es evidencia de la ausencia», y es posible que en el futuro se descubran modelos interpretables que resuelvan problemas complejos con igual o mayor precisión que las redes neuronales profundas.

5. Contra la objeción de Goodfellow

En el año 2017 el científico de la computación Ian Goodfellow (uno de los creadores de las redes neuronales generativas adversariales, GANs), escribió lo siguiente con respecto a la importancia de la interpretabilidad en inteligencia artificial (Goodfellow, 2017):

Creo que la interpretabilidad es importante, pero no creo que debería reducir la adopción del aprendizaje de máquinas. Los humanos no somos interpretables tampoco, porque no sabemos realmente lo que nuestros cerebros están haciendo. Hay un montón de evidencia en el campo de la psicología que indica que las explicaciones que damos sobre el porqué de nuestras decisiones no corresponden a las razones que realmente están operando detrás de ellas. Recomendando un muy buen libro en el tema: «The Illusion of Conscious Will», que trata justamente sobre cómo creemos que nuestras vidas están controladas por nuestras mentes conscientes, pero en la práctica muchas de nuestras decisiones están guiadas por el subconsciente. [...] Desde este punto de vista, la IA nos da la oportunidad de tomar decisiones verdaderamente interpretables y explicables por primera vez, porque tenemos acceso a la descripción completa del modelo.

A mi juicio, el problema con esta objeción es que, incluso aceptando como premisa que las decisiones humanas no son realmente interpretables, los estándares de interpretabilidad para modelos de IA no debiesen ser los mismos que para seres humanos. En particular, creo que esta discusión es similar a la clásica

¹¹ Un desafío concreto en esta línea sería sobrepasar 80% de precisión en CIFAR10 (Krizhevsky and Hinton, 2009) mediante árboles de decisión (with Code, 2024b). Aún más complejo parece ser obtener más de 25 % de precisión en ImageNet (Deng et al., 2009) mediante árboles de decisión, donde redes neuronales profundas han sobrepasado 90 % (with Code, 2024a).

discusión sobre el «voto electrónico»; si bien es cierto que el voto tradicional, mediante papeles en urnas, no es perfectamente seguro (es casi imposible garantizar que ningún participante será capaz de introducir votos falsos), esto no implica que sea razonable adoptar el voto electrónico sin tener un altísimo estándar de seguridad. El argumento clásico en el caso del voto electrónico es que, en el caso en que algún actor logra efectivamente «hackear» el sistema, el daño potencial es infinitamente mayor que en el voto manual. Una vulnerabilidad electrónica podría permitir a un actor malicioso la posibilidad de determinar a su voluntad el porcentaje exacto de votos que recibe cada candidato, mientras que en el voto manual, el daño potencial está naturalmente limitado por restricciones físicas del fraude; incluso, si una persona lograra introducir miles de votos falsos en una urna, tal evento sería (con altísima probabilidad) insuficiente para cambiar el resultado de una elección nacional. De la misma manera, en otro escenario, si bien los procesos de entrevistas laborales conducidas por humanos están sujetas a sesgos de los entrevistadores y las entrevistadoras, el despliegue de modelos de IA en esta materia podría traer consigo sesgos afectando a millones de personas, y su detección sería potencialmente mucho más compleja si los modelos utilizados no son interpretables.

6. Preguntas para guiar nuestra convivencia

Para concluir este ensayo, quisiera retornar la atención al dilema de la interpretabilidad y la precisión. Como he dicho en la sección 2, creo que, en ausencia de regulaciones legales, las empresas utilizarán los modelos que les otorguen mayores beneficios financieros, típicamente medidos indirectamente a través de la precisión de los modelos utilizados. Por tanto, si queremos que las empresas e instituciones utilicen modelos interpretables, necesitamos regulación o alguna otra forma de incentivos que penalice el uso de modelos no interpretables. Para las ingenieras e ingenieros del presente y el futuro, la pregunta fundamental, desde un punto de vista ético, pareciera ser *¿cuánta precisión estamos dispuestos a sacrificar en pos de la interpretabilidad?*, o equivalentemente, *¿cuánta interpretabilidad estamos dispuestos a sacrificar en pos de la precisión?*

La respuesta a esta pregunta, por supuesto, dependerá del dominio de aplicación. Como dice Rudin (2019), en dominios médicos, legales o financieros es particularmente nocivo utilizar modelos no interpretables, y es probable, además, que sus sesgos afecten negativamente a poblaciones tradicionalmente desventajadas. Un ejemplo en la dirección opuesta es el trabajo de Romera-Paredes et al. (2024), en que se utilizan *«Large Language Models»*, modelos con billones de parámetros, para descubrir mejores soluciones a problemas matemáticos. En este caso, a pesar de la falta de interpretabilidad de los modelos utilizados, estos fueron capaces de descubrir pequeños conjuntos de vectores (i.e., «admissible sets») que pueden ser analizados manualmente y que resultaron en una mejor cota para el famoso «Cap Free Set Problem». En este caso, las soluciones encontradas por los modelos tienen valor en sí mismas, y si bien sería interesante tener una mejor comprensión del funcionamiento de los modelos utilizados, esto no parece ser fundamental por el momento en aplicaciones matemáticas. Las y los futuros ingenieros deberán cuestionarse seriamente, cada vez que deseen utilizar un modelo de IA, el grado de interpretabilidad que su aplicación requiere para garantizar una convivencia saludable con los seres humanos que serán afectados por las decisiones que tomen sus modelos.

Referencias bibliográficas

- Arenas, Marcelo, Báez, Daniel, Barceló, Pablo, Pérez, Jorge and Subercaseaux, Bernardo (2021). «Foundations of Symbolic Languages for Model Interpretability». In *Advances in Neural Information Processing Systems*, volume 34, pages 11690–11701. Curran Associates, Inc.
- Arenas, Marcelo, Barceló, Pablo, Romero Orth, Miguel and Subercaseaux, Bernardo (2022). «On Computing Probabilistic Explanations for Decision Trees». In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28695–28707. Curran Associates, Inc..
- S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2006. ISBN 978-0-521-42426-4. URL <https://theory.cs.princeton.edu/complexity/book.pdf>.
- Barceló, Pablo, Monet, Mikaël, Pérez, Jorge and Subercaseaux, Bernardo. «Model Interpretability through the lens of Computational Complexity». In *Advances in Neural Information Processing Systems*, volume 33, pages 15487–15498. Curran Associates, Inc., 2020a.
- Barceló, Pablo, Pérez, Jorge and Subercaseaux, Bernardo. *Foundations of Languages for Interpretability and Bias Detection*. AFCI, 2020b.
- Arriagada Bruneau, Gabriela (2024). *Los sesgos del algoritmo*. La Pollera Ediciones: Santiago. ISBN 9789566267256.
- Buolamwini, Joy and Gebru, Timnit. «Gender shades: Intersectional accuracy disparities in commercial gender classification». In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24, Feb 2018. URL <https://proceedings.mlr.press/v81/buolamwini18a.html>.

- Burke, Lilah. U of Texas will stop using controversial algorithm to evaluate Ph.D. applicants — insidehighered.com. <https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd>, 2020. [Accedido el 15-05-2024].
- Cobham, Alan. The intrinsic computational difficulty of functions. In Yehoshua Bar-Hillel, editor, *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress (Studies in Logic and the Foundations of Mathematics)*, pages 24–30. North-Holland Publishing, 1965.
- Darwiche, Adnan and Hirth, Auguste. «On the Reasons Behind Decisions». In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), pages 712–720, 2020. doi: 10.3233/FAIA200158.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Markus D. Dubber, Frank Pasquale, and Sunit Das, editors. *The Oxford Handbook of Ethics of AI*. Oxford University Press, July 2020. ISBN 978-0-19-006739-7. doi: 10.1093/oxfordhbk/9780190067397.001.0001.
- Edmonds, Jack. *Paths, trees, and flowers*. Canadian Journal of Mathematics, 17:449–467, 1965. doi: 10.4153/CJM-1965-045-4.
- Hinton, Geoffrey. Question — twitter.com @geoffreyhinton. <https://twitter.com/geoffreyhinton/status/1230592238490615816?lang=en>, 2020. [Accedido el 15-05-2024].

- Goodfellow, Ian. *How important is interpretability for a model in Machine Learning?* — quorasessionwithiangoodfellow. quora.com. <https://quorasessionwithiangoodfellow.quora.com/How-important-is-interpretability-for-a-model-in-Machine-Learning?ch=1&share=b5056dcf>, 2017. [Accedido el 19-05-2024].
- Goodman, Bryce and Flaxman, Seth. *European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”*. AI Magazine, 38(3):50–57, October 2017. doi: 10.1609/aimag.v38i3.2741.
- Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001). *The Elements of Statistical Learning. Springer Series in Statistics*. Springer New York Inc., New York, NY, USA.
- Izza, Yacine, Ignatiev, Alexey and Marques-Silva, Joao (2020). *On explaining decision trees*.
- Kirchner, Lauren, Larson, Jeff, Angwin, Julia and Mattu, Surya (2016). *How We Analyzed the COMPAS Recidivism Algorithm* — propublica.org. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. [Accedido el 15-05-2024].
- Kirchner, Lauren, Angwin, Julia, Larson, Jeff and Mattu, Surya (2016). *Machine Bias. There’s software used across the country to predict future criminals. And it’s biased against blacks*— propublica.org. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. [Accedido el 16-05-2024].
- Karatsuba, Ekaterina. Fast computation of the riemann zeta function. *s/ for integer values of s. Problems of Information Transmission*, 31:353–362, 01 1995.
- Kelly, Jack (2024). *Your Next Job Interview May Be With ‘Alex,’ The AI Interviewer* — forbes.com. <https://www.forbes.com/sites/jackkelly/2024/05/10/your-next-job-interview-may-be-with-alex-the-ai-interviewer/?sh=7cac4c4f76f2>. [Accessed 15-05-2024].

- Knuth, Donald (2014). *Twenty Questions for Donald Knuth* | | *InformIT* — [informit.com](https://www.informit.com/articles/article.aspx?p=2213858&WT.mc_id=Author_Knuth_20Questions). https://www.informit.com/articles/article.aspx?p=2213858&WT.mc_id=Author_Knuth_20Questions. [Accedido el 18-05-2024].
- Kozachinskiy, Alexander (2023). *Inapproximability of sufficient reasons for decision trees*.
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto.
- Lipton, Zachary C. *The mythos of model interpretability*. Queue, 16(3):31–57, 2018.
- Malmon, Judy (2023). *How Are Algorithms Used in the Criminal Justice System?* — [superlawyers.com](https://www.superlawyers.com/resources/criminal-defense/how-are-algorithms-used-in-the-criminal-justice-system/). <https://www.superlawyers.com/resources/criminal-defense/how-are-algorithms-used-in-the-criminal-justice-system/>. [Accedido el 15-05-2024].
- Marques-Silva, Joao (2023). Logic-based explainability in machine learning.
- Marques-Silva, Joao and Ignatiev, Alexey. *Delivering Trustworthy AI through Formal XAI*. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12342–12350, June 2022. doi: 10.1609/aaai.v36i11.21499.
- Marques-Silva, Joao and Ignatiev, Alexey. *No silver bullet: Interpretable ML models must be explained*. Frontiers in Artificial Intelligence, 6, 2023. ISSN 2624-8212.
- Miller, George A. *The magical number seven, plus or minus two: Some limits on our capacity for processing information*. Psychological Review, 63(2):81–97, 1956. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/h0043158.
- Miller, Tim. *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence, 267: 1–38, 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007.

- Molnar, Christoph (2022). *Interpretable Machine Learning*. 2 edition. URL <https://christophm.github.io/interpretable-ml-book>.
- O'Neil, Cathy (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA. ISBN 0553418815.
- Pearl, Judea and Mackenzie, Dana (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition. ISBN 046509760X.
- Piltaver, Rok, Luštrek, Mitja, Gams, Matjaž and Martinčič-Ipšić, Sanda (2016). *What makes classification trees comprehensible? Expert Systems with Applications*, 62:333–346, 2016. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.06.009.
- Pringle, Eleanor. *Bumble founder says your dating 'AI concierge' will soon date hundreds of other people's 'concierges' for you* — fortune.com. <https://fortune.com/2024/05/10/bumbles-whitney-wolfe-herd-dating-concierge-artificial-intelligence/>, 2024. [Accedido el 15-05-2024].
- Romera-Paredes, Bernardino, Barekatin, Mohammadamin, Novikov, Alexander, Balog, Matej, Kumar, M. Pawan, Dupont, Emilien, J. R. Ruiz, Francisco, Ellenberg, Jordan S., Wang, Pengming, Fawzi, Omar, Kohli, Pushmeet and Fawzi, Alhussein. *Mathematical discoveries from program search with large language models*. *Nature*, 625(7995): 468–475, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06924-6.
- Roughgarden, Tim (2021). *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press.
- Rudin, Cynthia. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.
- Sapia.ai. *Sapia.ai | Hire top talent, faster, with AI Smart Interviewing* — sapia.ai. <https://sapia.ai/>, 2024. [Accedido el 15-05-2024].

- Semenova, Lesia, Rudin, Cynthia and Parr, Ronald. *On the existence of simpler machine learning models*. In 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533232. URL <http://dx.doi.org/10.1145/3531146.3533232>.
- Talently.ai. Talently.ai: Your AI Interviewer — interview.talently.ai. <https://interview.talently.ai/>, 2024. [Accedido el 15-05-2024].
- Wegener, Ingo. *Branching Programs and Binary Decision Diagrams*. Society for Industrial and Applied Mathematics, January 2000. doi: 10.1137/1.9780898719789.
- Papers with Code. Papers with Code - ImageNet Benchmark (Image Classification) — paperswithcode.com. <https://paperswithcode.com/sota/image-classification-on-imagenet>, 2024a. [Accedido el 19-05-2024].
- Papers with Code. Papers with Code - CIFAR-10 Benchmark (Image Classification) — paperswithcode.com. <https://paperswithcode.com/sota/image-classification-on-cifar-10>, 2024b. [Accedido el 19-05-2024].

Un mundo nuevo descubierto a través de los datos

Camila Henríquez Beltrán¹

En los últimos años hemos presenciado cómo la inteligencia artificial [IA] y la ciencia de datos han tomado un rol protagónico en nuestra sociedad (Giattino, Mathieu, Samborska y Roser, 2023). Gracias a que estas herramientas son altamente eficientes, podemos verlas aplicadas por profesionales de distintas disciplinas y en diversos escenarios, desde el deporte, pasando por las finanzas, hasta la medicina (MIT Technology Review Insights, 2023). Según Pacheco (2021), Pep Guardiola, director técnico del equipo de fútbol Manchester City, contrató a cuatro astrofísicos para realizar diversos análisis basados en los datos obtenidos durante los partidos, utilizando Inteligencia Artificial y Big Data. Ante esto nacen las siguientes preguntas: *¿qué tienen que ver los astrónomos con los equipos de fútbol? ¿Por qué no contrataron especialistas en fútbol para esta labor?*

Cuando la acción de Guardiola fue noticia, me encontraba a la mitad de la licenciatura en astronomía; el grueso de mis estudios giraba alrededor de la física, las matemáticas y también la computación. Contrariamente a lo que se puede creer, hoy en día los astrónomos casi no pasan tiempo observando el cielo, ni mucho menos viendo directamente las estrellas o planetas a través de un telescopio. Según reportes de la Associated Universities Inc. [AUI] (s.f), los astrónomos de los Estados Unidos pasan entre 10 y 20 horas

¹ Licenciada en ciencias, mención astronomía. Científica de Datos. Dirección de Innovación del Instituto Milenio Fundamentos de los Datos. camila.henriquez.beltran@gmail.com

al año ocupando instrumentos de observación. Solo en contadas ocasiones, cuando los datos no están disponibles en los grandes bancos de datos astronómicos, se opta por tiempo de observación en alguno de los muchos observatorios repartidos en todo el mundo. Muchas de las tareas que se realizan a diario consisten en analizar grandes volúmenes de datos que se encuentran disponibles en la web, crear simulaciones computacionales, implementar modelos matemáticos y, más recientemente, en desarrollar soluciones que integran inteligencia artificial.

En este contexto, la contratación de analistas astrofísicos en el Manchester City desdibujó el camino profesional que se extendía frente a mí. El camino que en un momento creí único, que consistía en continuar la formación como investigadora, hacer un magíster en Chile, postular a programas de doctorado en el extranjero y realizar unos cuantos postdocs investigando en un área en particular, ya no era la única opción. Ahora veía otras posibilidades para el desarrollo de todas las habilidades que fui adquiriendo a lo largo del pregrado. Existía un mundo fuera de la academia y de la astronomía, lleno de oportunidades para aplicar estas herramientas y seguir aprendiendo sobre cómo funciona el mundo moderno. Para hacer esta transición hacia la industria, ahora en el rol de una cientista de datos, sería necesario considerar mucho más que solo mejorar las habilidades técnicas y traducirlas para poder aplicarlas a contextos fuera de la física en el espacio.

Ética en astronomía

La formación en ética que recibí a lo largo de la licenciatura estaba orientada a la ética de la investigación y al rol medioambiental del ingeniero. Gran parte de los casos estudiados se referían al manejo correcto de desechos para no afectar poblados cercanos. En otras ocasiones se conversaba sobre evitar el plagio o uso de información privilegiada en investigaciones, etc. Si bien estas temáticas y las discusiones que surgen a partir de ellas despertaron un interés en mí, sentía que este interés era más a nivel personal que algo directamente relacionado con mi campo de trabajo. En

este sentido, al observar los casos discutidos, notaba que ninguno hacía referencia a trabajos de astrónomos o físicos. Era posible identificar cómo compañeros que participaban en investigaciones de biotecnología se podían beneficiar directamente del estudio de casos, como la responsabilidad que se tiene al participar con sujetos de prueba en investigaciones farmacológicas. Sin embargo, en ese momento, a simple vista no lograba dilucidar cómo eso se traducía a investigaciones astronómicas.

Uno puede pensar que tanto la física como la astronomía consisten en investigaciones más teóricas, con menos aplicaciones prácticas, y que, dado que el objeto de estudio a veces ni siquiera se ubica en nuestro planeta, es difícil ver cómo estos trabajos afectan directamente a las personas. Sin embargo, incluso la física se ha visto envuelta en este tipo de conversaciones.

Un ejemplo emblemático es la historia del desarrollo de la bomba atómica, el cual es uno de los casos más discutidos sobre cómo una *idea* desarrollada por físicos teóricos terminó afectando a cientos de miles de personas (U.S. Department of Energy Office of Scientific and Technical Information [OSTI], s.f). Este caso surge de manera recurrente hasta el día de hoy. Incluso en 2024, a 79 años del suceso, la película que narra la historia y el conflicto interno que tuvo uno de los principales investigadores a cargo, obtuvo siete premios Óscar (Associated Press HOLLYWOOD, 2024). Además, pareciera que no es casualidad que, al estar hablando de las posibles implicancias éticas de modelos de inteligencia artificial, recurramos nuevamente es este escenario histórico (Kearns y Roth, 2019).

Sesgos en la formación de astrónomos

En relación con los estudios de sesgo, era común participar en discusiones que, en principio, hacen referencia a situaciones ajenas a la investigación o trabajo realizado. Estos casos generalmente se centran en la selección de individuos de grupos protegidos, ya sea por género, nacionalidad, u otros, para las becas de postgrado. Otra instancia clara en la que se observaba el sesgo de género era en las

mismas salas de clases, donde la mayor parte de los alumnos que conforman los cursos, y casi la totalidad de profesores que imparten clases, son hombres.

Quiero destacar que las instancias anteriormente mencionadas abordan la temática de los sesgos desde una perspectiva externa a las investigaciones y proyectos en los que participaba. Aunque estaban presentes en el ambiente y, en cierta medida influyen en mi rol de investigadora, me faltaba comprender la dirección contraria: en qué casos mi trabajo, «mi ciencia» como astrónoma, podía influir en los demás y, de esta manera, aprender efectivamente cómo mitigar los posibles sesgos que pudieran surgir.

Sesgos en detección de exoplanetas

Como he dicho, la astronomía no está exenta de sesgos, ni mucho menos. Como en cualquier investigación, ya sea observacional o experimental, es necesario mantener esta mirada siempre crítica respecto a las conclusiones que sacamos a partir de los datos, esto con el fin de asegurar la reproducibilidad de resultados, mantener el rigor y la responsabilidad de la ciencia en general (Ciubotariu, I.I. y Bosch, G., 2022).

Uno de mis casos favoritos en relación con el sesgo observacional se refiere a la técnica de detección de exoplanetas que utiliza la medición de Velocidades Radiales de las estrellas. De no ser por esta mirada crítica al momento de evaluar los resultados obtenidos con este tipo de detección, es probable que hubiésemos creído por unos cuantos años, que nuestro planeta Tierra era único en su composición y tamaño, ya que no lográbamos detectar otros planetas rocosos en las regiones del espacio que estábamos observando.

Un exoplaneta se define como cualquier planeta que esté ubicado fuera de nuestro sistema solar (European Space Agency [ESA], s.f). La tarea de encontrar planetas fuera de nuestro vecindario terrestre podemos compararla a encontrar una mota de polvo que

apenas refleja la luz de una linterna lejana. Debido a la lejanía de estos objetos y a su baja luminosidad, ya que no emiten luz propia y solo reflejan la de su estrella madre, es necesario recurrir a otros métodos que no sean detección directa por imágenes para encontrarlos (ESA, 2024). Uno de estos métodos es la medición de la velocidad radial de la estrella en relación con la Tierra. Este método utiliza la ley de gravitación universal, que describe la interacción entre dos cuerpos de masa m_1 y m_2 (Figura 1). Lo importante de esta ecuación es que la fuerza F que el cuerpo 1 ejerce en el cuerpo 2, es la misma que el cuerpo 2 ejerce en el cuerpo 1.

Para llevarlo a un escenario más familiar: así como el Sol ejerce una fuerza sobre la Tierra, haciendo que esta orbite a su alrededor, esa misma fuerza la está ejerciendo la Tierra sobre el Sol. Sin embargo, debido a la gran masa del Sol, prácticamente no vemos un movimiento en él.

$$F = G \frac{m_1 m_2}{r^2}$$

Figura 1: Ley de gravitación universal de Newton.

A medida que aumentamos la masa del planeta y mantenemos la masa de la estrella, esta fuerza F aumenta directamente. En nuestro sistema solar, donde Júpiter es el planeta más grande, podemos medir el impacto que tiene esta fuerza ejercida por Júpiter sobre el Sol mediante instrumentos espectroscópicos. Observamos los pequeños movimientos en el espectro de la estrella que se mueven a intervalos regulares y que siguen el ritmo del año joviano, o lo que es equivalente a 12 años terrestres (Hawai'i Space Grant Consortium, s.f.).

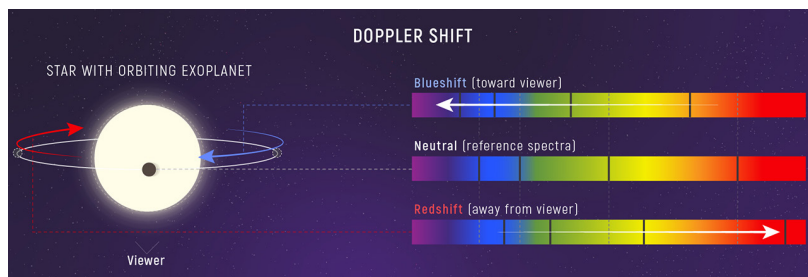


Figura 2: Ejemplo de efecto Doppler en espectro de una estrella debido a la fuerza gravitacional del planeta que la órbita.

Fuente: NASA, ESA, Leah Hustak (STScI) como se citó en Webb Space Telescope, 2021

Mediante la observación de los espectros de estas estrellas lejanas, podemos medir estos movimientos y, si no hay objetos luminosos cercanos, se puede inferir que es un planeta el que los causa (Figura 2). A mayor fuerza entre los objetos, mayor oscilación veremos en las líneas del espectro, lo que facilitará la detección de estos movimientos periódicos en la estrella observada. A partir de la ecuación, podemos notar que mientras más cercanos estén ambos cuerpos (es decir, un radio menor) y cuanto mayor sea la masa del planeta, más fácil será detectar este movimiento en el espectro.

En la Figura 3 podemos observar este sesgo heredado por nuestras limitaciones asociadas a la observación. Los exoplanetas descubiertos con el método descrito anteriormente están representados con una cruz. La mayoría de estos planetas tienen una masa del orden de 100 veces la masa de la Tierra o superior y se encuentran ubicados a menos de 100 parsecs de distancia.

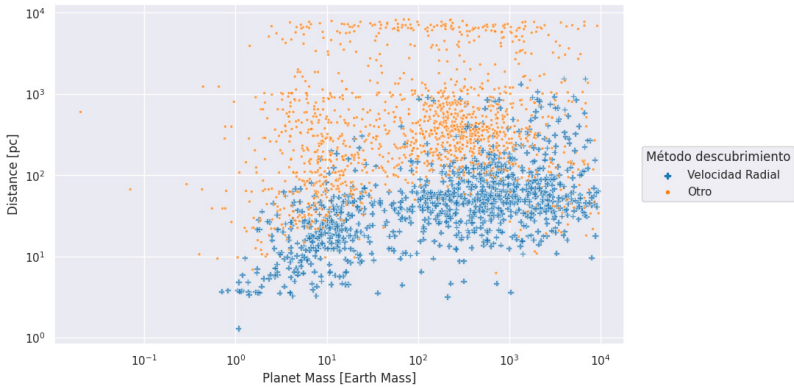


Figura 3: Relación entre la masa del exoplaneta y la distancia a la que se encuentra según el método de detección utilizado.

Fuente: NASA Exoplanet Archive, s.f.

La mayoría de los exoplanetas detectados por este método de velocidad radial están muy cerca de su estrella, planetas más masivos o ambos (NASA Exoplanet Archive, s.f.). Es posible detectar exoplanetas de masas similares a la Tierra si se observa a estrellas más cercanas. Sin embargo, si queremos alejarnos y observar este tipo de planetas en regiones más profundas del universo, serán necesarios telescopios más avanzados que puedan captar estos pequeños bamboleos de los espectros.

Hasta el año 2006, la mayoría de los exoplanetas se descubrieron con este método (NASA Exoplanet Archive, s.f.). Si en ese punto de la historia hubiésemos querido describir cómo son los planetas fuera del sistema solar, es probable que nos hubiéramos quedado con la imagen de planeta gigante, probablemente gaseoso, similar a Júpiter. Si pasamos por alto este sesgo observacional, y desarrollamos algún modelo basado en los datos de exoplanetas hasta esa fecha, los únicos «afectados» serían toda la variedad de exoplanetas que sabemos que existen gracias a la detección con otras metodologías. En este escenario, somos nosotros, los investigadores, los que nos vemos afectados por estas conclusiones apresuradas.

Puede que estas comparaciones dejen de ser válidas en unas décadas. Tal vez sea solo cuestión de tiempo para que nos encontremos en un mundo con una tecnología lo suficientemente avanzada, como para que las investigaciones en astronomía no solo afecten a estrellas, planetas u otros objetos celestes. Hasta que nosotros como especie estemos en contacto constante y habitemos las áreas en las que se desarrolla la investigación astronómica, no seremos plenamente conscientes del posible impacto que estos trabajos pueden tener en nosotros como individuos y como sociedad. Como astrónomos, solo nos queda esperar y estar atentos a los dilemas y problemáticas que surgen en áreas hermanas, con la esperanza de que un escenario como el de la bomba atómica no se nos escape de entre los dedos antes de poder preverlo.

Transición a la industria

Al momento de pasar de la academia a la industria, tuve que adaptar en ciertos aspectos mi metodología de trabajo. En la academia, los proyectos tienden a ser a largo plazo y tienen un enfoque en la exploración y en el descubrimiento de nuevos conocimientos, existiendo, también, la libertad de profundizar en una sola temática durante años. En contraste, en la industria, los proyectos suelen estar orientados a resultados concretos y prácticos, los plazos son más acotados y concentrados en la aplicabilidad inmediata. Además, en la industria, el trabajo se realiza usualmente con equipos multidisciplinarios, por lo que las habilidades de colaboración y comunicación son clave en este ambiente. Esta transición me obligó a ser más flexible y a desarrollar una mentalidad orientada a solucionar nuevos problemas de distintas áreas de manera rápida y eficaz. Al mismo tiempo que aprendía a adaptar mi forma de trabajo a las exigencias de la industria, necesitaba mantenerme al día con las investigaciones en curso y seguir formándome en inteligencia artificial, un campo que avanza a un ritmo acelerado.

La inteligencia artificial es un mundo en constante evolución, con nuevas técnicas, herramientas y enfoques surgiendo continuamente. Esto requería no solo estar al tanto de las

últimas investigaciones y desarrollos, sino también, aplicar estos conocimientos de manera práctica en mis proyectos. La combinación de adaptabilidad en el trabajo y la actualización constante en este campo dinámico, fue fundamental para mi crecimiento profesional y para aportar valor a los equipos en los que trabajaba.

Me gustaría reflexionar sobre cómo la necesidad de adaptarse rápidamente a la industria, puede relegar a un segundo plano la preocupación por cómo estas soluciones desarrolladas afectan a nivel de sesgos o implementaciones éticas. Esto es especialmente relevante, considerando que, como «investigadores de lo teórico», no es una tarea innata evaluar cómo lo que desarrollamos terminará impactando al usuario final.

En el ritmo acelerado de la industria, donde la eficiencia y los resultados rápidos son altamente valorados, es fácil que las consideraciones éticas y los posibles sesgos en nuestras soluciones pasen desapercibidos. La presión por entregar resultados de manera expedita y la emoción de trabajar con tecnologías de vanguardia pueden nublar nuestro juicio, haciendo que olvidemos evaluar el impacto más amplio de nuestro trabajo.

Invito a todos los recién graduados a que se adentren en el mundo de la inteligencia artificial en la industria y a no dejarse llevar únicamente por la emoción de las soluciones con IA, a no ignorar las posibles implicancias de sus desarrollos. Es crucial recordar que, ahora, nuestro trabajo no afecta a las estrellas, sino que tiene un impacto directo en los individuos y en la sociedad de la que formamos parte. Mantener siempre presente esta responsabilidad ética es fundamental para garantizar que nuestras contribuciones sean beneficiosas y justas.

Referencias bibliográficas

- Associated Press HOLLYWOOD. (11 de marzo de 2024). «Oscar 2024: Toda la lista de ganadores». *Los Angeles Times en Español*. <https://www.latimes.com/espanol/entretenimiento/articulo/2024-03-10/lista-de-ganadores-de-los-premios-oscar>
- Associated Universities Inc [AUI]. (s.f). So you want to be an astronomer?. <https://aui.edu/so-you-want-to-be-an-astronomer/>
- Ciubotariu, I.I. & Bosch, G. (15 mayo de 2022). Improving research integrity: a framework for responsible science communication. <https://bmccresnotes.biomedcentral.com/articles/10.1186/s13104-022-06065-5>
- European Space Agency [ESA]. (1 septiembre de 2019). Exoplanet Detection Methods. <https://sci.esa.int/web/exoplanets/-/60655-detection-methods>
- European Space Agency [ESA]. (s.f). ¿Qué son los exoplanetas? https://www.esa.int/Space_in_Member_States/Spain/Que_son_los_exoplanetas
- Giattino, C., Mathieu, E., Samborska, V. & Roser, M. (2023). Artificial Intelligence. Our World in Data. <https://ourworldindata.org/artificial-intelligence>
- Hawai'i Space Grant Consortium. (s.f). How Old Are You?. https://www.higp.hawaii.edu/spacegrant/old/class_acts/HowOld.html
- Kearns, M. & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.
- MIT Technology Review Insights. (5 octubre de 2023). Driving company wide efficiencies with AI. <https://www.technologyreview.com/2023/10/05/1079726/driving-companywide-efficiencies-with-ai/>

- NASA Exoplanet Archive. (s.f). Planetary Systems [Base de datos en línea]. https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS&constraint=default_flag%3E0&constraint=disc_method%20like%20%27%25rv%25%27
- Pacheco, J. (24 marzo de 2021). La innovación en el fútbol: los astrofísicos de Guardiola, los datos para fichar y las criptomonedas. *El Español*. https://www.elespanol.com/deportes/futbol/20210324/innovacion-futbol-astrofisicos-guardiola-datos-fichar-criptomonedas/568444762_0.html
- U.S. Department of Energy Office of Scientific and Technical Information [OSTI]. (s.f). The atomic bombing of Hiroshima. <https://www.osti.gov/opennet/manhattan-project-history/Events/1945/hiroshima.htm>
- Webb Space Telescope. (2 julio de 2021). Doppler Shift of a Star's Spectrum. <https://webbtelescope.org/contents/media/images/01F8GFCAM7Q0EMKZ3QFAF9AQJ3>

Propuesta de modelo para la formación ética mediante la discusión de dilemas morales

Pablo Ramírez Rivas, Josefa Cerda Maureira,
Jorge Martínez Silva¹

1. La formación ética

La ética consiste en un saber práctico que, mediante la *reflexión y la deliberación, orienta fundamentada y responsablemente las decisiones y acciones que tomamos y realizamos para obrar bien en el conjunto de nuestra vida, en contextos siempre complejos, según ciertos principios morales*. Ella requiere, para realizar su cometido, que las personas forjen su carácter, de forma tal, que las capacite para dicha deliberación. Se trata, por tanto, de un «saber práctico, preocupado por averiguar cuál debe ser el fin de nuestra acción, para que podamos decidir qué hábitos hemos de asumir, cómo ordenar las metas intermedias, cuáles son los valores por los que hemos de orientarnos, qué modo de ser o carácter hemos de incorporar, con objeto de obrar con prudencia, es decir, tomar decisiones acertadas» (Cortina 2000, pp. 20-21). Ese saber práctico no busca solo un objeto o producto concreto (que es lo que ocurre con la técnica y el arte, por ejemplo), sino que es «más ambicioso, [quiere] enseñarnos a obrar bien, racionalmente, en el conjunto de nuestra vida entera» (Cortina 1996, p. 22).

La enseñanza y el aprendizaje de la ética profesional, por su propia naturaleza, tiene un carácter más bien helicoidal que lineal.

¹ Pablo Ramírez Rivas, Doctor en Filosofía, profesor de ETHICS, FCFM-UChile. pabramirez@uchile.cl; Josefa Cerda Maureira, Ingeniera civil industrial por la Universidad de Chile, se desempeña en ETHICS, FCFM-Uchile, jofacerda@uchile.cl; Jorge Martínez Silva, Licenciado en Filosofía. Asesor Unidad de Ética de ETHICS, FCFM-UChile. jorgemarsil@uchile.cl.

Esto significa que el contenido procedimental de la ética –entendida como una capacidad decisional–, que es el discernimiento ético, se busca desarrollar cada vez que se realiza una actividad o situación de aprendizaje. Así, va progresando en la dificultad del proceso cognitivo que hace la persona. Cada nivel subsume los anteriores integradamente. Los principios y valores por promover (que constituye parte del contenido conceptual a enseñar), no varían en su desarrollo, pero sí en su aplicación, puesto que están siempre sujetos a una interpretación situada, siendo su aplicación adecuada a cada actividad curricular en que se propicie el desarrollo de la competencia.

Es posible determinar la progresión de la competencia, pues a medida que se avanza en la formación profesional se van introduciendo nuevos temas y marcos morales, estudiando distintos escenarios con complejidades diferentes. Por lo tanto, dicho metafóricamente, la enseñanza y aprendizaje de la ética, podrían verse como una hélice que va engrosando a medida que transcurren los niveles de formación.

El texto que sigue presenta un modelo didáctico pedagógico para la formación ética y el trabajo con dilemas morales que sistematiza el conocimiento, trabajo y experiencia de una unidad de ética de una universidad pública chilena.

2. Marco didáctico pedagógico

Desde el punto de vista didáctico, la ética tiene dos grandes modalidades de presentación en el aula universitaria: como contenido disciplinar y como contenido transversal. De modo general, la primera modalidad enfatiza la enseñanza de contenidos específicos de orden conceptual. La segunda modalidad, por su parte, se aboca a los contenidos actitudinales y procedimentales. En este sentido, para ambas modalidades se presenta la dificultad de realizar un trabajo sobre las ideas e intuiciones que ya traen los/as estudiantes, considerando que conocen y experimentan un significado previo de los conceptos de ética y moral, y su aplicación como norma o escala

de valores, producto de diferentes factores, que pasan por lo social, familiar, formativo, etc. (Aisenberg, 1994). En la formación ética es imprescindible tener en cuenta lo recién mencionado. Otro factor por considerar es la etapa de la vida y edad en que se encuentran: cuanto mayor es la edad de los/as estudiantes, mayor y más asentado es también el bagaje moral que traen al ámbito de la educación universitaria sistemática.

La formación ética posibilita que los/as estudiantes adopten una actitud crítica frente a contenidos y situaciones ya conocidos e interiorizados para que así, como consecuencia esperada, puedan modificar y/o matizar algunas de las ideas o intuiciones adquiridas previamente. De acuerdo con la literatura, una forma concreta de provocar un cambio conceptual en los/as estudiantes es a partir de un conflicto cognitivo, que les permite darse cuenta de que los conceptos existentes en sus esquemas mentales y morales no necesariamente son los únicos que existen (Weissmann, 1993; Carretero, Pozo y Asensio, 1989). Ello se puede dar, por ejemplo, enfrentándoles a situaciones conflictivas y/o dilemáticas sobre las cuales deben tomar decisiones o posturas, y que, además, deben debatir con sus pares y profesores/as. En ese sentido, una herramienta didáctica eficaz para la enseñanza de la ética, es la exposición y aplicación de dilemas morales. El intercambio y contraste de ideas, percepciones y sentimientos con pares, incrementa el conflicto cognitivo hasta el punto en que se hace preciso buscar una decisión consensuada que satisfaga a los/as participantes en una convergencia crítica (Meza, 2008).

El paradigma pedagógico en que se basa el modelo para el trabajo con dilemas, es el constructivismo. Según Pulgar (2005) el aprendizaje es el «proceso mediante el cual una persona adquiere destrezas o habilidades prácticas (motoras e intelectuales), incorpora contenidos formativos o adopta nuevas estrategias de conocimiento y/o acción» (p. 19). El aprendizaje es un proceso activo de construcción de conocimiento y es una actividad situada en un contexto particular y específico (Chadwick, 1999). Los dilemas morales proporcionan, precisamente, ese contexto, tal como lo expondremos más adelante.

El concepto de *aprendizaje significativo*, acuñado por el psicólogo y pedagogo David Ausubel, dice relación con el aprendizaje que logra un sujeto al relacionar nuevas ideas con aquellas que tenía previamente (Ortiz, 2015). De esta combinación surge una significación única que se distingue del aprendizaje repetitivo y memorístico. Es así como los *conocimientos previos* desempeñan un rol importante en la adquisición de nuevos conocimientos. Por lo tanto, en el trabajo con dilemas es de vital importancia tenerlos en cuenta, ya sea para potenciar y asentar la fundamentación de una determinada postura o para cambiarla.

Para la formación de estudiantes, en general, y el trabajo con dilemas en particular, el/la profesor/a o docente debe considerar que el sujeto interactúa constantemente con el ambiente que lo rodea (Araya, Alfaro y Andonegui, 2007). Por lo tanto, debe proveer de oportunidades de aprendizaje a través de un ambiente estimulante que permita que el/la estudiante supere las etapas que se correspondan con su desarrollo cognitivo. Al ir adquiriendo nueva información respecto a las propias decisiones y los criterios con que estas se toman, el/la estudiante incorpora estos esquemas y estructuras preexistentes a su red cognitiva, contribuyendo así a la construcción de nuevas estructuras de pensamiento (Tunnermann, 2011) que aportarán a sus actuaciones o decisiones posteriores.

En línea con lo anterior, el constructivismo social promovido por Vygotsky sostiene que la creación de conocimiento es una actividad compartida más que individual, existiendo una relación recíproca entre el sujeto y su contexto. Según este autor, la *zona de desarrollo próximo* es la distancia que existe entre lo que una persona puede aprender por sí misma y lo que podría aprender con la ayuda de un experto (Papalia, Wendkos y Duskin, 2007). Esta es la zona donde la acción de el/la profesor/a, guía o formador es de especial importancia, concibiendo así a la/el docente como «facilitador/a» del desarrollo de estructuras mentales que aumentan la complejidad del aprendizaje en el/la estudiante. Se da así un *intercambio dialéctico entre docentes y estudiantes*, el cual tiene como producto, una síntesis productiva –el aprendizaje– que se logra por medio del diálogo, la confrontación de ideas y la discusión; aspectos fundamentales en la formación ética

y del trabajo que proponemos con dilemas morales. Según Nieda y Macedo (1997), «el profesor adquiere especial protagonismo al ser un agente que facilita el andamiaje para la superación del propio desarrollo cognitivo personal» (p. 43). Esto lo logra por medio de acciones tales como: la estimulación de la autonomía y la iniciativa; el uso de las respuestas de los estudiantes para orientar las clases y modificar el contenido junto con las estrategias de enseñanza; la generación de un diálogo desequilibrante entre estudiantes y entre profesor/a y estudiantes; el estímulo de la curiosidad a través de preguntas abiertas, desafiantes y profundas; y la recopilación de la comprensión de los/as estudiantes antes de entregar su propia comprensión.

Siguiendo la idea de la doble formación propuesta por Vygotsky, en la que se afirma que todo aprendizaje surge primero en el plano interpersonal, para luego ser reconstruido en el plano intrapersonal (Tünnermann, 2011), es fundamental que los/as estudiantes tengan la oportunidad de contrastar sus propias posturas con la de otras personas, para así asentar e internalizar el conocimiento, ya que, como ya se mencionaba, cuando el proceso de aprendizaje se controla internamente y se integran las nuevas competencias a la estructura cognitiva, entonces se produce el desarrollo cognitivo. Esto tiene como consecuencia una mayor adaptación al medio ya que si el/la estudiante universitario/a realiza un proceso constructivo en el que parte de sus intuiciones, su moral, sus experiencias, su propio sentir, las analiza, las contrasta, hace inferencias y luego vuelve sobre sus intuiciones morales iniciales, este recorrido supone un estado de conocimiento mayor al dado en la situación inicial. En este aspecto, el trabajo con dilemas debe ser una actividad compartida en una comunidad de aprendizaje, en la cual, estudiantes y profesores/as están implicados e interactúan recíprocamente en pos del conocimiento colectivo y personal de cada integrante de esa comunidad.

3. Los dilemas morales como medios para la formación ética

Investigaciones como la de Giménez (2018), afirman que sería un error introducir asignaturas en el plan de estudios que trataran el tema ético de manera general, sin diferenciar el contexto en que estos se imparten. Se trataría, más bien, de articular una ética aplicada y, a partir de ella, estudiar casos reales –tanto de la vida estudiantil como de la profesional– para llegar a desarrollar la capacidad de tomar decisiones y reflexionar sobre el propio modo de actuar de forma consciente, responsable y argumentada. Esta ética aplicada ya fue delineada tempranamente por Cortina (1996b), quien propone la siguiente articulación de temas y contenidos: determinación del fin específico, el bien interno por el que cobra sentido y legitimidad una profesión; determinar los medios adecuados para producir ese bien; indagar qué virtudes y valores es necesario incorporar a la propia identidad, carácter y perfil profesional; identificar los valores de la moral cívica del contexto en que se desarrolla una persona y, con ello, el marco normativo relevante (los derechos que ese contexto reconoce); averiguar qué valores de justicia exige realizar el ámbito en el que se desarrollan las personas (vida académica, profesión, etc.) desde una perspectiva crítica de la ética del discurso; incorporar a las personas afectadas a las decisiones que las afectan. Los dilemas morales –relatos ante los cuales se debe elegir entre dos posibles cursos de acción (o de postura)– posibilitan precisamente esa articulación requerida por una ética aplicada, pues nos permiten trabajar integral y articuladamente estos aspectos señalados por la autora.

Otro elemento central para el aprendizaje es el trabajo situacional (Omart, 2010). Tal como se mencionó anteriormente, el aprendizaje es un proceso activo de construcción de conocimiento y está situado en un contexto particular y específico. El abordaje de la ética en un plan de estudios, ya sea en una asignatura específica, ya sea de forma transversal en el interior de cursos de especialidad de la carrera de la que se trate, realizado mediante el trabajo con dilemas morales, permite desarrollar la capacidad de la deliberación o discernimiento ético, lo cual constituye en sí mismo, un tema central para la ética. En situaciones dilemáticas también se debe

considerar el peso que tiene la experiencia previa en los aprendizajes de los/as estudiantes universitarios, en particular, cuando estos/as se encuentran en etapas avanzadas en sus carreras o que estén haciendo o hayan hecho ya alguna práctica profesional.

Hay diversos modelos de dilemas morales (Ruiz-Cano et al., 2015), pero los elementos en común que contienen todos ellos son los siguientes: 1) contener dos vías de acción (y/o dos posturas posibles); 2) cada una de esas vías contiene aspectos, implicancias y/o consecuencias morales positivas y negativas; 3) ante el dilema no se puede elegir una vía de acción, es decir, se impone la necesidad de tomar una decisión.

Las dos posibles vías de acción (o de postura) de un dilema moral acarrear consigo una carga positiva (por sus consecuencias, las motivaciones que haya, la adecuación a normas, etc.) y una negativa. Generalmente los dilemas descansan, en lo que respecta a su valoración positiva o negativa, en las consecuencias y/o impactos de la decisión que se tome. Pero la dimensión dilemática puede gravitar también en las normas que entran en conflicto, o bien, en las motivaciones y/o asuntos de conciencia del sujeto que debe tomar una decisión. Los dilemas ficticios que no se atienen a situaciones reales y/o verosímiles y cercanas a la experiencia de las personas, no suelen ser tan adecuados para el desarrollo y la formación de un juicio moral, aunque pueden ser un primer paso para mostrar que hay situaciones, en las cuales, la decisión acarrea una complejidad que obliga a detenerse para reflexionar más profundamente sobre la decisión a tomar. Por ello, es más apropiado un relato dilemático que sea real o verosímil, sirviendo ambos para la formación ética (Meza Rueda, 2008).

El trabajo con dilemas de este tipo obliga, por una parte, a decidir y, por otra, a tener que reflexionar y fundamentar la propia elección. La utilidad de la aplicación y exposición de dilemas morales deriva en la necesidad de reconstruir una trama de significados que surge de la vida cotidiana y se orienta a la futura vida profesional. En este proceso siempre es sugerible la construcción del conocimiento a través de la relación entre ideas previas e ideas adquiridas, para lograr

una diferenciación clara entre sesgos y condicionamientos y, así, incluir como contenido crítico, reflexivo y problematizador, las propias experiencias y prácticas sobre lo justo, lo injusto, las normas, el valor, las recompensas, las formas de vida, etc. Reflexionar sobre la propia práctica, en términos de Maliandi (1991), implica problematizar el propio *ethos*, y supone un compromiso mayor que la reflexión sobre prácticas ajenas. La formación ética mediante dilemas (y humanista, en general), permite a los/as estudiantes contextualizar los proyectos que desarrollan, o plantearse cuestiones sobre, por ejemplo, el efecto de su obra en las futuras generaciones (Giménez, 2018), o sobre la diversidad de culturas y etnias y la importancia de valorarla, sobre los límites de la responsabilidad profesional, etc.

Las ventajas de trabajar con dilemas se traducen en que, siendo la persona un ser pensante, sensible y actuante, ellos colaboran en reconocer e integrar las propias disposiciones afectivas, así como, la propia racionalidad y subjetividad; también tienen un efecto positivo en la comunicación con otros/as; sirven para distinguir la calidad de los argumentos que presentan las partes (entre los cuales se encuentran los propios) al defender sus propias posiciones; se aprecian el valor del razonamiento y las críticas de quienes toman postura distinta para el propio desarrollo; se reconocen los valores morales que entran en juego o se pretenden salvaguardar dentro de una determinada toma de decisiones; se diferencian los conflictos morales (y los dilemas propiamente dichos) de situaciones problemáticas que necesitan simplemente de una solución técnica; se utiliza el diálogo y la argumentación, expresada en procesos razonados y razonables, como medio para resolver los conflictos y, por demás, deslegitiman el uso de la imposición de la fuerza, verbal o física, como forma más o menos común para solucionar un conflicto (Lind, 1985).

Cuando se discuten dilemas morales, los sujetos se implican de alguna forma con sus personajes y situaciones. Eventualmente, se corre el riesgo de que se abran espacios de consejería en donde se dice lo que hay que hacer o lo que se debe evitar. Frente a esto, hay que tener claro que la discusión con dilemas no es una técnica terapéutica (Meza, 2008), aun cuando pueda serlo para algunas personas; sin embargo, el/la docente no es el profesional idóneo para este fin. Por

lo tanto, el foco de interés no debe estar centrado en «sanar» heridas producidas por decisiones erradas de los sujetos participantes, aunque la propia experiencia y la reflexión, en retrospectiva, puede ser un valioso insumo para resolver los dilemas. Contrariamente, se debe privilegiar el proceso de razonamiento moral que le permitirá a los sujetos llegar a juicios morales más robustos. Para lograrlo, es necesario que el caso presentado destaque y diferencie claramente las alternativas planteadas y su legitimidad, por lo cual también requiere del conocimiento de marcos legitimadores como teorías éticas, normativa vigente y valores cívicos ampliamente aceptados, además de que el relato contenga datos, información y situaciones reales o verosímiles.

En el caso de relatos verosímiles, es importante que contengan datos e información real, aunque la situación pueda ser imaginada. Debe aparecer, también, claramente delineado, el/la personaje principal y las opciones sobre las cuáles debe elegir. La construcción del dilema ha de ser clara y unívoca. El/la profesor/a debe tener seguridad de que el dilema contenga la mayor y mejor información posible y que ha sido entendido por el grupo. Adicionalmente, debe proporcionar las explicaciones necesarias para que todos/as los/as participantes tengan claridad sobre la materia a discernir. El exceso de detalles puede distraer a los sujetos sobre aspectos que no interesan verdaderamente. Es preciso disponer de preguntas exploratorias que permitan comprobar que se ha entendido el problema planteado. Este tipo de preguntas pueden, también, ser útiles cuando el diálogo derive hacia cuestiones poco relevantes o cuando los/as asistentes necesiten ayuda (Meza, 2008).

Los sucesos que se presentan en los dilemas morales deben estar centrados sobre cuestiones de valor moral y, adicionalmente, el planteamiento del dilema debe terminar siempre con preguntas del tipo: «¿Qué debe/debería hacer?» / «¿Qué es lo éticamente aconsejable en esta situación?», etc., con el objetivo de que aparezca de forma clara la obligación moral de quien protagoniza el relato (Meza, 2008). Una vez que los sujetos han dado a conocer su posición sobre lo que el agente debería o no debería hacer, es muy importante explorar las razones por las cuales se tiene tal posición.

De forma previa a la discusión del dilema, parece adecuado que se destine un tiempo para exponer las bases conceptuales a los/as estudiantes, aunque no sea de manera exhaustiva. El grupo ha de tener claridad sobre lo que se entiende por valores, ética, moral, desarrollo y conciencia moral, discernimiento ético, marcos normativos aceptados y teorías éticas pertinentes. Ahora bien, de la misma manera que un paciente no está obligado a saber sobre los diferentes tratados de medicina, pero sí a saber si tiene buena salud, tampoco es absolutamente necesario que los/as estudiantes sean filósofos/as morales; lo que es esperable es que sí reconozcan cómo es su razonamiento moral y el grado de desarrollo de su conciencia moral.

Finalmente, como la discusión de dilemas morales ocurre regularmente en un ambiente grupal, siempre será conveniente tener en cuenta las normas comunes para el diálogo grupal: el respeto por el/la otro/as y el uso de la palabra, la participación de las minorías, la escucha atenta de los argumentos, el rol imparcial del facilitador, el manejo del tiempo, la atención sobre la motivación del grupo y la posible dispersión en torno al tema en discusión.

4. Estructura de un dilema

Existen distintos tipos de dilemas o distintas maneras de construir y presentarlos (cfr. (Ruiz-Cano et al., 2015). La forma que adquieren dependerá de diversos factores, tales como, el nivel de complejidad que se le quiere dar, el formato en el que se trabajará (mediante plataformas digitales, presencialmente, etc.), el número de estudiantes con que se trabajará, entre otros.

El siguiente es un ejemplo de estructuración de un dilema para su presentación (Meza, 2008):

Fase 1: introducción al dilema

- Entrar en contacto con el dilema moral particular.
- Clarificar la situación que genera el dilema.
- Identificar los argumentos iniciales y discutir en grupo.

Estos propósitos se convierten en actividades que se desarrollan en un breve lapso (15-20 minutos).

Fase 2: debate inicial

- Poner sobre la mesa los argumentos existentes para analizar el dilema.
- Identificar los componentes morales contenidos en la situación dilemática, explorar y tomar conciencia de la propia postura y los argumentos que la sostienen.
- Confrontar con otros argumentos diferentes e introducir estrategias que cultiven competencias del juicio moral, actitudes, valores y dinámicas democráticas en el grupo.

Fase 3: profundización para estimular el desarrollo moral

- Complejizar el dilema, agregando facetas al problema.
- Evidenciar circunstancias concomitantes de la situación añadiendo nueva información, real o hipotética que precise o amplíe el problema original.

Es muy importante que el/la docente no conduzca esta parte del análisis a una u otra posición, pero que sí esté atento/a los diversos procesos que ocurren en la discusión para mediar en caso de necesidad (como, por ejemplo, que la discusión se desvíe a ámbitos no esenciales del dilema, o que, en vez de presentar las posturas y argumentarlas, se vuelva una conversación anecdótica).

La complejización lleva a la profundización del tópico generador. Como cada dilema tiene un «foco», y se trata de profundizar en este, es necesario aclarar cada uno de los puntos relevantes que lo componen. En este sentido, se puede buscar información en situaciones similares, así como teorías o datos que ayuden a comprender mejor lo que está en juego. El objetivo consiste en llegar a asumir responsablemente cada una de las situaciones o respuestas a las que se pueda llegar.

La profundización del dilema entrega la base para la diferenciación y la coordinación de perspectivas. Cuando se entienden las consecuencias, se logra deducir a quién pueden hacer daño o bene-

ficiar. Se debe pensar: si esto nos pasa a nosotros, ¿qué sucedería? (Meza, 2008), pues cuando son otras las personas que están en juego es relativamente fácil elegir, pero cuando somos nosotros/as el centro o quienes son afectados/as, la perspectiva cambia y, por lo general, se enriquece. Esto es lo que conduce a la «desestabilización» del juicio moral (Meza, 2008). Se trata de pensar en un nivel más complejo en comparación con aquel en que se ubica originalmente el sujeto. De este modo, se conduce al grupo para que piense en situaciones diversas a las que posiblemente esté acostumbrado. Se pueden introducir proyecciones para la vida profesional, para situaciones similares o casos conocidos. Para esto es necesario guiar con preguntas más contundentes y situaciones cada vez más complejas y ricas de matices que hagan pensar a la/el estudiante en posibilidades diversas a las que ya puede haber encontrado.

4.1 Ejemplo de dilema moral

En el trabajo con dilemas, la Unidad de ética de ETHICS ha confeccionado una serie de relatos (cfr. Ramírez, Guerrero, Cerda, Ross y Flores, 2022; Cerda, Carvajal, Ramírez y Hurtado, 2023) que contienen y cumplen con las condiciones descritas. A continuación, exponemos un ejemplo de dilema que puede ilustrar lo planteado hasta aquí. Se trata del *Caso Amanda*², el cual contiene un dilema moral sobre la integridad académica, en particular, sobre la honestidad:

La tuberculosis es una de las metas del objetivo 3 de los Objetivos de Desarrollo Sostenible (ODS) de la Organización Mundial de la Salud (OMS). La morbilidad total de tuberculosis en todas sus formas (TBTF) -la suma de casos nuevos y las recaídas- fue de 15,6 por 100.000 habitantes para el año 2017 (MINSAL, 2018a). Una incidencia menor a 10 por 100.000 habitantes define a un país de baja incidencia (MINSAL, 2018a). Respecto a los grupos de riesgo, se tiene que adultos mayores y extranjeros (principalmente latinoamericanos) presentan la mayor proporción de casos de TBTF (21,4% y 15,5% respectivamente) en Chile (MINSAL, 2018a).

² Dilema elaborado por la Unidad de Ética de ETHICS en conjunto con AEDIA.

Amanda es ingeniera en computación y cumple la función de cientista de datos en IA-Salud, una consultora que presta asesoría a distintas organizaciones privadas y públicas ligadas al mundo de la salud. En esta ocasión está realizando un trabajo de asesoría para un hospital perteneciente a uno de los servicios de salud del país. Para realizar una detección y vigilancia temprana de pacientes con tuberculosis (una infección bacteriana contagiosa que afecta principalmente a los pulmones), le piden que implemente un algoritmo de inteligencia artificial que ayude a determinar si un paciente tiene o no tuberculosis, utilizando distintos datos y mediciones del paciente. Para esto, Amanda entrenó una variedad de algoritmos con datos de la población de salud que atiende el hospital, tomando una proporción similar de pacientes con tuberculosis que pacientes sin tuberculosis.

Actualmente ella está en la etapa de probar y seleccionar el algoritmo a utilizar para la predicción de tuberculosis. Debe entregar los resultados en un mes al hospital y hay cuatro algoritmos susceptibles de ser utilizados para la labor encomendada: el algoritmo X, el algoritmo Y, el algoritmo W y el algoritmo Z. Al examinar sus desempeños, advierte que el algoritmo X es más exacto³ en promedio del total de la muestra, pero tiene una tasa de error⁴ más elevada en un subconjunto de la muestra correspondiente a la población migrante. Por otra parte, el algoritmo Y tiene tasas de error similares en la población de distintas nacionalidades y orígenes, pero tiende a fallar más en promedio. En cuanto a los algoritmos restantes, encuentra que el algoritmo W tiene mayor exactitud, pero menor explicabilidad⁵ mientras que el algoritmo Z es menos exacto pero los resultados que entrega son explicables.

En relación con la predictibilidad de un algoritmo, que las predicciones lleguen a ser erradas provoca dos consecuencias. La primera es que, si una persona no tiene tuberculosis, pero el algoritmo predice que la tiene, se realizaría un mayor gasto en tratamientos innecesarios. La segunda es que, si una persona tiene tuberculosis, pero el algoritmo predice que no

³ La exactitud se refiere a la proporción de datos bien predichos con respecto al total de datos.

⁴ La tasa de error se refiere a la proporción de datos *mal* predichos con respecto al total de datos. En este caso a detectar la enfermedad cuando no la tiene o viceversa.

⁵ La explicabilidad se refiere al grado en el que el comportamiento del modelo se puede explicar en términos humanos, considerando tanto el resultado como todo el proceso de la toma de decisión. (<https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2019/09/3T19-Quarterly-iDANAE.pdf>)

la tiene, entonces su enfermedad no sería detectada y, por lo tanto, no se trataría, lo que podría provocar su muerte. Si se utiliza un algoritmo menos exacto en la población entera o en una muestra, ambos problemas mencionados se presentan en mayor cantidad.

En cuanto a la explicabilidad de un algoritmo, si este no es explicable entonces ni el equipo médico ni el equipo de ingenieras/os podrían estudiar sus resultados para entender potenciales errores afectando así su confiabilidad. Además, se estaría infringiendo el derecho a una explicación del paciente, reconocido legalmente en otras legislaciones⁶ como el derecho que tiene un/a usuario/a a pedir una explicación de la salida de un algoritmo cuya decisión pueda tener un impacto importante en su vida, como ocurre en este caso por las consecuencias graves de un posible error.

En este dilema las/os estudiantes deben escoger una postura entre dos posibles, que son irreconciliables. Para ello, se dispone de una escala de valoración que va del 1 al 6, partiendo de la izquierda (1) y terminando en (6) a la derecha. Para el trabajo con dilemas se propone una actividad en tres etapas. En la primera el/la estudiante responde de manera individual al dilema presentado marcando una opción y justificando dicha preferencia. En la segunda etapa se conforman grupos con un criterio de heterogeneidad en las respuestas de la primera etapa, para generar una discusión e intercambio de ideas que aporten a la construcción de nuevas perspectivas en grupos pequeños, esto se puede hacer manualmente, y en grupos numerosos, se pueden usar diversas herramientas digitales como la aplicación web *EthicApp* (Álvarez, Zurita y Baloian, 2022). En la tercera etapa, el/la estudiante responde nuevamente en forma individual al dilema presentado luego de la discusión grupal. De esta manera se movilizan y desarrollan los elementos mencionados en los apartados 2 y 3 de este trabajo.

⁶ En la Unión Europea, aplica el *General Data Protection Regulation* (<https://www.privacy-regulation.eu/en/r71.htm>). También existe una regulación similar en Estados Unidos para puntuación crediticia automatizada, por el *Equal Credit Opportunity Act* (<https://www.ecfr.gov/current/title-12/chapter-X/part-1002/subpart-A/section-1002.9>).

La estructura mencionada se presenta de la siguiente manera:

1. Considerando la información recabada por Amanda mediante el entrenamiento y evaluación de los algoritmos y el impacto que tiene la tuberculosis en la población, ¿qué algoritmo debería escoger?

Algoritmo X: menor error detectando la enfermedad en promedio, pero con mayor error detectando la enfermedad en población migrante	o	o	o	o	o	o	Algoritmo Y: mayor error detectando la enfermedad en promedio, pero con tasas de error similares para cada nacionalidad
	1	2	3	4	5	6	

2. Teniendo en cuenta el propósito del proyecto (detectar tempranamente a pacientes con tuberculosis mediante la implementación de un algoritmo de inteligencia artificial) de ciencia de datos que dirige Amanda, ¿qué algoritmo debería escoger?

Algoritmo W: mayor exactitud y menor explicabilidad	o	o	o	o	o	o	Algoritmo Z: menor exactitud y mayor explicabilidad
	1	2	3	4	5	6	

Esta estructura puede, eventualmente, variar. Lo importante es crear una tensión en la que sea menester decidir una postura y poder fundamentarla a partir de las propias convicciones y de una discusión grupal, así como, desde la introducción de otra información, teorías y marcos éticos y normativos relevantes para la situación dada.

5. Orientaciones para evaluar el trabajo con dilemas

Con el fin de presentar orientaciones para evaluar el trabajo con dilemas, a continuación, se exponen dos elementos que permiten diseñar indicadores de evaluación que contribuyan a esta tarea: el discernimiento ético y el mapa de progreso de la competencia ética.

a. El discernimiento ético

Para actuar éticamente, se debe adquirir la capacidad de tomar decisiones de forma reflexiva y fundamentada, considerando un marco de valores y principios determinado. Esta capacidad es el discernimiento ético. Es un ejercicio que hacemos en vista a una decisión o postura que hay que tomar, cuando hay dos posturas distintas y una excluye a la otra. La base de esta metodología la encontramos en Mifsud (2012), de la cual hacemos una adaptación.

El proceso de discernimiento ético consiste en:

- (1) Identificar los hechos.** Se requiere identificar con claridad el problema, siendo capaz de describir la realidad sin confundirla con los propios juicios y de la forma más precisa posible, incluyendo todos los elementos necesarios para saber qué se va a discernir. Se debe definir con detalle el dilema y/o conflicto ético, delimitar la materia a discernir y explicitar quién es el sujeto que discierne. Se debe objetivar la experiencia y los sentimientos personales y colectivos. Es recomendable terminar esta etapa con una pregunta que oriente el discernimiento.
- (2) Comprender los hechos.** Es un momento de investigación para darle sentido a los hechos y comprender la realidad de manera integral, teniendo la capacidad de investigar e indagar, considerando las distintas maneras de explicar la realidad del hecho en su totalidad, sus implicaciones y consecuencias. Para esto es esencial la dimensión transdisciplinaria (aspectos sociales, económicos, políticos, éticos y/o ambientales), pues la realidad no es unidimensional. Es importante, por ejemplo, conocer elementos comparativos

históricos y/o legales y casos similares. En esta etapa se deben evitar, también, los juicios de valor e incorporar nociones teóricas que iluminen el problema. El hecho debe ser comprendido desde el punto de vista de todos los actores y perspectivas involucradas, objetivando influencias significativas del contexto personal y social del sujeto que discierne.

- (3) **Precisar implicancias éticas.** Identificar, mediante evaluación ética, cuáles son los valores (personales, sociales y/o institucionales), principios, virtudes y deberes (culturales y/o legales) subyacentes al problema respecto al cual hay que discernir. En este punto se incluyen teorías y/o racionalidades éticas (por ejemplo: ética de las virtudes, ética utilitarista, ética deontológica, ética procedimentalista, ética del cuidado, del reconocimiento, etc.). Aquí corresponde identificar los «nudos» que debemos desenredar, en este caso, de naturaleza ética. Conviene precisar qué conflictos y/o dilemas éticos hay en la situación descrita comprendida.
- (4) **Elementos para el discernimiento.** Se debe tomar una decisión o dar posibles soluciones, sugerir vías de acción o posiciones a tomar, tomando en consideración lo trabajado en los tres puntos anteriores. Para esto se deben jerarquizar los valores, definir los principios, virtudes y deberes en juego, así como considerar los marcos normativos pertinentes, todo ello basado en la(s) racionalidad(es) ética(s) en juego.

b. Dimensiones de la competencia

Según el Modelo Educativo de la Universidad de Chile (2021), las competencias:

son desarrolladas progresivamente durante el proceso formativo y su grado de logro es susceptible de ser evaluado. Implican, en definitiva, un conjunto de saberes que se desarrollan a lo largo de la formación y permiten a los egresados y egresadas tomar decisiones fundamentadas y actuar en coherencia, en distintos ámbitos de desempeño (p. 37).

Dentro de los perfiles de egreso se distinguen tres competencias: (1) las competencias específicas, (2) las competencias genéricas y (3) las competencias sello. La competencia de compromiso ético es una competencia sello del Modelo Educativo institucional. Su definición señala que corresponde a la:

capacidad de actuar responsablemente en la vida personal y profesional y de participar activamente en la vida social y ciudadana de acuerdo con los principios y valores democráticos fundamentales, el respeto irrestricto a los derechos humanos, al bien común y la justicia social. Implica tanto el desarrollo de la responsabilidad sobre sí, el propio actuar y sus consecuencias respecto de las demás personas, el entorno y la sociedad en su conjunto, así como el desarrollo de la capacidad de agencia personal y colectiva para la transformación social. (p. 40)

Con el fin de operativizar esta competencia se crea un mapa de progreso (ver sección Anexo) que contiene distintos elementos de progresión de la competencia de compromiso ético para su transversalización en la formación ética del plan de estudios de Ingeniería y Ciencias. En primer lugar, se desglosa la competencia en dos: (I) competencia plan común y (ii) competencia terminal. En segundo lugar, se definen tres niveles (i) básico, (ii) intermedio y (iii) avanzado, cada uno de ellos con hitos evaluativos asociados que se identifican en determinados momentos del plan de estudio. En tercer lugar, la competencia se divide en tres dimensiones: (i) responsabilidad, (ii) integridad y (iii) respeto, que corresponden a los principios en los que se quiere formar. En cuarto lugar, se definen indicadores de logro para cada dimensión y nivel. En el apéndice se puede encontrar el mapa de progreso construido para el desarrollo de la competencia en cuestión.

Referencias bibliográficas

- Aisenberg, B. (1994) Didáctica de las ciencias sociales: aportes y reflexiones / coord. por Silvia Alderoqui, Beatriz Aisenberg, 1994, ISBN 950-12-2110-5.
- Alvarez, C., Zurita, G. & Baloian, N., (2022) Applying the concept of implicit HCI to a groupware environment for teaching ethics. *Pers Ubiquit Comput* 26, 1373–1391. doi: 10.1007/s00779-020-01495-z
- Araya, V., Alfaro, M., & Andonegui, M. (2007). Constructivismo: orígenes y perspectivas. *Laurus*, 13(24), 76-92.
- Carretero, M.; Pozo, J.; Asensio, M. (1989). La enseñanza de las Ciencias Sociales. Madrid: Aprendizaje-Visor.
- Chadwick, Clifton B. (1999). La psicología del aprendizaje desde el enfoque constructivista. *Revista Latinoamericana de Psicología*, 3(31), 463-475.
- Cerda, J., Carvajal, C, Ramírez, P. y Hurtado, E. (octubre de 2023). «Profundizando en la formación ética canalizada mediante la tecnología: análisis exploratorio de juicios morales en la discusión de dilemas». Ponencia presentada en el XXXV Congreso Chileno de Educación en Ingeniería (SOCHEDI): la urgencia de una formación en ingeniería que contribuya a los ODS. Universidad de Valparaíso, Valparaíso, 11 a 13 de octubre.
- Conferencia de las Naciones Unidas sobre el Medio Ambiente y de Desarrollo (Río de Janeiro, B., & Naciones Unidas. (1992). *Declaración de Río sobre el Medio Ambiente y el Desarrollo: Principios relativos a los bosques*. Rio de Janeiro: Departamento de Información Pública de las Naciones Unidas.
- Cortina, A. (2000). «Presentación. El sentido de las profesiones». En A. Cortina y J. Conill, *10 palabras clave en ética de las profesiones*. Ed. Verbo Divino: Navarra.

- Cortina, A. (1996a). *El quehacer ético: guía para la educación moral*. Ed. Santillana: Madrid.
- Cortina, A. (1996b). «El estatuto de la ética aplicada: Hermenéutica crítica de las actividades humanas». En *Isegoría*/13, 119-134.
- García López, R.; Ferrández Berrueco, M.; Sales Ciges, M. y Moliner García, M. (2006). Elaboración de instrumentos de medida de las actitudes y opiniones del profesorado universitario hacia la ética profesional docente y su papel como transmisor de valores. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa*, vol. 12, núm. 1, 2006, pp. 129-149 Universitat de València.
- Giménez, E. (2018). Una propuesta de articulación de la ética de la ingeniería civil. A. Richart, M. García-Granero, C. Ortega Esquembre y L. de Tienda Palop (Eds.): *Juicio moral y democracia. Retos de la ética y la filosofía política*. Granada: Comares, 2018.
- Lind, G. (1985). The theory of moral-cognitive judgment: A socio-psychological assessment. En G. Lind, H.A. Hartmann y R. Wakenhut, *Moral development and the social environment. Studies the philosophy and psychology of moral judgment and education* (pp. 21-53). Chicago: Precedents Publishing.
- Maliandi, R. (1991). *Ética: conceptos y problemas*. Buenos Aires, Biblos.
- Meza Rueda, J. L. (2008). Los dilemas morales: una estrategia didáctica para la formación del sujeto moral en el ámbito universitario. *Actualidades Pedagógicas*, (52), 13-24.
- Mifsud, T. (2012). *Decisiones responsables: una ética de discernimiento*. Santiago: Ediciones Universidad Alberto Hurtado.
- Nieda, J. y Macedo, B. (1997). *Un currículo científico para estudiantes de 11 a 14 años*. Madrid: UNESCO – OEI.
- Ortiz Granja, D. (2015). El constructivismo como teoría y método de enseñanza. *Sophia, Colección de Filosofía de la Educación* (19), 93-110.

- Ormart, E. (2001). El aprendizaje de la ética en las instituciones de Educación Superior En Anuario del Instituto de investigaciones. Número IX p. 40.
- Ormart, E. (2010). *El aprendizaje de nociones éticas en estudiantes universitarios*. Universidad de Morón, ARK: <https://n2t.net/ark:/13683/p70c/GBs>
- Papalia, D., Wendkos, S., & Duskin, R. (2007). *Desarrollo humano*. México D.F.: McGraw-Hill Interamericana.
- Pulgar, J.L., (2005). *Evaluación del aprendizaje no formal. Recursos prácticos para el profesorado*. Madrid: Narcea.
- Ramírez, P., Guerrero, S., Cerda, J., Ross, J. y Flores, G. (2022). «La formación ética canalizada mediante la tecnología. Experiencia y resultados preliminares del uso de la herramienta web Ethicapp». *Cuadernos de Beauchef*, 6(2), 73-87.
- Richart, M. García-Granero, C. Ortega Esquembre y L. de Tienda Palop (Eds.). (2018). Juicio moral y democracia. Retos de la ética y la filosofía política. Granada: Comares.
- Ruiz-Cano, J.; Cantú-Quintanilla, G.; Ávila-Montiel, D.; Gamboa-Marrufo, J.D.; Juárez-Villegas, L.E.; de Hoyos-Bermea, A.... Grupo de Estudios sobre Dilemas Éticos HIMFG-IPADE (2015). *Revisión de modelos para el análisis de dilemas éticos. Boletín Médico del Hospital Infantil de México*, Volumen 72(2), 89-95.
- Tünnermann Bernheim, C. (2011). El constructivismo y el aprendizaje de los estudiantes. *Universidades* (48), 21-32.
- Universidad de Chile (2021). *Modelo Educativo de la Universidad de Chile*. Recuperado de: <https://repositorio.uchile.cl/handle/2250/151685>
- Weissmann, H. (1993). Qué enseñan los maestros cuando enseñan ciencias naturales, y qué dicen querer enseñar, en Weissmann, H. (Comp.). *Didáctica de las Ciencias Naturales. Aportes y Reflexiones*, pp. 37-65. Buenos Aires: Paidós

Anexo

Mapa de progreso

En este apartado se presenta la progresión de la competencia «compromiso ético». En la siguiente tabla, se desglosa la competencia ética en lo que respecta a plan común y a la etapa terminal de la formación.

Competencia Plan Común	Reflexionar sobre el propio actuar y sus consecuencias, en el marco de la integridad, la responsabilidad y el respeto, buscando la excelencia y rigurosidad en su proceder en contextos académicos, en las relaciones interpersonales y con su entorno.
Competencia Terminal	Actuar de manera responsable y honesta, dando cuenta en forma crítica de sus propias acciones y sus consecuencias, en el marco del respeto hacia la dignidad de las personas y el cuidado del medio social, cultural y natural.

Fase 2: Descripción de cada nivel, según integración de criterios			
Competencia: ¿cómo progresa cada dimensión de la competencia?	Nivel 1	Nivel 2	Nivel 3
	Primer a cuarto semestre	Quinto a séptimo semestre	Octavo al noveno semestre
1. Responsabilidad: cuidado solícito ⁷ y precautorio ⁸ de las tareas y relaciones que establecemos en el ámbito académico y profesional dando cuenta de nuestras acciones y decisiones, asumiendo sus consecuencias.	1.1. Identifica y aplica conceptualmente los aspectos de la responsabilidad en un contexto académico y de la vida estudiantil explicitando impactos y/o consecuencias. 1.2. Contrasta los efectos de sus acciones, en un contexto académico y estudiantil en el marco del principio ético de la responsabilidad.	1.3. Analiza los efectos de su acción en un marco de precaución en un contexto académico y profesional identificando ámbitos de aplicación de la responsabilidad.	1.4. Discierne éticamente en situaciones de la vida académica y profesional, en un marco de responsabilidad respecto a sus propias acciones y decisiones, proyectando sus consecuencias.

⁷ Según la RAE: diligente, atento.

⁸ Del Principio Precautorio consagrado en la Declaración de Río sobre el Medio Ambiente y Desarrollo: «Con el fin de proteger el medio ambiente, los Estados deberán aplicar ampliamente el criterio de precaución conforme a sus capacidades. Cuando haya peligro de daño grave o irreversible, la falta de certeza científica absoluta no deberá utilizarse como razón para postergar la adopción de medidas eficaces en función de los costos para impedir la degradación del medio ambiente» (Declaración de Río 1992).

Propuesta de modelo para la formación ética mediante la...

Fase 2: Descripción de cada nivel, según integración de criterios			
Competencia: <i>¿cómo progresa cada dimensión de la competencia?</i>	Nivel 1	Nivel 2	Nivel 3
	Primer a cuarto semestre	Quinto a séptimo semestre	Octavo al noveno semestre
2. Integridad: actuar en el ámbito académico y profesional acorde a los valores de la honestidad, la confianza, la equidad, el respeto, la responsabilidad y el coraje.	2.1. Identifica los distintos aspectos de la integridad involucrados en la vida estudiantil con proyección a la vida profesional. 2.2. Aplica los marcos normativos relevantes que regulan la vida universitaria en la FCFM [código de ética, misión y visión de la universidad] 2.3. Planifica y presenta sus trabajos, basándose en sus capacidades, sin incurrir en plagio, copia, suplantación de identidad en contextos estudiantiles y profesionales. 2.4. Se relaciona de manera honesta cuando interactúa en un ámbito académico, siendo consecuente en su desempeño, en las diversas situaciones formativas y evaluativas.	2.5. Evalúa autónomamente el propio actuar a la luz de los valores de la integridad académica, considerando la retroalimentación obtenida en el desempeño de su vida académica. 2.6. Se relaciona de manera honesta cuando interactúa en un ámbito académico y profesional, siendo consecuente en su desempeño, en las diversas situaciones formativas y evaluativas.	2.8. Discierne autónomamente y acorde a los valores de la integridad sobre vías de acción en situaciones dilemáticas y/o éticamente conflictivas propias de la vida profesional y social. 2.9. Aplica marcos normativos relevantes que regulan la vida profesional y social [código de ética del Colegio de Ingenieros, ISO, regulaciones por especialidad, entre otros]. 2.10. Utiliza marcos normativos, estableciéndolos como principios rectores de la propia actividad profesional y social.

Fase 2: Descripción de cada nivel, según integración de criterios			
Competencia: <i>¿cómo progresa cada dimensión de la competencia?</i>	Nivel 1	Nivel 2	Nivel 3
	Primer a cuarto semestre	Quinto a séptimo semestre	Octavo al noveno semestre
3. Respeto: disposición a relacionarse con otros/as y valorar sus opiniones, aunque difieran de las propias. No solo incluye la tolerancia a las ideas de los demás, sino que implica consideración positiva hacia el otro y un interés por comprenderlo con el fin de contribuir a la sana convivencia.	3.1. Identifica las distintas dimensiones e implicancias del respeto en situaciones de la vida académica y estudiantil. 3.2. Expresa, en forma oral y escrita, la aceptación de ideas y realidades distintas, sin discriminar a sus pares, funcionarios y cuerpo docente.	3.3. Considera la opinión de los demás, también cuando ella es divergente, aportando a la construcción de un clima de tolerancia en la convivencia. 3.4. Promueve autónomamente el respeto activo en las diversas actividades que desempeña en su vida estudiantil con proyección a la vida profesional.	3.5. Discierne éticamente en situaciones de la vida académica y profesional, considerando el marco del respeto activo en las diversas actividades que desempeña.

Recomendaciones para una IA responsable

Ricardo Baeza-Yates¹

Introducción

Este ensayo resume la mayoría de las propiedades de un sistema de software, en particular, los basados en inteligencia artificial (IA), y entrega recomendaciones para el análisis de algunas de ellas. No todas estas propiedades son necesarias para tener una IA responsable y algunas son más generales, pero todas ellas pueden generar algún impacto legal o ético.

Es importante señalar también que se habla de IA confiable o IA ética, pero ambos términos tienen problemas. El primero es tendencioso, pues sabemos que los sistemas de IA no son confiables todo el tiempo y, además, pone el peso del problema en el usuario y no en los creadores. Asimismo, la IA no puede ser ética pues la ética es humana y, por lo tanto, no es una característica de una tecnología. Dado que el tema de sistemas de IA responsable es un área en pleno desarrollo, todo lo presentado puede cambiar significativamente en el futuro, cuando ya no sea una tendencia, sino una necesidad. En particular, cuando hablamos de inteligencia artificial nos referimos a técnicas de aprendizaje automático. Pero en el futuro existirán sistemas híbridos que mezclan estas técnicas con bases de conocimiento y otras de inferencia más avanzadas, al igual que de comprensión semántica y razonamiento lógico.

¹ Instituto de IA Experiencial, Northeastern University Silicon Valley, EE. UU. Departamento de Ciencias de la Computación, Universidad de Chile, Santiago, Chile.

Propiedades de Sistemas de IA

A partir de un estudio bibliográfico y conocimiento previo, se recolectaron 30 propiedades relacionadas con datos y software, independientes de si ellas eran necesidades éticas o no, ya que en muchos casos esto depende de la aplicación específica. Estas propiedades se analizaron desde dos puntos de vista:

- **Aplicación:** referidas a los datos, los modelos/algoritmos principales de la aplicación, al sistema completo de IA o a la gobernanza de este mismo, sin incluir la gobernanza de los datos. La Tabla 1 muestra nuestro análisis inicial en estos cuatro aspectos.
- **Impacto:** son importantes para la justicia, para el gobierno, para los usuarios del sistema o para la sociedad en general. La Tabla 2 muestra nuestro análisis inicial de estos cuatro actores.

En ambas tablas hemos traducido del original en inglés, aunque algunas de ellas no tienen traducción directa como *accountability*, que es rendición de cuentas (esto es un ejemplo de sesgo semántico codificado en el lenguaje). Con respecto a este análisis debemos señalar:

- Respecto a la aplicación, hay que advertir que, en el caso de un sistema basado en aprendizaje automático supervisado, lo importante es el modelo que se usa. Sin embargo, un modelo por sí solo no resuelve nada y necesita una aplicación que lo procesa, interpreta y ejecuta. Por lo tanto, el modelo en producción también es un algoritmo.
- Respecto al impacto, aunque hay propiedades que tienen implicancias legales y éticas, la relevancia de ellas dependerá del uso particular que se dé al sistema de IA y del contexto de este. Por ejemplo, las implicancias legales son muy distintas para un país específico o un uso a nivel mundial, en el cual muchas más propiedades serán necesarias.

Propiedad	Datos	Modelo/ Algoritmo	Sistema	Gobernanza
Procedencia de los datos	✓			✓
Privacidad	✓		✓	✓
Control de calidad	✓		✓	✓
Trazabilidad	✓		✓	✓
Acceso y corrección	✓		✓	✓
Mantenimiento	✓	✓	✓	✓
Equidad y sesgos	✓	✓	✓	✓
Cumplimiento legal	✓	✓	✓	✓
Compleitud		✓	✓	✓
Conciencia		✓	✓	✓
Eficiencia		✓	✓	
Validación y testeo		✓	✓	
Interpretabilidad		✓	✓	
Explicabilidad		✓	✓	
Accesibilidad			✓	
Rendición de cuentas			✓	✓
Responsabilidad			✓	✓
Integridad y confianza			✓	✓
Seguridad			✓	✓
Proporcionalidad			✓	✓
Interoperabilidad			✓	✓
Autonomía			✓	✓
Transparencia			✓	✓
Documentación			✓	✓
Beneficiosa			✓	✓
Resiliencia			✓	✓

Usabilidad	✓	✓
Sostenibilidad	✓	✓
Auditabilidad	✓	✓
Reproducibilidad	✓	

Tabla 1: Propiedades indicando su dependencia de distintas partes del sistema de IA.

Las propiedades en negrita son las 7 inicialmente definidas por la *Association for Computing Machinery* (ACM) para lograr la *transparencia* de los sistemas de IA [1], aunque transparencia no es uno de los principios definidos y solo aparece en el título. La ACM es la asociación de profesionales de computación más grande del mundo, con más de 100 mil miembros, y por ende, es un referente de primer nivel.

Propiedad	Justicia	Gobierno	Usuarios	Sociedad
Procedencia de los datos	✓	✓	✓	✓
Privacidad	✓	✓	✓	✓
Control de calidad			✓	✓
Trazabilidad				
Acceso y corrección				
Mantenimiento				
Equidad y sesgos	✓	✓	✓	✓
Cumplimiento legal	✓	✓	✓	✓
Compleitud			✓	✓
Conciencia			✓	✓
Eficiencia			✓	✓
Validación y pruebas				
Interpretabilidad				
Explicabilidad	✓	✓	✓	✓
Accesibilidad	✓	✓	✓	✓

Rendición de cuentas	✓	✓	✓	✓
Responsabilidad	✓	✓	✓	✓
Integridad y confianza	✓	✓	✓	✓
Seguridad	✓	✓	✓	✓
Proporcionalidad	✓		✓	✓
Interoperabilidad			✓	
Autonomía			✓	
Transparencia			✓	✓
Documentación			✓	✓
Beneficiosa			✓	✓
Resiliencia			✓	✓
Usabilidad			✓	✓
Sostenibilidad	✓	✓		✓
Auditabilidad	✓	✓		
Reproducibilidad	?	?		

Tabla 2: Propiedades dependiendo de a quién impactan o a quién les importa.

En estas tablas podemos ver que hay propiedades de más alto nivel que otras. También vemos que algunas propiedades son transversales en ambas tablas (es decir, la fila correspondiente está completa) y las hemos ordenado verticalmente de modo de que se vean los grupos de propiedades que tienen aspectos y/o actores similares. Sobre esta base, proponemos una estructura jerárquica de estas propiedades en la Tabla 3, en la que hemos destacado en negrita las propiedades relacionadas con la ética (si es una propiedad principal, todas las propiedades secundarias también lo son). Igualmente, hemos incluido aquí el término original en inglés.

Propiedad Principal	Propiedades Secundarias	Notas
Conciencia (<i>Awareness</i>)	Validez ética y legal Validez científica Autonomía (<i>autonomy</i>) Integridad (<i>integrity</i>)	Legitimidad e identidad del sistema
Proveniencia de datos (<i>Data provenance</i>)	Control de calidad (<i>quality assurance</i>) Equidad y sesgo (<i>equity & bias</i>) Trazabilidad (<i>traceability</i>) Acceso y corrección (<i>access & redress</i>)	Representan el ciclo de vida de los datos
Robustez (<i>Robustness</i>)	Control de calidad (<i>quality assurance</i>) Adaptabilidad (<i>adaptability</i>) Escalabilidad (<i>scalability</i>) Extensibilidad (<i>extensibility</i>) Interoperabilidad (<i>interoperability</i>)	Representan la completitud del sistema
Usabilidad (<i>Usability</i>)	Eficiencia (<i>efficiency</i>) Accesibilidad (<i>accessibility</i>) Resiliencia (<i>resilience</i>) Reproducibilidad (<i>reproducibility</i>)	Permiten la satisfacción del usuario
Transparencia (<i>Transparency</i>)	Validación y testeo (<i>validation & testing</i>) Documentación (<i>documentation</i>) Interpretabilidad (<i>interpretability</i>) Explicabilidad (<i>explainability</i>) Auditabilidad (<i>auditability</i>)	Permiten la transparencia del sistema
Responsabilidad (<i>Responsibility</i>)	Conformidad legal (<i>Legal compliance</i>) Rendición de cuentas (<i>Accountability</i>) Proporcionalidad (<i>Proportionality</i>) Privacidad (<i>Privacy</i>) Seguridad (<i>Security & safety</i>) Integridad y confianza (<i>Trustworthy</i>) Mantenibilidad (<i>Maintenance</i>) Sostenibilidad (<i>Sustainability</i>) Beneficiosa (<i>Beneficial/wellbeing</i>)	Permiten que el sistema cumpla con principios éticos y normas legales

Tabla 3: Agrupación jerárquica de propiedades.

Puede observarse que el control de calidad se ha separado en dos para aplicarlo tanto a los datos como al sistema. Otras propiedades están ya incluidas dentro de las anteriores. Por ejemplo,

la protección de datos es el resultado de la privacidad y la seguridad. Del mismo modo, la eficacia del sistema o el uso racional de recursos es parte de la sostenibilidad. Otras propiedades relacionadas con los datos están también en responsabilidad. Por ejemplo, la recolección mínima y el almacenamiento de datos por un tiempo acotado es parte de la proporcionalidad. Quiero recordar que interpretabilidad se refiere a entender cómo el sistema llega a una decisión, mientras que explicabilidad significa que el sistema debe poder explicar una respuesta específica a un usuario dado. Esta agrupación en el original en inglés [4], fue uno de los insumos usados para generar los nuevos principios para sistemas algorítmicos responsables de la ACM publicados en octubre de 2022 [5], y en particular, el primer principio de legitimidad y competencia.

Gobernanza

La gobernanza de un sistema de IA incluye los datos y sus metadatos, los modelos y todos los procesos para entrenarlos, validarlos y evaluarlos, el software en producción con todos los mecanismos de registro y control subyacentes, y toda la información necesaria para mantener y evaluar las propiedades antes descritas.

Hay pocas propuestas concretas para la gobernanza. Una de ellas es la de Ben Shneiderman [12], que la divide en tres partes. La ingeniería de software como tal, que es a la que se refiere el párrafo anterior, que concierne al equipo de desarrollo. Por encima de ella está el diseño organizacional, que es la gobernanza de la institución misma. Finalmente, agrega un tercer nivel, que es el de certificación externa, el cual incluye regulación gubernamental y auditorías de software.

Sin embargo, en la práctica, es mejor ver la gobernanza en forma temporal respecto a los principios instrumentales que se definen. Si usamos los 9 nuevos principios de la ACM, tenemos el diagrama de la Figura 1 [6]. En negrita se destacan las herramientas principales que ayudan a la gobernanza. Si la gobernanza es adecuada, evitaremos la frecuencia de la última herramienta: auditoría algorítmica.

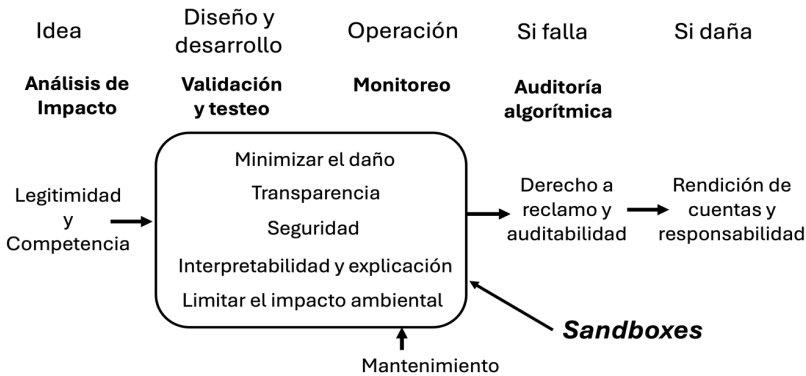


Figura 1: Gobernanza temporal usando los principios de la ACM.

Control de Calidad

Un modelo basado en aprendizaje automático supervisado debe ser robusto. Esto incluye una serie de análisis que no siempre se realizan, el que se resume a continuación. Para ello supondremos que en ciertos casos hay una noción de complejidad del problema o al menos una medida sustituta (*surrogate, proxy*) de ella; si no, habría que usar la valoración de expertos del área de aplicación.

Calidad de los Datos

Cobertura de los datos de entrenamiento: ¿tenemos suficientes datos en todo el espacio de posibles conjuntos de entrada? Debemos recordar que la complejidad de cada instancia es distinta, nunca es uniforme. Por lo tanto, el número de datos deberá ser al menos proporcional a esta complejidad. Si no es posible conocer la complejidad, una forma de analizar esto es escoger las variables más importantes de la entrada y dividir su rango de valores en dos o más partes de acuerdo con el problema. Por ejemplo, si uno usa la edad, podría dividir en niños y adolescentes, adultos hasta los 65 años y mayores de 65 años. ¿Tenemos una cantidad de datos en cada rango

proporcional a la población en ese rango de acuerdo con el censo del país en que se usa la aplicación?

Análisis de sesgos de datos: ¿analizó posibles sesgos en los datos? Si tiene datos demográficos, estos pueden incluir sesgo de género, edad, etnia, nivel económico, nivel educativo, etc. En algunos casos los sesgos pueden ser justificados dado el problema a resolver, pero habrá que verificar que sea el esperado. Por ejemplo, si se espera que haya más representación masculina, no debiera estar sobrerrepresentado. Es decir, hay que decidir cuál es la distribución neutral para el problema dado y eso no siempre es simple. Es fundamental tener en cuenta que en muchos casos se trata de una decisión social, no una decisión de los desarrolladores, así que debe ser validada con expertos del área de aplicación. Por otro lado, hay sesgos que pueden ser desconocidos. Si usamos el mismo ejemplo de la edad anterior, podemos hacer muestras al azar del mismo tamaño en esos tres rangos. ¿Son las distribuciones similares? Si no lo son, ¿hay una justificación para que no lo sean o hay un sesgo que no esperábamos?

Calidad del Modelo

Análisis de sesgos algorítmicos: el modelo mismo puede generar sesgos no previstos. Un buen ejemplo es el caso de Deliveroo [10], una aplicación para repartir comida, la cual, al intentar maximizar la ganancia económica, asignaba menos trabajo a los repartidores que no podían o no querían trabajar a las horas de más pedidos (la cena). En este caso el modelo no consideró que debía distribuir equitativamente los pedidos, atendiendo a las limitaciones personales válidas de los repartidores (atención de niños o personas dependientes, horas no hábiles, etc.).

Otra fuente de sesgos algorítmicos es el ciclo de realimentación sistema-usuario debido a su uso. Por ejemplo, los sistemas de recomendación solo exponen al usuario a un número limitado de alternativas, generando un sesgo de popularidad en el cual los ítems más frecuentes tienen ventajas. También la posición en la pantalla de cada ítem genera un sesgo y el orden de los resultados genera

otro sesgo (sesgo de *ranking*). A esto hay que agregarle el problema de la burbuja (*filter bubble*) o de la cámara de eco, referido al conocimiento parcial de las preferencias de los usuarios (que afecta a la personalización de la experiencia). Todo esto se agrava con los sesgos cognitivos de los usuarios y sesgos de segundo orden, producto del uso de los resultados de un sistema para alimentar otro sistema (por ejemplo, usar el resultado de los buscadores para generar contenido nuevo en la Web).

Punto de operación: el modelo debe ajustarse al objetivo del problema y esto no necesariamente significa usarlo en el punto de mayor exactitud (*accuracy*). De hecho, el mejor modelo es el que maximiza la exactitud en el punto de operación, el que no es necesariamente el mejor posible en todos los puntos de operación posibles (curva precisión/exactitud o similar). Por ejemplo, en aplicaciones médicas es mejor tener más falsos positivos (es decir, personas que no están enfermas que luego de una visita al médico son descartadas) que falsos negativos (personas que están enfermas pero que nunca se enterarán). Un ejemplo de este tipo de análisis se presenta en predicción de dislexia [11].

Calidad de los Resultados

Análisis de sensibilidad: un modelo debe ser robusto. Esto significa que pequeños cambios en la entrada causan pequeños cambios en la salida. Viceversa, grandes cambios en las variables principales de la entrada deberían producir grandes cambios en la salida (para esto debemos realizar antes un análisis del impacto de las variables o características de la entrada, *feature analysis*). En general, modelos más sensibles tendrán mayores niveles de error.

Análisis de error: las medidas promedio como la exactitud, no muestran cómo el modelo se comporta para distintas instancias del problema. Es muy probable que el error sea insignificante cuando la instancia es fácil y mucho mayor cuando la instancia sea compleja. Este segundo caso es importante investigarlo, pues aquí están los errores con mayor impacto en la operación del sistema y que

son los que al final pueden afectar a personas específicas, grupos minoritarios, o incluso, a una gran parte de la sociedad (por ejemplo, mujeres).

Interpretabilidad y explicabilidad

No hay una definición única para estos dos términos, pero hay consenso en que un modelo de aprendizaje automático es *interpretable* si un ser humano puede comprender cómo el modelo toma una decisión. Esto implica conocer el proceso de cómo el modelo llega a un resultado. Por otro lado, un modelo es *explicable* si un ser humano puede comprender por qué se tomó una decisión específica. Esto implica conocer qué atributos o variables influyen en el resultado y en qué medida. Más aún, esta explicación debería poder darse en lenguaje natural.

En el caso de modelos opacos, como aprendizaje profundo, donde los atributos son generados por el sistema, es más difícil dar una explicación. En otros casos más transparentes, es necesario realizar el mismo análisis de características mencionado para la sensibilidad (*feature analysis*), el que depende del método de aprendizaje automático utilizado.

La facilidad para dar una explicación depende, como es esperable, de la complejidad del problema. Más complejo, más difícil será explicarlo. Es probable, pero no es siempre cierto, que, si el modelo es más complejo, probablemente la exactitud de la respuesta será menor. Por esta razón, las técnicas para generar explicaciones dependen del método de aprendizaje automático usado. En el caso de que sean genéricas, estas pueden ser locales (explican un resultado) o globales (intentan explicar el modelo completo). En [9] se presenta una taxonomía de modelos interpretables y una revisión extensa de todos los trabajos de explicabilidad a la fecha.

Conclusiones

Para una exposición más completa de todos los sesgos posibles en un sistema, recomendamos los sesgos de la Web [2]. Gran parte del problema de sesgos viene de categorías sociales que no tienen justificación científica, tal como se muestra en el documental Coded Bias [3]. Los ejemplos más clásicos son raza [8] o preferencia sexual [7], respecto de los cuales, la biología nos enseña que estamos simplificando un espacio multidimensional complejo en unas pocas categorías arbitrarias. Por esta misma razón, cuando usamos interseccionalidad (intersección de dos o más categorías), intentamos recuperar los casos más complejos, pero al mismo tiempo, en cierto sentido, validamos las categorías. La solución, por supuesto, es eliminar este tipo de categorizaciones, pero con esto tenemos la paradoja de que, si no las conocemos, no podemos medir los daños que han provocado sus sesgos en los datos.

En las consideraciones éticas, suponiendo que ya están cubiertos todos los requerimientos legales, la primera pregunta que debemos hacernos es: ¿usarías tu producto si pertenecieras al grupo objetivo de tu aplicación? Si la respuesta es no, ya tenemos un problema ético. Si la respuesta es sí, es de esperar que no sea porque sabemos que en nuestro caso nos favorece, ya que si no, entonces, habría otro problema ético. El resto de las preguntas dependerá de los resultados de los análisis mencionados en este ensayo.

Referencias bibliográficas

- Association for Computing Machinery. Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf. 1/12/2017.
- Baeza-Yates, Ricardo. Bias on the Web. *Communications of ACM*, vol. 61 (6), pp. 54-61, 6/2018. Presentación disponible en YouTube.
- Baeza-Yates, Ricardo, Muñoz, Catherine. Sesgos codificados. *Ciper Académico*, 8/5/2021.
- Baeza-Yates, Ricardo. Some thoughts on Responsible AI, sin publicar, 13/4/2022.
- Baeza-Yates, Ricardo, Matthews, Jeanna, et al. Principles for Responsible Algorithmic Systems, *ACM*, 24/10/2022. En castellano en: <https://www.acm.org/binaries/content/assets/public-policy/spanish-statement-ai.pdf>.
- Baeza-Yates, Ricardo. Introduction to Responsible AI. *European Review* 31 (4), 3/8/2023.
- Helm, Rebecca. Let's talk about biological sex. Twitter, 19/12/2019.
- Kolbert, Elizabeth. There is no scientific base for race – it's a made-up label. *National Geographic*, 4/2018.
- Linardatos, Pantelis, Papastefanopoulos, Vasilis, Kotsiantis, Sotiris. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 18, 2021.
- Lomas, Natasha. Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm. *TechCrunch*, 4/1/2021.
- Rello, Luz, Baeza-Yates, Ricardo, Ali, Abdullah, Bigham, Jeffrey P, Serra, Miquel. Predicting risk of dyslexia with an online gamified test. *PlosOne*, 2/12/2020.
- Shneiderman, Ben. Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4, 10/2020.

Lengua, computadoras y emociones: interdisciplinariedad en la era de los Large Language Models (LLMs)¹

Amanda Cercas²

Introducción

El filósofo escocés David Hume dijo que la razón es, y debe ser, solo esclava de las pasiones (Hume, 1739). Hume se dirige a las tendencias filosóficas del momento, y que aún existen hoy en día, por las que sobrevaloramos la racionalidad por encima de nuestras emociones. Sin embargo, las emociones desempeñan un papel realmente importante en nuestras vidas: desvelan nuestros valores y guían nuestras acciones. Por ejemplo, si en una apuesta tienes un 80% de probabilidad de hacerte millonario, pero un 20% de perderlo todo, puede ser razonable apostar; sin embargo, ¿qué ocurre con el miedo a perderlo todo?

Esta tendencia a sobrevalorar la racionalidad permeó la Inteligencia Artificial (IA) hasta hace poco. En los años noventa, Rosalind Picard, en su libro *Affective Computing* (Picard, 2000), reconoce cómo las emociones nos ayudan a entender el mundo y propone la *computación afectiva*, una rama de la IA dedicada a entender las emociones humanas. Según Picard, las emociones forman una parte integral de las funciones cognitivas humanas y,

¹ Esta investigación es parte del estudio titulado “Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution”, llevado a cabo por el grupo MilaNLP de la Universidad Bocconi en Italia, la Universidad de Heriot-Watt en Edimburgo y la Universidad de Leeds en el Reino Unido. Ha sido aceptado para su publicación en el congreso más prestigioso en el campo de la Computación Lingüística y el Procesamiento del Lenguaje Natural: Association for Computational Linguistics.

² Investigadora Postdoctoral en Universidad Bocconi, amanda.cercas@unibocconi.it.

por lo tanto, deben formar parte integral de la IA. Además, Picard propone que las emociones merecen mayor consideración en el campo de la interacción humano-computadora, con sistemas que puedan reconocer, interpretar e, incluso, suscitar emociones en los usuarios. Desde la publicación del libro de Picard, el ámbito de la computación afectiva ha explotado, con cientos de artículos publicados cada año, además de la aparición de regulaciones contra el reconocimiento de las emociones en ámbitos educativos y laborales en la Unión Europea.

Las emociones

Las emociones son respuestas fisiológicas, cognitivas y conductuales a eventos. Según Paul Ekman, pionero del estudio de las emociones, son un tipo particular de valoración automática influida por nuestro pasado evolutivo y personal, según el cual percibimos que ocurre algo importante para nuestro bienestar (Ekman, 1972). Las emociones tienen funciones importantes a nivel social, facilitando las interacciones con otros/as, y nos ayudan a responder adecuadamente a situaciones nuevas, promoviendo nuestra supervivencia.

Las emociones tienen funciones tanto epistémicas como conativas (Curry, 2022). A nivel epistemológico, desempeñan al menos tres papeles: (1) señalan al individuo mismo lo que valoran y cómo ven el mundo, (2) señalan a nuestros compañeros lo que valoramos, y (3) las interacciones emocionales señalan a una tercera parte cómo los miembros de una interacción ven el mundo y lo que valoran. Por ejemplo, cuando una persona se lamenta, indica a sí mismo y a los demás que ha perdido algo de valor (dado el refrán «no sabes lo que tienes hasta que lo pierdes», no es inconcebible que no fuera consciente de cuánto lo valoraba). Finalmente, una tercera parte puede aprender sobre la dinámica del lamento, tus valores y los de tu familia observando vuestra interacción.

A nivel conativo, por otra parte, las emociones son la raíz de nuestra motivación, guiando nuestra forma de actuar. Ciertas

emociones funcionan como una «llamada» a actuar; la ira, por ejemplo, puede empujarnos a corregir algo que consideramos injusto. Asimismo, las emociones nos invitan a formar conexiones sociales, nos ayudan a tomar decisiones y mucho más.

La IA y las emociones

Teniendo en cuenta el rol tan significativo que tienen las emociones en nuestras vidas y en nuestra inteligencia, empezamos a entender la tesis de Picard: la Inteligencia Artificial no puede ser realmente inteligente sin entender nuestras *pasiones*. Sin embargo, aún estamos lejos de entender completamente las emociones en humanos, entonces, ¿cómo podemos enseñárselas a la IA y, más concretamente, a los modelos de lenguaje?

Hasta hace poco, la IA dependía de conjuntos de datos anotados específicamente para una tarea. Para cada ejemplo en nuestro conjunto de datos, anotadores humanos deciden a qué etiqueta o clase pertenece. Estos conjuntos de datos se ven limitados a un cierto número de clases a causa de los recursos necesarios para recolectarlos: las fuentes de datos aptas para cierto propósito son escasas, y cuando las encontramos, requieren tiempo y dinero para anotarlas. En el caso de las clases referidas a emociones, estas se concentran únicamente en seis emociones básicas, propuestas por Paul Ekman: ira, tristeza, alegría, asco, miedo y sorpresa (Plaza-del-Arco et al., 2024). Ekman plantea que estas seis emociones son universales y todos los humanos de todas las culturas las experimentan y expresan de la misma forma.

El marco teórico ofrecido por Ekman es ventajoso en la medida que presenta solo seis clases que, además, tienen valor *universal*. Sin embargo, podemos advertir varias deficiencias en este: las emociones básicas de Ekman están basadas en expresiones faciales, y nos dicen poco sobre las expresiones de emoción en el lenguaje; asimismo, estudios psicológicos más recientes, como el trabajo de Lisa Feldman-Barret, han puesto en cuestión la universalidad de las emociones, que es característica de propuesta

de Ekman (Barrett, 2017). Finalmente, varios filósofos han criticado la teoría de Ekman por la falta de dirección: cuando uno siente ira, siente ira hacia *algo*, pero el marco de Ekman ignora por completo el objeto de la emoción (Brady, 2019). Por otra parte, a nivel de la IA, podemos poner en cuestión la utilidad de seis emociones básicas en aplicaciones específicas –¿son siempre relevantes y suficientes?– En ciertos contextos, es probable que se manifiesten otras emociones en nosotros, tales como la melancolía, la soledad o, incluso, el amor, por ejemplo, si hacemos un poema.

A raíz de críticas como las señaladas, y de los avances en el modelamiento del lenguaje, vemos surgir nuevas técnicas, haciendo uso de los Large Language Models (LLMs) o modelos de lenguaje a gran escala, entre los cuales el más conocido es ChatGPT. Explicado de manera sencilla, estos modelos predicen la siguiente palabra en un texto. Para que esta predicción ocurra, cada modelo incorpora grandes conjuntos de datos «genéricos» –según se informa, unos 10 trillones de palabras en el caso de GPT-4 o cientos de millones de libros (Arya, 2023)–, y han demostrado mejoras en casi todas las tareas de Procesamiento de Lenguaje Natural (PLN). En el reconocimiento de emociones, no solo suponen una mejora frente a modelos anteriores a nivel de reconocimiento, sino que son capaces de atribuir una gran variedad de emociones a un texto sin necesidad de un conjunto de datos preanotados, por lo que abren la puerta a muchos más usos (Plaza del Arco et al., 2024). Sin embargo, los LLMs no son una solución milagrosa a todos los problemas que han plagado el reconocimiento de emociones (ni a otras tareas de PLN): como hemos visto en otros modelos del lenguaje natural, y otras tareas, tienden a reflejar los mismos estereotipos que los humanos (e.g., Kotek et al., 2023 y Shrawgi et al., 2024).

Los estereotipos de género y las emociones

Los estereotipos de género relacionados con las emociones se remontan al menos a los tiempos de Aristóteles, quien teorizaba que las mujeres eran más propensas a los excesos emocionales (Stauffer, 2008). Más adelante, Darwin (1874) asociaba la agresividad con la

masculinidad, y emociones como la empatía y aquellas relacionadas con el cuidado, con las mujeres. Hoy en día estos estereotipos aún nos afectan socialmente: la disparidad de género en campos como la ingeniería y la enfermería, refleja las capacidades «naturales» de cada género para ser racional o empático, respectivamente. Dado que estos estereotipos han acompañado a los humanos durante toda su historia, resulta poco sorprendente verlos aparecer en los modelos de lenguaje.

En un estudio reciente (Plaza-del-Arco et al., 2024), investigamos, precisamente, esos estereotipos de género. La metodología es sencilla: dada una situación, podemos pedirle al modelo que tome el rol de una mujer o de un hombre, y preguntarle cómo se sentiría en dicha situación. Si hacemos a ChatGPT que simule ser una mujer y le preguntamos qué emoción la mujer sentiría tras una discusión seria con un ser querido, por ejemplo, en la mayoría de las ocasiones responde «tristeza». Sin embargo, cuando se le pide simular ser un hombre, la respuesta es «ira». Repetimos este experimento con miles de eventos y descubrimos un patrón claro: los modelos asocian a las mujeres con emociones pasivas como la tristeza y la alegría, y a los hombres con emociones activas asociadas con la agencia y la autoestima, tales como la ira y el orgullo. Estas diferencias demuestran cómo los estereotipos que existen en nuestra sociedad se reflejan y se amplían en los modelos de lenguaje.

Cada día hay más avances en el PLN y más sesgos documentados. Sin duda, poco a poco la investigación encontrará soluciones a estos problemas, pero no siempre es fácil saber cuál es la solución ideal. En el caso de las emociones, son algo extremadamente subjetivo. La emoción que sentimos en cierto momento depende no solo de nuestra genética, sino de nuestra experiencia, nuestros valores y crianza. Teniendo esto en cuenta, es posible que (hasta cierto punto) las mujeres y los hombres, *en general*, sí sientan emociones diferentes frente a un evento similar ya que habrán recibido educaciones diferentes. Por ejemplo, estudios en psicología demuestran que las mujeres habitualmente tienden a ser más empáticas que los hombres (Jolliffe, 2006). Pero si esas

diferencias surgen a partir de una educación sexista, ¿los modelos deberían reflejarlas? Además, no debemos pasar por alto la falacia ecológica: las diferencias entre dos grupos no nos dicen mucho sobre los individuos. ¿Cuál es la mejor forma de abordar algo tan subjetivo e intentar generalizar? Estas son preguntas que los ingenieros o la informática no pueden resolver solos, sino que requieren una colaboración interdisciplinaria.

La IA y la empatía

El que los LLMs sean capaces de generar respuestas, una característica esencialmente antropomórfica, es lo que los lanzó a la popularidad. Pero ¿cómo deben responder a las emociones de los humanos? Hasta el momento, la tendencia en PLN ha sido responder con empatía como herramienta para regular las emociones del usuario, amplificando aquellas positivas, como el orgullo, y calmando emociones negativas, como puede ser el enfado. Sin embargo, esto puede resultar problemático (Curry y Curry, 2023).

Jaswant Singh Chail tenía 19 años cuando fue arrestado en 2021 por planear un atentado contra la Reina Elizabeth II de Inglaterra. Durante el juicio declaró que un chatbot le había dicho que estaba «impresionado» cuando aquel expresó orgullo sobre sus planes y se vio alentado a llevarlos a cabo. Aunque este puede resultar un caso algo extremo, no es aislado (Cuadra et al., 2024) y demuestra uno de los problemas con la empatía en la inteligencia artificial: estos modelos no *entienden* las emociones, cómo nos motivan, ni las posibles consecuencias. Amplificar emociones positivas puede ser malo si esa emoción no es adecuada, y viceversa, calmar emociones negativas puede traer consecuencias significativas para las personas. Por ejemplo, la ira nos remarca que algo no es justo y es un ingrediente muy importante en el activismo (Lorde, 1984).

Más allá de la IA, filósofos como Prinz (2011), Bloom (2017) y Breithaupt (2019), han problematizado ya la empatía entre humanos, por ser fácil de manipular y parcial (somos más empáticos

con gente de nuestro grupo social). Sin embargo, no existen estudios explorando la empatía en el contexto de la IA, tampoco sobre los tipos de contextos en los que surge y puede llegar a ser apropiada, o investigación acerca de las consecuencias de su aparición. Bloom, por ejemplo, propone una emoción más distante, la compasión; sin embargo, en la actualidad resulta prácticamente imposible obtener una respuesta que no sea empática cuando tratamos con un LLM –incluso cuando hablamos de cometer crímenes o de lenguaje del odio–.

Si retomamos el argumento de Picard sobre la inteligencia y las emociones, resaltan las carencias de los modelos de hoy en día en términos de inteligencia emocional, pero también, la falta de consideración de conocimientos y teorías fuera de la IA, que nos puedan orientar a la hora de mejorar nuestros modelos como herramientas.

Conclusión

Los recientes avances en el ámbito del PLN gracias a la introducción de modelos de lenguaje a gran escala han abierto las puertas a la interdisciplinariedad. Hasta hace un tiempo, la tecnología no nos permitía pensar en cuestiones como los sesgos de género emocionales, ni la empatía; los modelos como chatGPT parecían posibles solo en la ciencia ficción. El PLN sigue teniendo problemas técnicos abiertos, pero ahora debemos pensar también en el mundo en el que existen nuestras tecnologías y cómo las usamos, y es aquí donde las humanidades y las ciencias sociales nos pueden ayudar a ser mejores informáticos.

La IA Afectiva es una línea de investigación inherentemente interdisciplinaria que une a la informática, la lingüística, la psicología, la filosofía, etc., y todas son necesarias para entender el fenómeno de las emociones y para enmarcar los modelos que estamos construyendo en el ambiente social que estamos modelando. Por definición, los modelos nos ofrecen una visión simplificada de un fenómeno, pero un buen modelo no puede ser demasiado

simplista: debe representar el fenómeno con la fidelidad *necesaria*. La historia de las emociones en PLN nos muestra cómo podemos sobresimplificar si no preguntas tales como para qué es nuestro modelo, quién va a usarlo y cómo. Estas no son interrogantes que debamos resolver solos; nuestras co-disciplinas llevan haciéndoselas hace siglos, si no, milenios. Sin embargo, no es una conversación unilateral: la informática puede aprender de otras disciplinas, pero también, podemos involucrarlas en nuestro trabajo y pedirles lo que necesitamos de ellas. ¿Qué necesitamos de las humanidades para mejorar la IA?

Referencias bibliográficas

- Arya, N. (2023). GPT-4 Details Have Been Leaked! - KDnuggets. *2KDnuggets*. Publicado el 19 julio 2023, accedido el 2 julio 2024.
- Barrett, L. F. (2017). *How Emotions are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.
- Bloom, P. (2017). *Against Empathy: The Case for Rational Compassion*. HarperCollins.
- Brady, M. (2019). *Emotion: The Basics* (1st ed.). Routledge. <https://www.routledge.com/Emotion-The-Basics/Brady/p/book/9781138081390>
- Breithaupt, F. (2019). *The Dark Sides of Empathy* (A. B. B. Hamilton, Trans.). Cornell University Press.
- Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-18).
- Curry, A. (2022). *An Apologia for Anger With Reference to Early China and Ancient Greece* (Doctoral dissertation, UC Riverside).
- Curry, A. C., & Curry, A. C. (2023). Computer says “no”: The case against empathetic conversational AI. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 8123-8130).
- Darwin, C., & Griffith, T. (1874). *The descent of man* (Vol. 4). New York: Prometheus Books.
- Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elsevier.
- Hume, B. (1739) *Tratado de la Naturaleza Humana* 1739–40, T II.3.1 399

- Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of adolescence*, 29(4), 589-611.
- Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12-24).
- Lorde, A. (1984). The uses of anger: Women responding to racism. *Sister outsider*, 127, 131.
- Picard, R. W. (2000). *Affective computing*. MIT press.
- Plaza del Arco, F. M., Curry, A., Cercas Curry, A., & Hovy, D. (2024, May). Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5696-5710. <https://aclanthology.org/2024.lrec-main.506.pdf>
- Prinz, J. (2011). Against Empathy. *Southern Journal of Philosophy* 49 (s1):214-233.
- Shrawgi, H., Rath, P., Singhal, T., & Dandapat, S. (2024, March). Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1841-1857).
- Stauffer, D. J. (2008). Aristotle's Account of the Subjection of Women. *The Journal of Politics*, 70(4), 929-941. <https://doi.org/10.1017/s0022381608080973>

Usuarios de IA generativa responsables de obras mal atribuidas a grandes artistas

Javier Moyano¹

Imagina que tienes la oportunidad de crear la próxima obra de uno de tus pintores o músicos favoritos. Bien. Sería razonable suponer que esta nueva creación no cumpla con los estándares técnicos ni estéticos de sus obras anteriores y, con cierta probabilidad, que caiga en el olvido. Con esta idea se asume la enorme dificultad que sería imitar el estilo de unas obras o clonar el talento creativo de un gran artista. Pero te pido que consideres lo contrario: que realmente tengas la posibilidad de crear, mediante el uso de tecnología muy avanzada, una obra que se integre a la perfección al trabajo de tu artista favorito.

Ahora, imagina que eres ese artista que tenías en mente. Cada obra que has creado a lo largo de tu carrera es el resultado de tu talento creativo y refleja tu estilo característico e inconfundible. Sin embargo, te encuentras en una situación desconcertante: te das cuenta de que no solo alguien más, sino todo el mundo, puede crear una nueva obra a tu manera, donde se aprecian los elementos distintivos de tu arte y se te atribuya la autoría. Esta situación plantea una potencial confusión para tu público y un posible daño a tu reputación como artista.

Lo que inicialmente parecía un sueño, y luego una pesadilla, ahora se ha convertido en una realidad sumamente preocupante. Todo esto se debe a los considerables avances en inteligencia artificial

¹ Realizador en cine y televisión por la Universidad de Chile. Comunicador Audiovisual, Investigador de Realidad Virtual y Ensayista de Arte e Inteligencia Artificial Generativa. Correo: yosoyjaviermoyano@gmail.com

(IA), especialmente en el campo de la IA generativa. Esta tecnología facilita el proceso creativo de nuevas obras en áreas artísticas como la pintura y la música. Las obras de arte generadas por las distintas herramientas tecnológicas de IA muestran resultados muy convincentes similares a los de los artistas humanos, tanto en términos visuales como sonoros.

Las IA generativas tienen la capacidad de crear nuevas obras mediante la síntesis de grandes cantidades de datos con los que han sido entrenadas y los datos introducidos por los usuarios en forma de texto o *prompt*, imágenes y audios. Estas herramientas realizan análisis de datos y reconocimiento de patrones sobre obras existentes, incluyendo las de grandes artistas, así como, las creaciones esperadas. En general, pueden recombinar las características distintivas de la obra de un artista para generar algo nuevo, ya sea un cuadro o una canción. Además, pueden transferir el estilo visual de una imagen a otra o el estilo vocal de un cantante a un nuevo audio.

En la actualidad, existe una amplia variedad de IA generativas que tienen la capacidad de generar obras y contenido infinito a partir de todo lo que ha sido creado (Guerrero-Solé, 2024). Entre estas herramientas, también conocidas como «generadores», destacan Dall-e, Midjourney y Stable Diffusion para creaciones visuales, así como AIVA, Kits AI y Suno para creaciones sonoras y musicales. Todas estas IA están disponibles en la web y son accesibles para cualquier usuario interesado en explorar su potencial creativo. Su uso no requiere de competencias avanzadas en arte y puede compararse con el proceso de encargar una obra. Sin embargo, los usuarios también tienen la opción de involucrarse más en el proceso creativo, pudiendo asumir un rol más propio de un artista al aportar datos elaborados, iterar (es decir, repetir varias veces el proceso generativo hasta alcanzar la creación deseada) y realizar ediciones a la obra generada.

Actualmente, estas obras son generadas con facilidad y se difunden rápidamente por la web sin necesidad de un contexto, una etiqueta o una marca de agua que evidencie su origen, lo que favorece la confusión y el engaño del público. La información sobre el origen

de estas creaciones, que se obtiene a partir del título o la descripción, puede ser incorrecta, manipulada o incompleta, lo que genera preocupación, especialmente, cuando se identifican características distintivas que serían atribuibles a artistas específicos. Esta situación puede llevar al público a entenderlas, no como lo que realmente son, sino como auténticas obras de los artistas originales o como creaciones falsas, debido a atribuciones erróneas, ya sea por error involuntario o con la deliberada intención de confundir y engañar.

Este escenario, en el cual se imita la obra de un artista o se clona el talento del propio artista, puede compararse con los *deepfakes* o ultrafalsos, que son imágenes, audios y videos modificados mediante IA para cambiar o reemplazar la apariencia o las acciones de una persona. La interpretación como algo real de este contenido, especialmente las obras de arte que parecen auténticas pero que en realidad no lo son, podría seguir en aumento. Sin embargo, también existe la posibilidad de que, debido a la creciente dificultad para distinguir entre lo auténtico y lo falso, todo tipo de contenido, incluso el auténtico, podría ser percibido en adelante como falso (Abd-Almageed, 2023, citado en O'Brien y Lajka, 2023).

Para las próximas décadas se espera una intensa competencia entre la inteligencia humana y la IA, y en este escenario, la responsabilidad ética emerge como un pilar fundamental para abordar los desafíos asociados (Terrones, 2018). Desde la difusión del contenido falso y la atribución errónea hasta la devaluación de obras de arte, el reemplazo de los artistas y el descrédito del arte en su totalidad (Albar, 2024). Para los usuarios de IA generativa, tanto la creación como la publicación de obras atribuibles al talento creativo de un artista, pueden parecer actividades recreativas e inofensivas. Sin embargo, como dice la conocida frase, «un gran poder conlleva una gran responsabilidad». Y a cada persona le corresponde hacerse responsable, tanto de sí misma como de la sociedad a la que pertenece, la humanidad en su conjunto y también el mundo (Polo, 2019).

Frente a esta situación, sería prudente enviar una señal de alerta a la comunidad interesada en la ética de la IA generativa, con el propósito de buscar formas de asegurar la responsabilidad

de los usuarios. En otras palabras, que aquellos que participen en la creación y publicación de obras atribuibles a grandes pintores y músicos, deban actuar con responsabilidad, adoptando medidas para evitar posibles consecuencias negativas. Y, además, deban asumir la responsabilidad de sus acciones y omisiones, reconociendo y aceptando las posibles repercusiones que estas puedan tener.

Hasta ahora, ¿tú consideras posible que una nueva obra creada mediante IA pueda competir con un original de Goya o Takashi Murakami, de Beethoven o Kanye West? Hace unos años, antes de que las IA generativas se volvieran populares en la web, ya se observaban casos en los que empresas tecnológicas, experimentando con las IA, comenzaban a entrenarlas con grandes cantidades de datos de ciertos artistas. El objetivo era poder generar con precisión nuevas obras que imitaran sus estilos o clonaran sus talentos creativos, es decir, reflejar su hacer, desde la elección de la paleta de colores y pinceladas, hasta la manera de componer y de cantar.

En el ámbito de la pintura, un primer gran ejemplo de creación mediante IA de una obra atribuible a un gran artista fue el cuadro *The Next Rembrandt*, generado en 2016, a cargo de la agencia JWT Amsterdam. Esta obra fue posible gracias al análisis de datos y el reconocimiento de patrones en píxeles de más de trescientos cuadros del pintor barroco Rembrandt, fallecido hace más de tres siglos. Apoyándose en esa información visual, la IA buscó replicar el estilo característico e inconfundible de la obra del artista. En el ámbito musical, un ejemplo comparable ocurrió en 2019, a cargo de Huawei. En este caso, el objetivo fue crear los dos movimientos faltantes de la *Sinfonía n.º 8 en Si menor, D. 759*, de Franz Schubert, compositor clásico y romántico cuya obra había quedado incompleta hace dos siglos.

Para ambos casos, la creación de estas obras mediante IA pudo generar opiniones sobre si son dignas de comparación con las creaciones originales y auténticas, si efectivamente guardan similitud con ellas, si están bien ejecutadas, si logran evocar

emociones o si son del agrado del público, entre otros aspectos. Sin embargo, en general, el hecho de generar estas obras mediante IA y luego publicarlas, no parece plantear mayores problemas éticos o legales. Las dos creaciones no deberían ser motivo de controversia debido a una presunta transgresión de valores ni por infringir la ley de algún Estado.

Por un lado, los derechos de autor existen y sirven para preservar el vínculo personal del creador con su obra y regular su explotación y retribución económica, pero también, permiten ciertos usos legítimos, como la investigación o la parodia, siempre que se diferencie de la obra original y no signifique una explotación encubierta de ella. Por otro lado, los grandes artistas mencionados ya no están vivos y, tras setenta años de su muerte, los derechos sobre sus obras han expirado. Sin embargo, la situación cambia notablemente al considerar ejemplos de músicos y pintores vivos, cuyas obras, además, se encuentran protegidas.

Una vez más, te pido que imagines. Imagina un escenario donde tú o un usuario de IA generativa crea un cuadro que es indistinguible de las obras expuestas por un pintor en plena actividad, o bien, una canción que no se pueda diferenciar de las pistas del último álbum de un músico vivo y en activo. ¿Qué implicaciones crees que tendría la publicación de una de estas obras? Permitamos que la realidad misma nos dé una respuesta a partir de los casos que significaron un punto de inflexión en el desarrollo de la IA generativa.

En el ámbito de la pintura, no fue una sola obra, sino miles, las que se dieron a conocer en la web y podrían ser fácilmente atribuidas a un artista, confundiendo a más de alguien. Entre 2022 y 2023, Greg Rutkowski, pintor de cuadros de fantasía épica, se convirtió en el artista más utilizado como *prompt* por los usuarios de IA generativa. Hoy en día, resulta difícil distinguir entre su obra auténtica, como *Secret Pass - Eagle Nest*, y aquellas que no lo son, pero que igualmente aparecen en los resultados de búsqueda de imágenes. En este contexto, es importante recordar que la pintura generada mediante IA imita el resultado de un trabajo manual. Sin embargo, en la música, no se trata solo de replicar una acción particular, como

un intérprete tocando de cierta manera un instrumento, sino que también, está abierta la posibilidad de clonar una cualidad única, como el timbre de voz.

En el ámbito musical, dos obras en particular provocaron gran controversia debido a que daban la impresión de ser auténticas colaboraciones entre cantantes populares a nivel mundial. A principios de 2023, como parte del proyecto Ghostwriter, un usuario de IA creó mediante un generador la canción de rap *Heart on My Sleeve*, aunque con letra y producción propias. En ella, se puede escuchar la etiqueta del productor musical Metro Boomin, así como las voces de Drake y The Weeknd, las cuales fueron imitadas o, más bien, clonadas mediante un filtro de transferencia de estilo vocal. Hacia finales del mismo año, otro usuario de IA, bajo el proyecto FlowGPT, llevó a cabo una acción similar en la canción de reguetón *Demo 5: nostalgIA*, incorporando las voces de Justin Bieber, Bad Bunny y Daddy Yankee.

Estas dos canciones se volvieron rápidamente virales en la web, en parte, debido a la calidad superior de ambas obras en comparación con los últimos sencillos de los músicos involucrados. Luego de un tiempo, las dos canciones fueron retiradas de YouTube y Spotify debido a reclamaciones de los sellos discográficos por infracción de derechos de autor. A pesar de ello, lograron ser resubidas a varias plataformas web y siguen en línea hasta el día de hoy. Ahora, y aun conociéndose sus orígenes, tanto los creadores como los públicos que difunden las obras, han añadido etiquetas como (IA), (versión IA) y ([artista] IA) en los títulos o descripciones, aparentemente como un esfuerzo para contextualizar estas creaciones como claramente separadas del trabajo de los artistas originales.

Esto último guarda similitud con la implementación por parte de TikTok del requisito obligatorio de etiquetar como «Generado por IA» a las obras realistas creadas en gran parte o totalmente por IA. Además, se espera que pronto Meta implemente una etiqueta similar, «Hecho con IA». Estas acciones están en línea con una iniciativa de autenticidad de contenido, en la que participan empresas como Adobe, Canon, Microsoft, Nvidia y Universal

Music Group, que busca un estándar de marca de agua para estas obras, desde su creación hasta cuando son vistas o escuchadas. De momento, emerge la pregunta de si estas medidas serán suficientes o simplemente representan un pequeño avance en el camino hacia la atribución adecuada de las obras, con el fin de evitar la confusión y el engaño.

De todos modos, es fundamental que el usuario de la IA generativa, al crear y publicar obras atribuibles a un artista, lo haga con responsabilidad ética. Lo ideal es que el usuario actúe de manera responsable en lugar de depender de las decisiones de empresas desarrolladoras de IA, de los términos y condiciones de los generadores, de las iniciativas particulares como las del público o de los marcos legales de cada país para abordar todos los desafíos éticos. Además, es crucial que el usuario asuma la responsabilidad por la obra que genera, sin importar su grado de implicación y trabajo en el proceso creativo.

En el mejor de los casos, un usuario responsable debe entrenar la IA generativa con obras o identidades artísticas para las cuales tenga autorización, y comprender los alcances de los derechos de autor. Además, debe evaluar cuidadosamente lo que se genera para asegurarse de que cumpla con estándares de calidad artística y ética, lo que podría requerir ajustes en la obra y la decisión de si compartirla o no. Al publicar la creación, debe ser transparente sobre el uso de generadores y atribuir correctamente lo creado, dando crédito tanto a la IA generativa como a los artistas humanos originales. Y, por último, también considerar el impacto social y cultural que la obra pueda tener, anticipándose a las posibles consecuencias tanto positivas como negativas.

Los usuarios son responsables no solo de sus acciones, sino también, de sus omisiones, ya sean deliberadas o inadvertidas. En este sentido, todo depende de las intenciones de la creación y su publicación. Entre los usos bien intencionados de la IA se incluye la creación de una obra como experimentación con el estilo característico e inconfundible de un gran artista, como muestra de homenaje o parodia, y su publicación en esa misma línea. Aunque

esto también pueda parecer un aprovechamiento indebido de la imitación del estilo o la clonación del talento creativo del artista y su reputación para posicionarse como creador, un pintor o músico relevante.

Aunque las intenciones puedan ser buenas, estas no son suficientes. Incluso desde un uso bien intencionado pueden derivarse consecuencias negativas, como la confusión derivada de no proporcionar información necesaria sobre el origen de la obra. Específicamente, entre los usos malintencionados se encuentran: el plagio (es decir, apropiarse del trabajo ajeno y presentarlo como creación propia), omitiendo al artista, y la falsificación (crear una obra falsa para hacerla pasar como auténtica), suplantando al artista.

Todo lo anterior se refiere a las fases de creación y publicación de una obra generada mediante IA, en las que se espera que los usuarios actúen de manera responsable. El mal uso de la IA plantea la pregunta sobre la falta de responsabilidad del usuario. Algunas posibles causas de esto podrían ser la ausencia de valores como el respeto y la honestidad, la indiferencia hacia las implicaciones legales de los derechos de autor, e incluso, la comprensión insuficiente del rol creativo en el uso de la IA y sus consecuencias.

Con todo lo expresado, se ha querido lanzar una señal de alerta sobre el uso de la IA generativa en la creación y publicación de obras de arte atribuibles a grandes artistas. Este interés surge porque estas obras pueden ser tan convincentes y similares a las creaciones auténticas que resulta difícil distinguirlas, lo que afectaría al artista original y provocaría confusión en el público. Además, el hecho de que estas obras se vuelvan cada vez más populares podría tener un impacto negativo en la valoración del arte y los artistas humanos. Sin embargo, con un enfoque responsable por parte de los usuarios de la IA, se podría evitar en gran parte las repercusiones negativas y contribuir positivamente al campo artístico.

Con el avance y la mayor integración de la IA generativa en la creación artística, es esencial abordar de alguna manera las

preocupaciones éticas que esto conlleva. Esta tecnología brinda a los usuarios la oportunidad de expresar obras imaginativas y muy realistas haciendo uso del estilo o el talento artístico de otros. Esto exige reconsiderar la relación con la herramienta y buscar formas de aprovechar su potencial creativo sin afectar negativamente a los artistas originales ni al público. Además, es importante encontrar un equilibrio en el cual se aprovechen las capacidades de los generadores mientras se respetan los valores fundamentales y los derechos de autor.

En este sentido, no se busca limitar la creatividad o restringir la innovación tecnológica, sino más bien hacer un llamado a los usuarios de IA generativa para que actúen de manera responsable y asuman la responsabilidad, en lugar de imponer cargas adicionales a los artistas. En otras palabras, se pretende dar validez a la creación y publicación de este tipo de obras, siempre y cuando los usuarios actúen con cuidado y atención respecto al artista original. Los usuarios deben ser honestos y transparentes sobre el origen de sus creaciones, informando claramente al público que la obra no es creación del artista original, e implementar etiquetas o marcas claras que indiquen cuando una obra ha sido generada mediante IA y no por el artista original. De esta manera, se evitan confusiones y cualquier forma de engaño mediante el uso de IA generativa.

En síntesis, se ha planteado que la creación y publicación de este tipo de obras artísticas, conlleva desafíos éticos y potencialmente legales en cuanto a la atribución creativa y su impacto sobre la identidad artística de los artistas originales y en el público. Además, se ha enfatizado que la responsabilidad ética es fundamental para elaborar alternativas para estos desafíos y sus respectivas consecuencias. Y, en este sentido, tomando como referencia las repercusiones de los casos más destacados, tanto en pintura como en música, se ha buscado una manera de asegurar un uso responsable.

¿Te habías imaginado que crear y publicar lo que sería el próximo cuadro de uno de tus pintores favoritos o la nueva canción de uno de tus músicos preferidos pudiera desencadenar tantos desafíos éticos y potencialmente legales? Es posible que lo hayas

considerado, pero nunca está demás comunicarlo. Espero que este texto pueda contribuir tanto a la comunidad interesada en la ética de la IA generativa, como a sus futuros usuarios responsables de obras atribuidas a grandes artistas. Además, que pueda ser un aporte para sentar las bases para abordar los desafíos éticos que puedan surgir no solo en la pintura y la música, sino también en diversas expresiones artísticas y formas de entretenimiento. El avance de la IA generativa está transformando el panorama artístico como nunca, con un impacto preocupante también en el ámbito audiovisual.

Como realizador en cine y televisión, me siento con la obligación de destacar el rápido avance de las IA generativas en el sector, con generadores de video como Gen-2, Pika y Sora, y la responsabilidad que debe acompañar su uso. Estas herramientas ofrecen a los usuarios la capacidad de generar escenas imaginativas y muy realistas a partir de un *prompt*. Aunque esta tecnología es reciente, ya promete cambios significativos. En el futuro, podrá facilitar considerablemente el proceso creativo de obras con múltiples escenas y de larga duración, resolviendo todas las variables necesarias para la creación audiovisual a gran escala, ya sea para televisión o cine.

La evolución de estos generadores de video podría llevarnos a un punto en el que la producción de obras televisivas y cinematográficas, con los estilos característicos e indistinguibles de nuestros realizadores y estudios preferidos, esté al alcance de usuarios como tú o yo. Podría llegar el día en que podamos indicarle a una IA: «crea una película o una serie de televisión dirigida por tal director, con tal actor como protagonista y basada en tal historia», y verla hecha realidad en nuestra pantalla. Tareas que usualmente requieren meses o años de trabajo humano se realizarían en segundos, desde la escritura de guion, la selección del elenco y el diseño de producción hasta la dirección de fotografía, la composición musical y el montaje, ya sea para obras de ficción o documental, tanto en formato *live action* como animación. Esto sería fascinante, pero al mismo tiempo, aterrador.

Hoy en día, la distinción entre obras creadas por humanos y aquellas generadas mediante inteligencia artificial es cada vez más difusa. En los próximos años, puede que ya no sea necesario especificar si las obras fueron creadas con la ayuda de IA, sino solo cuando sean creaciones humanas de artistas originales. O, tal vez, ni siquiera importe quién haya creado la obra. ¿Será esta la última oportunidad de revalorizar el modo humano de hacer arte? Para enfrentar este y otros desafíos, será crucial que expertos en arte, ética e inteligencia artificial sostengan un diálogo fructífero y con visión de futuro.

Referencias bibliográficas

- Albar, P. (2024). La inteligencia artificial de generación de imágenes en arte: ¿Cómo impacta en el futuro del alumnado en Bellas Artes? *Encuentros. Revista de Ciencias Humanas, Teoría Social y Pensamiento Crítico*, (20), 145–164. doi: 10.5281/zenodo.10052355
- Guerrero-Solé, F. (2024). La comunicación ante el desafío de la inteligencia artificial generativa. Reflexiones, retos y oportunidades en un escenario disruptivo. Barcelona: Cultura Difusa.
- O'Brien, M., y Lajka, A. (19 de enero de 2023). En las obras de IA, ¿quién es el verdadero artista? *Los Angeles Times*. Recuperado de <https://www.latimes.com/espanol/entretenimiento/articulo/2023-01-19/en-las-obras-de-ia-quien-es-el-verdadero-artista>
- Polo, M. (2019). La responsabilidad ética. *Veritas*, (42), 49–72.
- Terrones, A. (2018). Inteligencia artificial y ética de la responsabilidad. *Cuestiones de Filosofía*, 4(22), 141–170. doi:10.19053/01235095.v4.n22.2018.8311

Lavender: la máquina de IA que dirige los bombardeos de Israel en Gaza

Yuval Abraham¹

En 2021, se publicó el libro *The Human-Machine Team: How to Create Synergy Between Human and Artificial Intelligence That Will Revolutionize Our World*, firmado bajo el seudónimo de «Brigadier General Y.S.». El autor –el actual comandante de la unidad de élite 8200 de la inteligencia israelí– defiende el diseño de una máquina que podría procesar rápidamente ingentes cantidades de datos con el fin de generar miles de potenciales «objetivos» para ataques militares en una guerra. Dicha tecnología, escribe, solucionaría el «atasco humano tanto para localizar nuevos objetivos como para la toma de decisiones que aprueben dichos objetivos».

Resulta que una máquina así existe. El ejército israelí ha desarrollado un programa basado en inteligencia artificial (IA) conocido como *Lavender*. Agentes del servicio de inteligencia israelí que han servido en el ejército durante la actual guerra en la Franja de Gaza², usaron la IA con el fin de generar objetivos para asesinar. *Lavender* ha desempeñado un papel fundamental en el bombardeo sin precedentes de palestinos, especialmente, durante las primeras etapas de la guerra. Era tan influyente, que se trataban los resultados de la IA «como si fuera una decisión humana».

¹ Este texto corresponde a una investigación periodística publicada en *+972 Magazine* (03.04.2024) que recoge las declaraciones de seis agentes de la inteligencia israelí identificados en el texto con una inicial. La presente versión ha sido editada con autorización del autor.

² Periodista y director de cine israelí.

² El 7 de octubre de 2023 militantes liderados por Hamás lanzaron un ataque mortal contra las comunidades del sur de Israel, en el que mataron a unas 1.200 personas y secuestraron a 240.

Lavender está diseñado para marcar como posibles objetivos de bombardeo a todos los sospechosos de pertenecer al brazo armado de Hamás y la Yihad Islámica Palestina, incluidos los de rango inferior. Durante las primeras semanas de la guerra, el ejército dependió casi por completo de *Lavender*, que registró hasta 37.000 palestinos –y sus hogares– como presuntos militantes para posibles ataques aéreos.

Durante las primeras fases de la guerra, el ejército autorizó a los oficiales para que asumieran las listas de objetivos a matar que generaba *Lavender*, sin necesidad de comprobar las decisiones de la máquina o de analizar la información sin procesar en la que se basaban. El personal humano a menudo solo servía de «sello» para aprobar las decisiones de la máquina; en general, dedicaban unos «20 segundos» a cada objetivo antes de autorizar un bombardeo, únicamente para asegurarse de que el objetivo marcado era varón. Ello a pesar de saber que el sistema comete «errores» en aproximadamente el 10% de los casos, y de que, en ocasiones, marca a personas que apenas tienen una ligera o ninguna conexión con grupos militantes.

Además, el ejército israelí atacaba sistemáticamente a las personas marcadas como objetivos cuando se encontraban en sus hogares –generalmente de noche, con toda la familia presente– y no en sus actividades militares. Esto se debía a que, desde el punto de vista de los servicios de inteligencia, era más fácil localizarlas en sus domicilios. Otros sistemas automatizados, incluido uno llamado «¿Dónde está papá?», se utilizaban para localizar a personas concretas y llevar a cabo los bombardeos cuando entraban a sus domicilios familiares.

Así, miles de palestinos –la mayoría mujeres y niños o personas no combatientes– fueron aniquilados por los ataques aéreos israelíes, especialmente durante las primeras semanas de la guerra, debido a las decisiones de la IA.

No nos interesaba matar a los operativos [de Hamás] solo cuando estaban en un edificio militar o participaban en una operación militar. [...] Por el contrario, las FDI los bombardeaban en sus casas

sin dudarlo, como primera opción. Es mucho más fácil bombardear la casa de una familia. El sistema está diseñado para buscarlos en estas situaciones (A., agente de los servicios de inteligencia).

Además, cuando se trataba de atacar a supuestos militantes subalternos marcados por *Lavender*, el ejército prefería utilizar únicamente misiles no guiados (bombas «tontas», en contraste con las «inteligentes»), que pueden destruir edificios enteros con sus ocupantes y causan una cantidad importante de bajas. «No conviene desperdiciar bombas caras en gente sin importancia; es muy caro para el país y hay escasez [de esas bombas]», dijo C., otro agente de los servicios de inteligencia.

En una medida sin precedentes, durante las primeras semanas de la guerra, el ejército también decidió que por cada militante subalterno de Hamás marcado por *Lavender*, estaba permitido matar hasta 15 o 20 civiles. Antes los militares no autorizaban ningún “daño colateral” durante los asesinatos de militantes de bajo rango. En cambio, luego, si el objetivo era un comandante de batallón o brigada de Hamás, se autorizó en varias ocasiones la muerte de más de 100 civiles en el asesinato de un solo comandante.

La siguiente investigación está organizada por orden cronológico, de acuerdo con las seis etapas de la producción automatizada de objetivos por parte del ejército israelí, durante las primeras semanas de la guerra de Gaza. A saber:

1. Se explica qué es la máquina *Lavender*.
2. Se desvela el funcionamiento del sistema «¿Dónde está papá?», que rastreaba esos objetivos y avisaba al ejército cuando entraban en sus casas familiares.
3. Se describe cómo se eligieron las bombas «tontas» para atacar estos hogares.
4. Se explica cómo el ejército flexibilizó el número permitido de civiles que podían morir durante el bombardeo de un objetivo.
5. Se señala el modo en que el *software* automatizado calculó de manera inexacta la cantidad de no combatientes en cada hogar.
6. Se muestra cómo en varias ocasiones, cuando se atacaba una vivienda, generalmente de noche, el objetivo individual a veces no estaba dentro porque el personal militar no verificaba la información en tiempo real.

Paso 1: generación de objetivos

En el pasado, en el ejército israelí, el término «objetivo humano» se refería a un agente militar de alto rango que, según las reglas de su Departamento de Derecho Internacional, puede ser asesinado en su domicilio, incluso, con civiles cerca. Durante las guerras anteriores de Israel, dado que se trataba de una forma «especialmente brutal» de matar a alguien –en la que a menudo se mataba a toda una familia junto al objetivo–, dichos objetivos se marcaban cuidadosamente y solo se bombardeaba a altos mandos militares en sus hogares para mantener el principio de proporcionalidad del derecho internacional.

Después del 7 de octubre, el ejército adoptó un enfoque radicalmente distinto. En el marco de la «Operación espadas de hierro», decidió señalar a todos los agentes del brazo militar de Hamás como objetivos humanos, independientemente de su rango o importancia militar.

La nueva política también planteó un problema técnico para los servicios de inteligencia israelíes. En guerras anteriores, para autorizar el asesinato de un solo objetivo humano, un oficial tenía que pasar por un complejo y largo proceso de «incriminación»: comprobar que la persona era un miembro de alto rango del brazo militar de Hamás, averiguar dónde vivía, su información de contacto y, finalmente, saber cuándo estaba en casa en tiempo real. Cuando la lista de objetivos era de unas pocas docenas de altos mandos, el personal de los servicios de inteligencia podía encargarse individualmente de ese trabajo.

Sin embargo, una vez que la lista se amplió e incluyó a decenas de miles de agentes de rango inferior, el ejército israelí pensó que tenía que apoyarse en el *software* automatizado y la IA. El resultado fue que el papel del personal en la incriminación de palestinos combatientes fue descartado y, en su lugar, la IA hizo la mayor parte del trabajo. *Lavender* –que fue desarrollado para crear objetivos humanos en la guerra actual– ha marcado a unos 37.000 palestinos como presuntos «militantes de Hamás», la mayoría de

ellos subalternos, para asesinarlos. El oficial superior B. aclara el motivo para desarrollar esta particular máquina generadora de objetivos para la guerra actual:

No sabíamos quiénes eran los agentes subalternos porque Israel no les seguía la pista de forma rutinaria [antes de la guerra]. [...] Querían permitirtos atacar [a los agentes subalternos] automáticamente. Ese es el Santo Grial. En cuanto se automatiza, la generación de objetivos se vuelve loca.

La aprobación para asumir automáticamente las listas de *Lavender*, antes utilizadas solo como herramienta auxiliar, se dio unas dos semanas iniciada la guerra, después de que el personal de los servicios de inteligencia verificara «manualmente» la precisión de una muestra aleatoria de varios cientos de objetivos seleccionados por la IA. Cuando esa muestra descubrió que los resultados de *Lavender* habían logrado 90% de precisión en la identificación de personas que pertenecían a Hamás, el ejército autorizó su uso generalizado. A partir de ese momento si *Lavender* decidía que un individuo era militante de Hamás, se les pedía que lo trataran como una orden, sin necesidad de comprobar por qué la máquina tomaba esa decisión o examinar la información sin procesar en la que se basaba.

A las 5 de la mañana, [las fuerzas aéreas] llegaban y bombardeaban todas las casas que habíamos marcado. [...] Eliminamos a miles de personas. No los revisábamos uno a uno: pusimos todo en sistemas automatizados, y en cuanto uno [de los individuos marcados] estaba en casa, inmediatamente se convertía en un objetivo. Lo bombardeábamos a él y a su casa [Oficial superior B.].

Los resultados mortales de esta flexibilización de las restricciones en la primera fase de la guerra, fueron asombrosos. Según datos del Ministerio de Salud palestino en Gaza, del que el ejército israelí ha dependido casi exclusivamente desde el comienzo de la guerra, Israel mató a unos 15.000 palestinos –casi la mitad de la cifra de muertos hasta el momento– en las primeras seis semanas de la guerra, hasta que el 24 de noviembre se acordó un alto el fuego de una semana.

Lavender analiza la información recopilada sobre la mayoría de los 2,3 millones de residentes de Gaza a través de un sistema de

vigilancia masiva y, a continuación, evalúa y clasifica la probabilidad de que cada persona pertenezca al brazo armado de Hamás o a la Yihad Islámica Palestina. La IA otorga a casi todas las personas una calificación de 1 a 100 para expresar la probabilidad de que sean militantes.

La IA aprende a identificar las características de los agentes conocidos de Hamás y la Yihad Islámica Palestina, cuya información fue introducida en la máquina como datos de entrenamiento, y luego localiza esas características entre la población. Un individuo que tenga varias de esas características alcanzará una calificación alta y, por lo tanto, se convertirá en un posible objetivo de asesinato.

En la práctica, los agentes aquí entrevistados que han utilizado *Lavender* en los últimos meses sostienen que la mediación humana y la precisión, fueron sustituidas por la creación masiva de objetivos y por la letalidad. B., un oficial de alto rango que usó *Lavender*, reiteró que, en la guerra actual, los oficiales no estaban obligados a revisar de forma independiente las evaluaciones de la IA. El fin era ahorrar tiempo y permitir la producción en masa y sin obstáculos de objetivos humanos: «Todo era estadístico, todo estaba ordenado». Señaló que esta falta de supervisión estaba permitida a pesar de que los controles internos mostraban que los cálculos de *Lavender* se consideraban precisos, aunque se sabía de antemano que el 10 % de los objetivos humanos programados para ser asesinados no eran en absoluto miembros del brazo armado de Hamás.

Por ejemplo, las fuentes explicaron que *Lavender*, a veces, marcaba erróneamente a personas que tenían patrones de comunicación similares a los de agentes de Hamás o de la Yihad Islámica en Palestina, incluido personal de policía y de defensa civil, familiares de militantes, residentes que casualmente tenían un nombre y apodo idénticos al de un agente, y gazatíes que utilizaban un dispositivo que había pertenecido a un agente de Hamás. En tal sentido, como señala una fuente crítica con la inexactitud de *Lavender*:

¿Hasta qué punto tiene que estar una persona cerca de Hamás para que una máquina de los servicios de inteligencia artificial la

considere miembro de la organización? (...) Es una frontera imprecisa. ¿Una persona que no recibe un salario de Hamás, pero les ayuda en todo tipo de cosas, es un agente de Hamás? ¿Alguien que estuvo en Hamás en el pasado, pero que ya no lo está hoy, es un agente de Hamás? Cada una de estas especificaciones –características que una máquina consideraría sospechosas– es inexacta.³

El ejército sabía que la mínima supervisión humana establecida no descubriría estos fallos. «No existía una política de ‘cero errores’. Los errores se trataban estadísticamente», dijo una fuente que utilizó *Lavender*. «Debido al alcance y la magnitud de la situación, el protocolo era que, aunque no se estuviera seguro de que la máquina acertara, se sabía que estadísticamente acertaba. Así que se hacía.» Por su parte, B., la fuente principal, señala:

Ha demostrado su eficacia. (...) Hay algo en el enfoque estadístico que te fija en una determinada norma y estándar. Ha habido una cantidad ilógica [de bombardeos] en esta operación. No recuerdo nada similar en mi vida. Y confío mucho más en un mecanismo estadístico que en un soldado que ha perdido a un amigo hace dos días. Todos los presentes, incluido yo, perdimos personas el 7 de octubre. La máquina funcionaba con frialdad. Y eso lo hizo más fácil.

Otra fuente de los servicios de inteligencia, que defendió la confianza en las listas de asesinatos de sospechosos palestinos generadas por *Lavender*, argumentó que solo valía la pena invertir el tiempo de un agente de los servicios de inteligencia para verificar la información, si el objetivo era un alto mando de Hamás. «Pero cuando se trata de un militante subalterno, no conviene invertir tiempo ni personal en ello [...]. En la guerra, no hay tiempo para incriminar a todos los objetivos. Así que estás dispuesto a asumir el margen de error que supone utilizar inteligencia artificial, arriesgarte a que haya daños colaterales y a que mueran civiles, y a atacar por error y vivir con ello.»

³ Existen problemas similares, según señala una fuente, con la capacidad de las máquinas generadoras de objetivos para evaluar el teléfono utilizado por una persona marcada para ser asesinada: «En la guerra, los palestinos cambian de teléfono continuamente [...]. La gente pierde el contacto con sus familias, le da su teléfono a un amigo o a su esposa y tal vez lo pierde. No hay manera de confiar al 100 % en el mecanismo automático que determina a quién pertenece un número [de teléfono].»

B. dijo que la razón de esta automatización era el impulso constante de generar más objetivos para asesinar. «En un día sin objetivos [cuyo índice de calificación fuera suficiente para autorizar un ataque], atacábamos con un criterio más bajo. Nos presionaban constantemente: ‘Traednos más objetivos’. Nos gritaban mucho. Acabábamos con [la vida de] nuestros objetivos muy rápido.» Explicó que, al reducir el criterio de calificación de *Lavender*, se marcaba a más personas como objetivos:

En su punto álgido, el sistema logró generar 37.000 personas como potenciales objetivos humanos [...]. Pero las cifras cambiaban todo el tiempo, porque depende de dónde pongas el listón de lo que es un agente de Hamás. Hubo momentos en que la definición de un agente de Hamás se llevaba a cabo de manera más amplia y, de este modo, la máquina comenzaba a traernos todo tipo de personal de defensa civil, agentes de policía, con los que sería una lástima desperdiciar bombas. Ayudan al gobierno de Hamás, pero en realidad no ponen en peligro a los soldados.

Una fuente que trabajó con el equipo científico de datos militares que entrenó a *Lavender*, dijo que en la máquina también se introdujeron los datos recopilados de los empleados del Ministerio de Seguridad Interna dirigido por Hamás, a los que no considera militantes. «Me molestó el hecho de que, cuando se entrenó a *Lavender*, usaron el término ‘agente de Hamás’ de manera poco precisa y se incluyó a personas que eran trabajadores de defensa civil en el conjunto de datos de entrenamiento.» Así, continúa, incluso si uno cree que estas personas merecen ser asesinadas, entrenar al sistema en función de sus perfiles de comunicación, hizo que hubiera más probabilidades de que *Lavender* seleccionara civiles por error cuando sus algoritmos se aplicaron a la población general:

Al tratarse de un sistema automático que no es manejado manualmente por humanos, el significado de esta decisión es dramático: significa que se incluyen como objetivos potenciales a muchas personas con un perfil de comunicación civil.

Pese a esto, el portavoz de las FDI [Fuerzas de Defensa de Israel] negó el uso de los servicios de inteligencia artificial con el fin de incriminar objetivos, declarando que son «herramientas auxiliares

que ayudan a los oficiales en el proceso de incriminación.»⁴ Pero, las fuentes lo desmintieron. Afirmaron que el único protocolo de supervisión humana que se aplicaba antes de bombardear las casas de los presuntos militantes «subalternos», marcados por *Lavender*, consistía en llevar a cabo una única comprobación: que el objetivo fuera varón. En el ejército se partía de la base de que, si el objetivo era una mujer, probablemente la máquina había cometido un error: no hay mujeres en el brazo militar de Hamás y en la Yihad Islámica.

«Un ser humano tenía que hacerlo [verificar el objetivo] en tan solo unos segundos», dijo B., y explicaba que esto se convirtió en el protocolo después de darse cuenta de que el sistema *Lavender* «acertaba» la mayoría de las veces. Para llevar a cabo la comprobación hombre/mujer, B. afirmó que «en esa fase invertía 20 segundos por cada objetivo y hacía docenas al día. No aportaba nada como humano, aparte de ser un sello de aprobación. Ahorraba mucho tiempo. Si [el agente] aparecía en el mecanismo automatizado y yo comprobaba que era un hombre, habría permiso para bombardearlo, previo examen de los daños colaterales.»

En la práctica esto significaba que, para los hombres civiles que *Lavender* marcaba por error, no existía ningún mecanismo de supervisión para detectar dicho fallo. Según B., se producía un error común «si el objetivo [de Hamás] le daba [su teléfono] a su hijo, a su hermano mayor o simplemente a un hombre al azar. Esa persona será bombardeada en su casa con su familia. Esto sucedió a menudo. Estos fueron la mayoría de los errores causados por *Lavender*.»

Paso 2: vinculación de objetivos con hogares familiares

La siguiente etapa en el procedimiento de asesinato del ejército israelí consiste en identificar dónde atacar a los objetivos que genera *Lavender*.

⁴ La declaración de las FDI proseguía: «En cualquier caso, se requiere un examen independiente por parte de un analista [de los servicios de inteligencia], que verifique la legitimidad de atacar los objetivos identificados de acuerdo con las condiciones establecidas en las directivas de las FDI y el derecho internacional.»

El portavoz de las FDI afirmó que «Hamás coloca sus agentes y activos militares en el corazón de la población civil, utiliza sistemáticamente a la población civil como escudos humanos y lleva a cabo combates desde el interior de construcciones civiles, incluidos lugares vulnerables como hospitales, mezquitas, escuelas e instalaciones de la ONU. Las FDI están sujetas al derecho internacional y actúan de acuerdo con él y dirigen sus ataques únicamente contra objetivos y operativos militares.»

Las seis fuentes con las que hablamos, en cierta medida, repitieron esto mismo al decir que el extenso sistema de túneles de Hamás pasa deliberadamente por debajo de hospitales y escuelas; que los militantes de Hamás utilizan ambulancias para desplazarse; y que innumerables activos militares han sido situados cerca de edificios civiles. Las fuentes argumentaron que muchos ataques israelíes matan a civiles como consecuencia de estas tácticas de Hamás, una caracterización que, según advierten los grupos de derechos humanos, elude la responsabilidad de Israel de causar las bajas.

Sin embargo, en contraste con las declaraciones oficiales del ejército israelí, las fuentes explicaron que una de las principales razones del número sin precedentes de víctimas mortales, causado por los actuales bombardeos de Israel, es el hecho de que se ha atacado sistemáticamente objetivos en sus domicilios particulares junto a sus familias. En parte porque, desde el punto de vista de los servicios de inteligencia, era más fácil marcar las casas familiares empleando sistemas automatizados.

A diferencia de los numerosos casos de operativos de Hamás que participaban en actividades militares desde zonas civiles, en el caso de los ataques letales sistemáticos, el ejército, de forma rutinaria, tomaba la decisión activa de bombardear a militantes sospechosos cuando se encontraban dentro de hogares donde no tenía lugar ninguna actividad militar. Esta elección era un reflejo de la forma en que está diseñado el sistema israelí de vigilancia masiva en Gaza.

Puesto que todo el mundo en Gaza tenía una casa con la que se le podía asociar, los sistemas de vigilancia del ejército podían «vincular» fácil y automáticamente a las personas con las casas familiares. Para identificar en tiempo real el momento en que los agentes entran en sus casas, se han desarrollado varios *softwares* automáticos adicionales. Estos programas rastrean a miles de personas simultáneamente, identifican cuándo están en casa y envían una alerta automática al oficial encargado de señalar los objetivos, que luego marca la casa para bombardearla. Uno de varios de estos *softwares* de seguimiento se llama «¿Dónde está papá?».

Los datos también demuestran claramente esta política. Según cifras de la ONU, durante el primer mes de la guerra, más de la mitad de las víctimas mortales –6.120 personas– pertenecían a 1.340 familias, muchas de las cuales fueron exterminadas cuando se encontraban dentro de sus hogares. La proporción de familias enteras bombardeadas en sus casas en la guerra actual, es mucho mayor que en la operación israelí de 2014 en Gaza (que anteriormente fue la guerra más mortífera de Israel en la Franja), lo que sugiere, aún más, la relevancia que ha tomado esta política.

Otra fuente dijo que cada vez que disminuía el ritmo de los asesinatos, se añadían más objetivos a sistemas como «¿Dónde está papá?» para localizar a personas que entraban en sus casas, y que, por tanto, podrían ser bombardeadas. Afirmó que la decisión de a quién incluir en los sistemas de seguimiento podría ser tomada por oficiales de rango relativamente bajo en la jerarquía militar: «Un día, totalmente por mi cuenta, añadí unos 1.200 objetivos nuevos al sistema [de seguimiento], porque el número de ataques [que estábamos llevando a cabo] disminuyó [...]. Y esas decisiones no se tomaban a altos niveles.»

Las fuentes dijeron que, en las dos primeras semanas de la guerra, inicialmente se introdujeron «varios miles» de objetivos en programas de localización como «¿Dónde está papá?». Entre ellos se encontraban todos los miembros de la unidad de élite de las fuerzas especiales de Hamás, la Nukhba, todos los operativos antitanques de Hamás y cualquiera que hubiera entrado en Israel el 7 de octubre. Pero al poco tiempo, la lista de muertos se amplió drásticamente. «Al

final eran todos [los marcados por *Lavender*]», explicó una fuente. «Decenas de miles. Esto ocurrió unas semanas más tarde, cuando las brigadas [israelíes] entraron en Gaza y ya había menos personas no implicadas [es decir, civiles] en las zonas del norte.»⁵

De este modo, *Lavender* y sistemas como «¿Dónde está papá?» se combinaban con consecuencias letales que mataban a familias enteras. Al añadir al sistema de seguimiento domiciliario «¿Dónde está papá?» un nombre de las listas generadas por *Lavender*, explicó A., la persona marcada quedaba bajo vigilancia permanente y podía ser atacada tan pronto como pusiera un pie en su casa. Se tiraba la casa abajo con todos los que estuvieran dentro:

Digamos que se calcula que hay un agente militar de Hamás más 10 [civiles en la casa]. [...] Por lo general estos 10 serán mujeres y niños. De modo que, absurdamente, resulta que la mayoría de las personas que mataste eran mujeres y niños.

Paso 3: elección del arma

En el momento en que *Lavender* ha marcado un objetivo para asesinarlo, el personal del ejército ha verificado que se trata de varones y el *software* de rastreo ha localizado al objetivo en casa, la siguiente fase es elegir la munición con la que bombardearlos.

En diciembre de 2023, CNN informó que, según estimaciones de los servicios de inteligencia estadounidenses, alrededor del 45% de las municiones utilizadas por las fuerzas aéreas israelíes en Gaza eran bombas «tontas», las cuales se sabe que causan más daños colaterales que las bombas guiadas. En respuesta a CNN, un portavoz del ejército dijo: «Como militares comprometidos con el derecho internacional y con un código moral de conducta, estamos dedicando ingentes recursos a minimizar los daños a los civiles a los que Hamás ha obligado a asumir el papel de escudos humanos. Nuestra guerra es contra Hamás, no contra el pueblo de Gaza.»

⁵ *Lavender*, incluso, marcó a algunos menores de edad como objetivos de los bombardeos: «Normalmente, los combatientes tienen más de 17 años, pero no era una condición.»

Sin embargo, los combatientes de menor rango marcados por *Lavender* eran asesinados únicamente con bombas tontas, para ahorrar armamento más caro. El ejército no atacaría a un objetivo menor si vivía en un edificio alto porque no quería gastar una «bomba para destruir una planta» (con un efecto colateral más limitado), más precisa y costosa. Mas, si un objetivo menor vivía en un edificio de pocas plantas, el ejército estaba autorizado a matarlo y a todos los que estaban en el edificio con una bomba tonta. Según C., que ha utilizado varios programas automatizados en la guerra actual, «ocurrió así con todos los objetivos menores»:

La única pregunta era: ¿es posible atacar el edificio teniendo en cuenta los daños colaterales? Porque normalmente llevábamos a cabo los ataques con bombas tontas, y eso significaba literalmente destruir toda la casa encima de sus ocupantes. Pero, incluso, si se evita un ataque, te da igual: inmediatamente pasas al siguiente objetivo. Gracias al sistema, los objetivos nunca se acaban. Tienes otros 36.000 esperando.

Paso 4: autorización de bajas civiles

Al atacar a agentes de menor rango, incluidos aquellos marcados por sistemas de los servicios de inteligencia artificial como *Lavender*, el número permitido de civiles para matar junto a cada objetivo se fijó, durante las primeras semanas de la guerra, en 20 (otra fuente habló de 15). Estos «límites de daños colaterales», como los denominan los militares, se aplicaban de forma generalizada a todos los presuntos militantes subalternos, independiente de su rango, importancia militar y edad, y sin ningún análisis específico de cada caso para sopesar la ventaja militar de asesinarlos frente al daño que pudiera causarle a la población civil.

Según A., que en la guerra actual fue oficial en una sala de operaciones de objetivos, el Departamento de Derecho Internacional del ejército nunca había dado una «aprobación tan amplia» a un límite tan alto de daños colaterales. «No se trata solo de que se pueda matar a cualquier persona que sea soldado de Hamás, lo cual está claramente permitido y es legítimo en términos del derecho

internacional», dijo A., «sino que directamente te dicen: 'Puedes matarlos junto con muchos civiles'». De tal modo: «Cada persona que vistiera uniforme de Hamás en el último año o dos, podía ser bombardeada junto con 20 daños colaterales [civiles], incluso, sin un permiso especial». No existía «el principio de proporcionalidad». Según A., esta fue la política durante la mayor parte del tiempo que estuvo en el cargo. Solo más tarde, los militares redujeron el límite de daños colaterales. «En este cálculo, también podrían ser 20 niños para un agente militar de menor rango [...]. No era así en el pasado.» El motivo de esta política era la «Letalidad». El límite predeterminado y fijo de daños colaterales ayudó a acelerar la creación masiva de objetivos, utilizando la máquina *Lavender*, porque ahorró tiempo.

Ahora, en parte, debido a la presión estadounidense, el ejército israelí ya no genera en masa objetivos humanos de rango inferior susceptibles de ser bombardeados en viviendas civiles. El hecho de que la mayoría de las viviendas de la Franja de Gaza ya hayan sido destruidas o dañadas, y de que casi toda la población haya sido desplazada, también afectó a la capacidad del ejército para confiar en las bases de datos de los servicios de inteligencia y en los programas automatizados de localización de viviendas. Por otro lado, según E., ahora «Hay una economía de municiones», debido a que «Siempre temieron que hubiera [una guerra] en el norte [con Hezbolá en el Líbano]. Ya no atacan a este tipo de personas [de menor rango]».

Sin embargo, los ataques aéreos contra comandantes de alto rango de Hamás persisten, y las fuentes dijeron que el ejército está autorizando el asesinato de «cientos» de civiles por objetivo. Una política oficial para la cual no existe ningún precedente histórico en Israel, y ni siquiera, en las recientes operaciones militares estadounidenses.

«En el bombardeo del comandante del Batallón Shuja'iya sabíamos que mataríamos a más de 100 civiles», recordó B. sobre un bombardeo que tuvo lugar el 2 de diciembre que, según el portavoz de las FDI, tenía como objetivo asesinar a Wisam Farhat. «Para mí, psicológicamente, fue extraño. Más de 100 civiles: cruza una línea roja».

Amjad Al-Sheikh, un joven palestino de Gaza, dijo que muchos de sus familiares murieron en ese bombardeo. Residente de Shuja'iya, al Este de la ciudad de Gaza, se encontraba ese día en un supermercado local cuando escuchó cinco explosiones que hicieron añicos los cristales de las ventanas:

Corrí a casa de mi familia, pero ya no había edificios. [...] La calle se llenó de gritos y humo. Bloques residenciales enteros se convirtieron en montañas de escombros y agujeros profundos. La gente empezó a buscar en el cemento usando las manos, y yo también, buscando señales de la casa de mi familia.

La esposa y la hija pequeña de Al-Sheikh sobrevivieron – protegidas de los escombros por un armario que les cayó encima– pero encontró a otros 11 miembros de su familia, entre ellos sus hermanas, hermanos y sus hijos pequeños, muertos bajo los escombros. Según el grupo de derechos humanos B'Tselem, el bombardeo de ese día destruyó decenas de edificios, mató a decenas de personas y enterró a cientos bajo las ruinas de sus hogares.

No obstante, hubo ataques aún más mortíferos. Para asesinar a Ayman Nofal, comandante de la Brigada Central de Gaza de Hamás, el ejército autorizó el asesinato de aproximadamente 300 civiles y la destrucción de varios edificios en ataques aéreos contra el campo de refugiados de Al-Bureij el 17 de octubre. Todo lo cual se hizo basándose en una localización imprecisa de Nofal.

«En el ataque se destruyeron entre 16 y 18 casas», dijo Amro Al-Khatib, residente del campamento. «No podíamos distinguir un piso de otro; todos quedaron mezclados entre los escombros y encontramos partes de cuerpos humanos por todas partes.» Después, Al-Khatib recordó que se sacaron unos 50 cadáveres y unas 200 personas heridas de entre los escombros, muchas de ellas de gravedad. Pero eso fue solo el primer día. Los residentes del campo pasaron cinco días sacando a los muertos y heridos. Nael Al-Bahisi, sanitario, fue uno de los primeros en llegar. Contó entre 50 y 70 víctimas ese primer día:

En un momento dado, entendimos que el objetivo del ataque era el comandante de Hamás Ayman Nofal [...]. Lo mataron a él y también a mucha gente que no sabía que Ayman Nofal estaba allí. Mataron a familias enteras con niños.

El ejército israelí destruyó un edificio de gran altura en Rafah a mediados de diciembre, en el que mató a «docenas de civiles» para intentar matar a Mohammed Shabaneh, el comandante de la Brigada Rafah de Hamás (no está claro si murió o no en el ataque). A menudo, según la fuente israelí, los altos mandos se esconden en túneles que pasan debajo de edificios civiles, por lo que la decisión de asesinarlos con un ataque aéreo necesariamente mata a civiles:

Había una política absolutamente permisiva respecto a las víctimas de las operaciones [de bombardeo], tan permisiva que en mi opinión tenía un punto de venganza [...]. El objetivo era el asesinato de altos mandos [de Hamás y de la Yihad Islámica en Palestina] por los que estaban dispuestos a matar a cientos de civiles. Hicimos un cálculo: cuántos por comandante de brigada, cuántos por comandante de batallón, etcétera. (D., fuente de los servicios de inteligencia).

Había normas, pero eran muy tolerantes [...]. Hemos matado a personas con daños colaterales de dos dígitos, si no de tres dígitos. Cosas que no habían sucedido antes. (E., fuente de los servicios de inteligencia)⁶

Todas las fuentes entrevistadas para esta investigación afirmaron que las masacres de Hamás del 7 de octubre y el secuestro de rehenes influyeron mucho en la política para abrir fuego del ejército y en los límites de daños colaterales. «Al principio, el ambiente era desgarrador y vengativo», dijo B., que fue reclutado por el ejército inmediatamente después del 7 de octubre y sirvió en una sala de operaciones de objetivos:

Las reglas eran muy permisivas. Derribaron cuatro edificios cuando sabían que el objetivo estaba en uno de ellos. Era una locura [...]. Había una discordancia: por una parte, la gente estaba frustrada

⁶ Un índice tan alto de «daños colaterales» es excepcional no solo en comparación con lo que el ejército israelí consideraba aceptable anteriormente, sino también en comparación con las guerras emprendidas por Estados Unidos en Irak, Siria y Afganistán.

porque no atacábamos lo suficiente [...]. Por otra parte, al final del día veías que otros mil habitantes de Gaza habían muerto, la mayoría de ellos civiles.

«En las filas profesionales había histeria», dijo D., quien también fue reclutado inmediatamente después del 7 de octubre. «No tenían ni idea de cómo reaccionar. Lo único que sabían hacer era empezar a bombardear como locos para intentar desarticular el potencial de Hamás.» D. subrayó que no se les dijo explícitamente que el objetivo del ejército era la «venganza», pero expresó que «en el momento en que cada objetivo relacionado con Hamás se legitima y casi cualquier daño colateral se aprueba, tienes claro que se van a matar miles de personas. Aunque oficialmente todos los objetivos estén conectados con Hamás, cuando la política es tan permisiva, pierde todo sentido.» A. también utilizó la palabra «venganza» para describir el ambiente dentro del ejército después del 7 de octubre:

Nadie pensó en lo que habría que hacer después, cuando termine la guerra, ni en cómo sería posible vivir en Gaza y qué harían con ello [...] Nos dijeron: ahora tenemos que joder a Hamás, cueste lo que cueste. Siempre que podáis, bombardead.

B., la principal fuente de los servicios de inteligencia, dijo que, en retrospectiva, cree que esta política «desproporcionada» de matar a palestinos en Gaza también pone en peligro a los israelíes, y que esta fue una de las razones por las que decidió dar la entrevista:

A corto plazo estamos más seguros porque perjudicamos a Hamás. Pero creo que, a largo plazo, estamos menos seguros. Veo que en las familias desconsoladas de Gaza –que son casi todas– aumentarán los motivos para unirse a Hamás dentro de 10 años. Y será mucho más fácil [para Hamás] reclutarlos.

El ejército israelí negó gran parte de lo que nos dijeron las fuentes y afirmaron que «cada objetivo se examina individualmente, al tiempo que se hace una evaluación individual de la ventaja militar y daños colaterales esperados del ataque [...]. Las FDI no llevan a cabo ataques cuando los daños colaterales que se esperan del ataque son excesivos en relación con la ventaja militar.»

Paso 5: cálculo de los daños colaterales

Según fuentes de los servicios de inteligencia, el cálculo por parte del ejército israelí del número de civiles que se esperaba que murieran en cada casa junto a un objetivo, se llevó a cabo con la ayuda de herramientas automáticas e imprecisas. En guerras anteriores, el personal de los servicios de inteligencia dedicaba mucho tiempo a verificar cuántas personas había en una casa que iba a ser bombardeada y el número de civiles que podían morir figuraba como parte de un «archivo de objetivos». Después del 7 de octubre, esta verificación exhaustiva en gran medida se abandonó en favor de la automatización.

En octubre, *The New York Times* informó sobre un sistema operado desde una base especial en el sur de Israel, que recopila información de los teléfonos móviles en la Franja de Gaza y que proporcionó a los militares una estimación en tiempo real del número de palestinos que huyeron del norte de la Franja de Gaza hacia el sur. El general Udi Ben Muha declaró al *Times* que «no es un sistema 100% perfecto, pero te da la información que necesitas para tomar una decisión». El sistema funciona por colores: el rojo marca las zonas donde hay mucha gente, y el verde y el amarillo marcan las zonas que han quedado relativamente despejadas de residentes.

Las fuentes describieron un sistema similar para calcular los daños colaterales, que se utilizó para decidir si bombardear un edificio en Gaza. Dijeron que el *software* calculaba el número de civiles que residían en cada casa antes de la guerra y luego redujo esos números en la proporción de residentes que supuestamente evacuaron el vecindario. Por ejemplo, si el ejército calculaba que la mitad de los residentes de un barrio se habían ido, el programa contaría una casa que normalmente tenía 10 residentes como una casa que contenía cinco personas. Para ahorrar tiempo, el ejército no vigiló las casas para comprobar cuántas personas vivían allí realmente, como hizo en operaciones anteriores, para comprobar si las estimaciones del programa eran exactas. Según una fuente:

Este modelo no se correspondía con la realidad [...]. No había ninguna conexión entre los que estaban en la casa ahora, durante la guerra, y los que figuraban como que vivían allí antes de la guerra. [En una ocasión] bombardeamos una casa sin saber que dentro había varias familias escondidas.

La fuente dijo que, aunque el ejército sabía que tales errores podían ocurrir, de todos modos, se adoptó este modelo impreciso porque era más rápido. Por lo tanto, «el cálculo de los daños colaterales era totalmente automático y estadístico» e, incluso, generó cifras que no eran números enteros.

Paso 6: bombardeo de una vivienda familiar

Las fuentes explicaron que a veces había una brecha sustancial entre el momento en que sistemas de rastreo como «¿Dónde está papá?» alertaba a un oficial de que un objetivo había entrado en su casa, y del bombardeo en sí, lo que, finalmente, provocó la muerte de familias enteras, incluso sin alcanzar el objetivo del ejército. «En muchas ocasiones atacamos una casa, pero la persona ni siquiera estaba allí [...]. El resultado es que matabas a una familia sin ningún motivo.»

Tres fuentes de los servicios de inteligencia dijeron que habían sido testigos de un incidente en el que el ejército israelí bombardeó la vivienda privada de una familia, y más tarde resultó que el objetivo del asesinato ni siquiera estaba dentro de la casa, ya que no se había llevado a cabo ninguna verificación adicional en tiempo real.

A veces [el objetivo] estaba en su casa antes, y luego, por la noche, se iba a dormir a otro lugar, por ejemplo, bajo tierra, y no lo sabíamos [...]. Había ocasiones en las que volvías a verificar la ubicación y otras en las que simplemente decías: 'Muy bien, como estaba en la casa en las últimas horas, simplemente puedes bombardear'.

Otra fuente describió un incidente similar que le afectó y por el que quiso participar en esta investigación:

Entendimos que el objetivo estaba en casa a las ocho de la noche. Al final, las fuerzas aéreas bombardearon la casa a las tres de la madrugada. Luego descubrimos [en ese lapso] que había conseguido trasladarse a otra casa con su familia. En el edificio que bombardeamos había otras dos familias con niños.

En guerras anteriores en Gaza, después del asesinato de objetivos humanos, los servicios de inteligencia israelí llevaban a cabo procedimientos de evaluación de daños por bombas. Esto es una verificación rutinaria posterior al ataque para ver si el mando superior había muerto y cuántos civiles habían muerto con él. Esto implicó escuchar llamadas telefónicas de familiares que perdieron a sus seres queridos. Sin embargo, en la guerra actual, al menos en relación con los militantes de rango inferior marcados por IA, este procedimiento fue suprimido para ahorrar tiempo. Las fuentes dijeron que no sabían cuántos civiles murieron en cada ataque. En el caso de los presuntos operativos de rango bajo de Hamás y la Yihad Islámica marcados por IA, ni siquiera sabían si el propio objetivo había muerto.

No sabes exactamente a cuántos has matado ni a quién has matado [...]. Únicamente se sigue el procedimiento de evaluación de daños por bombas (BDA, por sus siglas en inglés) cuando se trata de mandos altos de Hamás. En el resto de los casos, no importa. Recibes un informe de las fuerzas aéreas sobre si el edificio ha sido bombardeado y ya está. No tienes ni idea del número de daños colaterales; inmediatamente después pasas al siguiente objetivo. El acento se ponía en crear tantos objetivos como fuera posible, lo más rápido posible.

Si bien el ejército israelí puede dejar atrás cada ataque sin detenerse en el número de víctimas, Amjad Al-Sheikh, el residente de Shuja'iya que perdió a 11 de sus familiares en el bombardeo del 2 de diciembre, dijo que él y sus vecinos todavía están buscando cadáveres:

Aún hay cadáveres bajo los escombros [...] Catorce edificios de viviendas fueron bombardeados con sus residentes dentro. Algunos de mis familiares y vecinos todavía están enterrados bajo los escombros.

Estrategia militar e inteligencia artificial: algunas consideraciones éticas

Roberto Cristián Urrutia¹

«Puede ser que no te interese la guerra, pero la guerra está interesada en ti» es una frase que suele atribuirse, sin que se encuentre en ninguno de sus escritos, a Leon Trotsky. Al parecer proviene del profesor emérito de ciencias sociales del Instituto de Estudios Avanzados de la Universidad de Princeton, Michael Walzer, quien en su libro original de 1977 *Just and Unjust Wars* parafrasea al revolucionario soviético (Walzer, 2006, p. 29). Como sea, la frase nos permite iniciar este ensayo en el ánimo de explorar un tópico muchas veces dejado de lado en el ámbito académico civil en Chile: el tema de la guerra.

En el ámbito académico civil occidental, la guerra como fenómeno político, social y cultural de estudio, ha sido paulatinamente apartada del ámbito académico formal por sus connotaciones negativas, dejando a las academias de formación de las Fuerzas Armadas como únicos lugares de estudio formal de este fenómeno. Y aunque las bases filosóficas, epistemológicas y éticas de la guerra son ampliamente estudiadas en prestigiosas universidades, especialmente en el mundo anglosajón, ellas no son de particular

¹ Ingeniero Civil en Obras Civiles de la Universidad Austral de Chile (UACH). Magíster en Historia Militar y Pensamiento Estratégico, Academia de Guerra del Ejército de Chile (ACAGUE). Miembro de la Academia de Historia Militar, Chile (AHM). Oficial de Reserva, Centro de Reservistas Escuela de Telecomunicaciones del Ejército de Chile (ESCTEL). Actualmente, se desempeña como Jefe de Operaciones de Nüyün_Tek SpA. (NYN), empresa chilena dedicada al diseño y fabricación de dispositivos de protección sísmica de estructuras. Correo electrónico: rurrutia@nuyuntek.com – robertourrutia1@gmail.com

interés desde la perspectiva de las humanidades en el ámbito latinoamericano en general, y chileno en particular, producto, posiblemente, de aprensiones tanto en el ámbito metodológico como político de su estudio (González y Llantén, 2020), sufriendo, como describió Michael Howard (Morgan-Owen y Finch, 2022), una «huida a los suburbios» (Howard, 2006; Bueno, 2021).

Sin embargo, la guerra entre Ucrania y Rusia, y las operaciones militares de Israel en Gaza nos muestran de forma trágica y lamentable, la recurrencia de este fenómeno en la historia humana y cómo, a pesar de que queramos limitar o derechamente evadir el estudio de estas temáticas, ellas terminan por reaparecer en el ámbito de las relaciones internacionales.

Por otra parte, la creciente preocupación por el uso de tecnologías basadas en Inteligencia Artificial (IA) en el ámbito militar, se ha convertido en un tema especialmente sensible.

El texto *Understanding Modern Warfare* indica que, lamentablemente, es poco probable que ignorar el fenómeno de la guerra haga que este desaparezca, y hacerlo fomenta la ignorancia sobre un tema que ha tenido, y sigue teniendo, un impacto importante en los asuntos humanos. Ya sea que uno desee evitar la guerra, mitigar su impacto o prepararse para llevarla a cabo de manera eficiente (cuestiones que no son excluyentes entre sí), es importante que se estudie de forma profunda y exponiendo las consideraciones éticas que en ella se ponen en juego. De hecho, se podría sugerir que en una democracia del siglo XXI es particularmente importante que el mayor número de personas comprenda la naturaleza de la guerra moderna, a fin de estar preparados para emitir juicios inteligentes sobre la forma en que sus propios gobiernos buscan emplear la fuerza militar (Jordan et al, 2016, p. 2).

Introducción

La Inteligencia Artificial (IA), término acuñado por John McCarthy en 1955, fue definida como «la ciencia y la ingeniería para

fabricar máquinas inteligentes». Muchas investigaciones indican que los humanos programan máquinas para que se comporten de manera inteligente, pero hoy en día hacemos hincapié en las máquinas que pueden aprender, al menos en cierto modo, como lo hacen los seres humanos (HAI, 2020).

La literatura surgida en los últimos años sobre el uso de IA en el dominio militar es numerosa, pero producida por y desde grandes potencias, o desde estados de nivel de poder medio. Chile, en su condición de pequeña potencia en el sistema internacional (*Small Power* o *Small State*, en la literatura especializada de las Relaciones Internacionales), debe tener presente esta condición en el contexto particular del estudio de la IA.

Carl Von Clausewitz en su tratado *De la guerra*, indica que «La guerra de una comunidad se origina de una situación política, estalla por un motivo político, y es por lo tanto, un acto político». En cambio, sus reflexiones y análisis asociados a la ética militar han sido de mucho menor alcance, aun cuando contempla importantes elementos de la tradición occidental referidos al legítimo uso de la fuerza y a la conducción de operaciones militares bajo las reglas de la guerra (Moliner, 2019).

Al tratarse de un acto político, la guerra es dirigida desde los más altos niveles estatales. En el caso de sociedades democráticas, son las más altas autoridades las que deben conducir la guerra en caso de conflicto. En Chile, son atribuciones exclusivas de la presidencia de la República el empleo de las Fuerzas Armadas y la conducción de la política exterior (Ortega, 2011, p. 69), siendo el deber de cada ciudadano el velar por el buen uso y control del instrumento militar.

El estudio de la guerra la ha dividido en «niveles» de conducción. Tal como expone Clausewitz, el nivel principal es el político, del cual se espera dé dirección a los niveles inferiores, el estratégico, operacional y táctico. Los dos primeros son niveles de «dirección» y los dos últimos de «planeamiento y ejecución» (De Vergara, 2012, p. 7). Esta clasificación suele variar entre países, pero mantiene una misma lógica de jerarquía, en que el nivel político

mantiene la supremacía. El nivel estratégico posee su propio campo intelectual: los «estudios estratégicos» (Baylis et al., 2019; Mahnken et al., 2014; Bueno, 2021).

En las democracias occidentales, la expansión de los estudios estratégicos ha sido posible gracias al aumento de académicos civiles especializados. Thomas C. Schelling (1960), estrategia nuclear que basó su estudio y comprensión del conflicto y la cooperación a través del análisis de la teoría de juegos, señaló que para la comunidad estratégica norteamericana, fue vital la existencia de una contraparte civil a la militar. En consecuencia, la construcción de un pilar civil de dicha comunidad es un elemento intrínseco y central en su evolución intelectual y, por lo tanto, clave para el fortalecimiento del poder y control civil del instrumento militar (Bueno, 2021, p. 19), un tema que a nuestro juicio debe ser fortalecido en Chile por medio de un modelo de relaciones cívico-militares, que contribuya a una sana intersección entre el Estado y la ciudadanía.

El concepto de estrategia militar

La etimología de la palabra «estrategia» se remonta al griego clásico. A lo largo de la Edad Media y en la Edad Moderna las referencias más relevantes tendieron a asociarla al «Arte de la Guerra». Comenzó a utilizarse en su significado actual en Gran Bretaña, Francia y Alemania durante la década de 1770, reflejando el optimismo ilustrado según el cual, la guerra, como el resto de las esferas de los asuntos humanos, podía beneficiarse del uso de la razón. Antes, los fines y los medios podían mezclarse en la mente del jefe militar, que era responsable tanto de la formulación como de la ejecución de la estrategia. Posteriormente, los gobiernos comenzaron a fijar los objetivos, esperando que los generales los consiguieran. Antes de la Primera Guerra Mundial, el término tenía un carácter específicamente militar. Sin embargo, no fue hasta el siglo XX que la estrategia adquirió los significados amplios que hoy se le atribuyen y que tienden a aplicarse, retrospectivamente, a los practicantes del pasado. Con el tiempo, el término se alejó de sus orígenes castrenses y observamos que hoy se aplica a todos los

campos del esfuerzo humano, desde los deportes hasta los negocios, por lo que se ha hecho necesario hablar de «estrategia militar» como una subcategoría (Freedman, 2016).

Lawrence Freedman definió elegantemente la estrategia militar como «el arte de crear poder». Llevada al ámbito militar, su definición sería «el arte de crear poder para obtener el máximo de objetivos políticos utilizando los medios militares disponibles». Arthur Lykke la definió como la suma de fines, formas y medios ($S=E+W+M$). Richard K. Betts como «un plan para utilizar medios militares para lograr fines políticos»; Colin S. Gray (2010) como «el puente que relaciona el poder militar con el propósito político»; Jordan et al. (2016) como el «proceso que convierte el poder militar en efecto político, proporciona propósito y significado a las actividades militares» y, recientemente, Carr (2024) como una forma de «resolución de problemas». En general, la estrategia militar se divide en cuatro componentes: fines (objetivos), modos (cursos de acción), medios (recursos) y riesgo. (Echevarría II, 2017, p. 1). Para Hew Strachan (2014), la estrategia tiene significado como «Dirección de la Guerra», buscando instrumentalizar la violencia. En este sentido, Freedman afirma que la estrategia es más que un plan para conectar fines y medios, sobre todo porque involucra a un antagonista, por lo cual debe ser flexible y fluida, reflejando objetivos que pueden cambiar a medida que se desarrolla el conflicto y cambian las circunstancias; asimismo, implica la toma de decisiones relacionadas con la violencia organizada en nombre de un grupo particular, cuya esencia es profundamente humana (Ayoub y Payne, 2016, p. 794). Cuando la estrategia se entiende en toda su amplitud, la transferibilidad del concepto puede relacionarse con la toma de decisiones estratégicas o con su implementación posterior (Urrutia, 2023, p. 94).

Inteligencia Artificial Limitada (ANI, Artificial Narrow Intelligence) y sus posibles conexiones con la estrategia militar

Menon (2023) se pregunta: ¿la Inteligencia Artificial cambiaría la base de la estrategia? La revista especializada *Military*

Strategy Magazine ha publicado artículos con preguntas del mismo tenor: ¿la inteligencia artificial general cambiará la naturaleza de la guerra? (Compagnoni, 2023) o ¿deberían los estrategas preocuparse por la filosofía de la inteligencia artificial? (Carchidi, 2023), indicativos de la atención que en este campo está alcanzando la irrupción de la IA.

La inteligencia humana se considera la capacidad de lograr objetivos a través de una combinación de habilidades analíticas, creativas y prácticas. Actualmente, los sistemas de IA que funcionan a nuestro alrededor son variedades de ANI. Se limitan a utilizar algoritmos para atender tareas singulares (reconocimiento de voz, clasificación de imágenes, etc.). La ANI, a través de aprendizaje reforzado, logra capacidades de nivel humano o superior en campos predeterminados, siendo eficaces para realizar tareas específicas, pero sin adaptarse de manera flexible para ejecutar tareas diferentes o modificar su funcionamiento para enfrentar un desafío nuevo (Winter-Levy y Trefethen, 2016, pp.105-111).

Varios ejemplos del uso de la ANI en el ámbito militar, con foco en el nivel táctico, han sido desarrollados. La velocidad de procesamiento de datos asociada con la ANI, su rápido procesamiento del ciclo OODA (Observar, Orientar, Decidir y Actuar), su increíble precisión y la capacidad de reemplazar a los humanos en el frente de guerra, beneficia las capacidades ofensivas (Horowitz, 2019), permitiendo a los tomadores de decisiones optar por objetivos políticos más audaces, lo que afecta a la estrategia.

Payne (2018) indica que la IA puede establecer paralelos e inferencias a partir de datos, posiblemente invisibles a los humanos. Estos conocimientos imprevistos y perceptivos de la ANI podrían afectar drásticamente la configuración de la estrategia. Siempre hay aspectos físicos y psicológicos en la elaboración de estrategias, de manera que quienes toman decisiones se ven afectados por factores como la carga cognitiva, las limitaciones de tiempo, el estrés y la fatiga (Ayoub y Payne, 2016). En este caso, la ANI puede proporcionar aportes decisivos a los tomadores de decisiones y, al mismo tiempo, evitar la heurística humana y los juicios sesgados.

También puede haber una tendencia a que los planificadores comiencen a considerar las observaciones generadas por IA como comparables, o incluso superiores, a las realizadas por humanos. Este sesgo de automatización o dependencia excesiva en la toma de decisiones militares probablemente fomentará circunstancias que conduzcan a la inestabilidad estratégica en ausencia de intuición, juicio y responsabilidad humana (Horowitz, 2019).

A diferencia de un juego con conjuntos de problemas bien definidos, Clausewitz (1989) afirmó que la guerra «es el reino de la incertidumbre», donde la ambigüedad de las situaciones crea la llamada «niebla de la guerra». Yarger (2006) calificó a los entornos estratégicos como sistemas «caóticos», sensibles a las variabilidades en las condiciones iniciales y sujetos a un comportamiento aparentemente aleatorio e impredecible del adversario, destacando la experiencia, anticipación y conocimientos flexibles y contextuales del elemento humano como factor de éxito detrás de la estrategia. La ANI puede proporcionar ventajas tácticas, pero se necesita genio humano para guiar a la IA a través del conocimiento peligroso e imperfecto disponible en entornos estratégicos.

La estrategia debe incorporar las «pasiones, valores y creencias humanas», que ciertamente no son cuantificables para la ANI (Clausewitz, 1989, pp. 134-135). Yarger (2006) también sostuvo que la estrategia es «esencialmente una empresa humana» y sugirió que las emociones, la ideología y la cultura influyen en los objetivos dinámicos impulsados por la política y en la estrategia diseñada para alcanzarlos. La IA podrá direccionarse a subobjetivos que faciliten el logro final de la estrategia.

La importancia táctica de la ANI se extenderá inexorablemente a la esfera estratégica, ya que perturbará el equilibrio de poder a favor de su poseedor. A medida que las naciones compiten por lograr la hegemonía en estas tecnologías, los gobiernos dependerán de empresas privadas, y muchas naciones utilizarán las mismas cadenas de suministro internacionales de IA, lo que podría generar conflictos de intereses competitivos y afectar el camino hacia la adquisición de medios militares.

Inteligencia Artificial General (AGI, Artificial General Intelligence) y sus posibles conexiones con la estrategia militar

AGI, al contrario de la ANI, representa una inteligencia amplia que podría, en teoría, alcanzar habilidades cognitivas dinámicas similares a las humanas, equipada para emplear de manera flexible su aprendizaje en múltiples esferas, cambiar de forma autónoma entre tareas y abordar desafíos complejos (Winter-Levy y Trefethen, 2016). Se espera que una vez desarrollada la AGI, se alcance un punto de «singularidad», lo que eventualmente resultará en una superinteligencia que superará la humana. Payne (2018) afirma que la IA inevitablemente afectará la naturaleza y la conducción de la guerra al alterar los factores detrás de la toma de decisiones humanas.

Investigadores con miradas optimistas respecto de la IA señalan el historial de éxitos limitados de esta como prueba de una mejora continua, lo que Erik Larson ha llamado la «tesis de la inevitabilidad», narrativa determinista que postula un salto fluido desde tecnologías ANI probadas hasta el punto de sistemas que pueden imitar e, incluso, superar el pensamiento humano.

Como la AGI aún no existe, la naturaleza especulativa en cualquier análisis que pretenda vincularla con la Estrategia es inevitable. Se prevé que la AGI sea un agente autónomo con capacidades de aprendizaje no supervisado por humanos, preparado para aplicarlo en diversos ámbitos, que podría ser capaz de convertirse en la primera entidad no biológica con pensamiento estratégico. Debido a que la acción cognitiva subyacente de la AGI es sustancialmente diferente a la humana (Payne, 2018, pp. 208-210), podría cambiar drásticamente la base de la Estrategia al producir acciones y esquemas hasta ahora impredecibles, aunque carecería de una evaluación «humana» de lo que sería «suficientemente bueno» en términos de resultados.

Sin embargo, los escenarios planteados para la AGI son limitados. Según Hunter y Bowen (2024), la razón por la que podemos estar seguros de que la fe en la maduración de la ANI hacia algo cercano a la AGI estaría «fuera de lugar», es debido

al problema del razonamiento inductivo. Como se trata de un problema epistemológico, y como la guerra no es lógicamente solucionable, no pueden mejorarse sistemas de pensamiento simplemente reuniendo más datos, computadoras más poderosas o convirtiendo más cosas en métricas cuantificables. En su forma más simple, la inducción significa «adquirir conocimiento a partir de la experiencia» y, específicamente, convertir esas experiencias u observaciones en afirmaciones generalizables sobre cómo funciona el mundo, obtenidas probabilísticamente. Sin embargo, todas las inferencias inductivas serían «sospechosas», porque un conjunto de observaciones que en algún momento parecieron indicar correlación, o incluso causa y efecto, no son en absoluto una «cadena de razonamiento», con conclusiones no necesariamente seguras, con reservas de que las generalizaciones basadas en la observación probablemente sean correctas, tal como planteó David Hume. Los acontecimientos del tipo «Cisne Negro», es decir, acontecimientos altamente improbables, pero cuya consumación tiene un enorme potencial disruptivo, son ejemplos que probarían que la inducción no puede predecirlo todo ni producir generalizaciones infalibles.

En este sentido, la estrategia requeriría lógica abductiva: la capacidad de pensar y tomar decisiones basadas en la presencia constante de incógnitas y cosas incognoscibles que tal vez nunca aparezcan en un conjunto de datos históricos o en experiencias pasadas. Por lo tanto, existirían dudas razonables para confiar en el potencial de la AGI en el campo de la estrategia en el estado actual de la investigación en esta área.

Consideraciones éticas en el uso de IA en la estrategia militar: de generales e ingenieros

La literatura concerniente a las consideraciones éticas es de gran profundidad tanto para la IA en general (Vieweg, 2021; Coeckelbergh, 2020), como para uso en el dominio militar en particular (Voenekey et al., 2022; French y Lindsay, 2022; Scholz y Galliot, 2020), teniendo para el caso de la estrategia militar su propio corpus ético (Lucas, 2020 y 2015, Moliner, 2019).

Dos grupos profesionales se interconectan entre estrategia militar e IA: el «generalato» como grupo de toma de decisiones a nivel estratégico y a los «ingenieros», representando a quienes desarrollan el avance de la IA.

En el caso del generalato, ¿dejaremos que las máquinas tomen las decisiones reemplazando a humanos a quienes delegamos el control del instrumento militar? Creemos que el uso de IA solo aumenta la importancia humana de supervisión tecnológica en el ámbito militar. La estrategia de Inteligencia Artificial de Defensa del Reino Unido de 2022, indica: «las máquinas son buenas para hacer las cosas bien, por ejemplo, procesar rápidamente grandes conjuntos de datos, mientras que las personas son buenas haciendo las cosas correctas, por ejemplo, evaluando información compleja, incompleta y que cambia rápidamente, guiadas por valores» (Hunter y Bowen, 2024, pp. 120-121).

Las afirmaciones que sugieren que la IA será capaz de igualar a un general humano en la toma de decisiones estratégicas, se basan en distinciones excesivamente rígidas entre niveles de la guerra. La «estrategia centauro», unión humanos/IA en la toma de decisiones, es una solución a las dificultades de la IA con la creatividad y el sentido común, mejorando la calidad de la toma de decisiones humanas con ideas y sugerencias potencialmente innovadoras, optimizando el papel del Oficial de Estado Mayor (Pontillo y Garay, 2000; Ortega, 2013).

La ANI actual, generalmente, recurre al aprendizaje profundo para «descubrir patrones y tendencias en los datos observados», siendo, por lo tanto, estadística y probabilística. Este enfoque se basa específicamente en la lógica inductiva que es, según Hunter y Bowen (2024), inherentemente defectuosa en decisiones estratégicas, que implican decidir con respecto a la lógica abductiva incluyendo la impredecibilidad de la guerra. Qué tan lejos pueda estar la IA de ser competente en esta área depende en gran medida de la conceptualización de la estrategia utilizada. Una definición estrecha e insuficiente de toma de decisiones estratégicas demasiado materialista, implicará, erróneamente, que una acción del adversario

produce consecuencias particulares, cognoscibles y calculables, a las que luego podemos responder con conocimiento (Hunter y Bowen, 2024, pp. 117-118).

Existen dudas razonables sobre la capacidad de generar, a pesar de los grandes avances técnicos, un sistema de mando con IA competente que pueda liderar o asesorar de manera confiable a los estrategas. Esto se debe a que la IA que existe y domina la literatura contemporánea de tipo ANI, de aprendizaje automático o reconocimiento de patrones, se basa en la lógica inductiva (decidir basándose en predicciones extraídas de la observación previa), mientras que el mando estratégico requiere lógica abductiva (decidir ante lo desconocido e incognoscible). El razonamiento abductivo es fundamental para la teoría estratégica, sin embargo, sus formas de aplicación y potencial integración en la IA enfocada a la guerra, no han sido suficientemente analizadas y explicadas. Las afirmaciones existentes de que asesores competentes basados en IA solo son lo bastantes persuasivos si se adopta una visión de la estrategia basada en juegos reglados (como el ajedrez o el go), que tienden a adoptar las explicaciones más optimistas de la IA en el ámbito militar en general, y estratégico en particular. (Hunter y Bowen, 2024: 119).

Aunque en teoría pudiera llegarse a tecnologías que iguallen o mejoren la toma de decisiones estratégicas, ¿dejaremos en manos de tecnologías basadas en la IA las decisiones finales de vida y muerte en la guerra? Por muchos fallos que los humanos tengan en la toma de decisiones, debemos recordar que la estrategia y los estrategas humanos se forman en un contexto institucional particular, y son, en un sentido sociológico, la expresión de dicha cultura, siendo por la tanto, una representación de la nación a la que sirven (Carter, 2015).

Un análisis de la ética de los ingenieros en la IA, profesión que está hoy a la vanguardia de su desarrollo, merece ciertos comentarios previos en sus conexiones con el dominio militar. Podemos partir por la separación hoy existente entre los ingenieros «civiles» y «militares». El antecedente histórico de esta división se remonta a John Smeaton, quien fue el primer ingeniero en describirse como «civil», siendo clave, además, para establecer

en 1771 la Sociedad de Ingenieros Civiles de Inglaterra (Petroski, 2002). Sin embargo, es importante recalcar que, desde sus inicios, ambas ingenierías han estado íntimamente vinculadas (Fernández, 2001). Además, si seguimos la etimología de la palabra ingeniero y su conexión con «ingenio», podemos usar una de las acepciones de la RAE que la define como «máquina o artificio de guerra para atacar y defenderse». El ensayo «Ingenio e Ingeniería» indica que algunos anglófilos han buscado el antepasado del ingeniero en la voz inglesa «engineer» y, desde allí, en «engine», es decir, máquina. Sin embargo, también «engine» proviene del latín *ingenium* que llegó a las islas británicas durante el siglo XII, a través del francés *engin*, es decir, habilidad, ingenio, inteligencia; pero también, truco, engaño, máquina de guerra, significación que es posible rastrear en el latín tardío, tal como muestran Isidoro de Sevilla o Tertuliano al referirse al *ingenium* como «máquina de guerra, ariete» (Torres, 2018, p. 52), volviendo a conectar con la raíz militar de la ingeniería como profesión. Asimismo, y para volver a una visión más generalizante de la profesión que abarque el ámbito civil y militar, aún antes de su nacimiento, existen antecedentes de que el término ingeniero se utilizaba para el técnico que, sin provenir de un campo definido, «generaba soluciones» (Pastrana-Salcedo, 2014, p. 13).

Y es en esta última vertiente, la del ingeniero como generador de soluciones, donde la ingeniería en la IA puede ser asociada a la estrategia. Andrés Ortega Klein indica que la ingeniería de software (gran parte de la IA), no se ha reconocido aún como una profesión con deberes fiduciarios respecto del público. Al hablar de los «ingenieros éticos», Ortega Klein cita el libro de Robert McGinn *The Ethical Engineer. Contemporary Concepts & Cases*, observando una desconexión entre la formación ética de los estudiantes de ingeniería (aspecto teórico) y la realidad del trabajo ingenieril que cada día, gracias a la tecnología, es más ubicuo (aspecto práctico). Según este autor, los ingenieros deben asumir como parte de su profesión un único deber fundamental: combatir el daño. Para ello, plantea cuatro responsabilidades éticas o FERE (Fundamental Ethical Responsibility of Engineers), de acuerdo con las cuales, el ingeniero sería *empleado* por la sociedad que le permite practicar su profesión (Ortega Klein, 2020, pp. 18-19). Sin embargo, es bastante claro que

llevar a la práctica y compatibilizar estas responsabilidades supone un desafío mayor en el ámbito castrense. Si nos planteamos, por ejemplo, que el interés legítimo de una sociedad en un momento dado sea «ganar una guerra», ya sea porque se considera una «guerra justa», o porque la sobrevivencia del propio Estado que da sostén al trabajo de los ingenieros se hace primordial en un momento dado, o simplemente porque un Estado y sus burocracias en tiempo de paz pretendan mantener una superioridad tecnológica que les permita disuadir a sus potenciales adversarios de cualquier intento futuro de dañarla, entonces podemos llegar a preguntarnos: ese potencial daño que agentes externos puedan realizar, ¿puede también hacer que «dañar a los demás» se justifique, en defensa propia o del grupo de pertenencia, infringiendo estas responsabilidades éticas? Importante es realizar estas reflexiones en tiempos de paz, y no en las caóticas circunstancias de la guerra.

Si nos enfocamos en ámbitos más específicos que tocan la ética y la estrategia, como el de los ingenieros que desarrollarán la IA para «la toma de decisiones» existen varios prismas de análisis. Uno de ellos es la mirada desde un punto de vista basado en la complejidad del fenómeno de la guerra y la importancia de realizar análisis que consideren su carácter de sistema adaptativo complejo (Solvit, 2011; Ryan, 2011). En primer término, está el peligro de la simplificación del fenómeno de la guerra como un hecho que puede ser reducido solo a variables matemáticas que no tengan en cuenta los aspectos humanos y psicológicos. Concebir la guerra como una especie de juego o sistema con una cantidad finita de variables, reduce la estrategia y la política a una versión de ellas centrada en la batalla (o directamente centrada en matar), a-geográfica, a-histórica, sin tomar en cuenta las fricciones propias del fenómeno social, considerando equivocadamente sus variables como cuantificables, repetibles, no paradójicas y computables, adoptando una visión falsa de la guerra.

Algunos autores también indican la brecha existente entre la educación y la experiencia de los ingenieros al poner en prácticas sus habilidades en el mundo del trabajo. Los ingenieros rara vez consideran la posibilidad de que algunos episodios dañinos de ingeniería puedan ser atribuidos a una conducta éticamente problemática por parte de

los participantes. La práctica contemporánea de la ingeniería no solo plantea muchos desafíos éticos, sino que los ingenieros rara vez están preparados adecuadamente para abordarlos de manera reflexiva. Por lo tanto, hay varias razones por las que es importante que los ingenieros tengan una comprensión clara de la ética de su campo, especialmente para aquellos que hoy están a la vanguardia del cambio tecnológico de la IA, ya que «el poder de cambiar el mundo viene con obligaciones éticas» (Peterson, 2020, p. 3). Una de las razones por las que los ingenieros deben recibir esta capacitación ética es que muchas de las decisiones que deben y deberán tomar en el futuro no están gobernadas por reglas y, por lo tanto, estas quedan libres al criterio de cada cual (Schlossberger, 2023).

Dos aplicaciones de la IA en el ámbito de la estrategia militar que creemos es posible comenzar a explorar a partir del conocimiento de estas consideraciones éticas, teniéndolas en cuenta en lo que respecta a sus posibles aplicaciones en Chile, son la «Toma de Decisiones Militares» basadas en IA (Zai, 2020; Sarcia y Colo, 2023; French y Lindsay, 2022; Kase et al., 2022; Dear, 2019; Kerbusch et al., 2018; Rasch et al., 2003), y su aplicación en los «Juegos de Guerra» (Van Creveld, 2013; Mason, 2018; Barzashka, 2023; Knack y Powell, 2023; Hinton, 2023; Davis y Bracken, 2022; Perla, 2022; Black y Darken, 2024; Van Oijen y De Marez Oyens, 2023; Schwartz et al., 2020; Jensen et al., 2018), tema especialmente importante ya que, tal como Basil Liddell Hart (2012) puntualizó en uno de sus trabajos, a diferencia de quienes ejercen otras profesiones, el soldado regular no puede ejercer «regularmente» su profesión, considerando, además, que la literatura en estos temas se enfoca en la aplicación de la toma de decisiones del nivel estratégico militar de grandes potencias (Edström y Westberg, 2023), que se diferencia profundamente de los tipos de estrategias militares que pueden aplicar, debido a los recursos de que disponen y los fines que persiguen, poderes medios o pequeños en el sistema de relaciones internacionales (Edström y Westberg, 2022; Edström et al., 2020), pudiendo ser perfectamente adaptadas estas aplicaciones a países como Chile y sus realidades propias, tal como se hizo a inicios del siglo XX en el Ejército de Chile, y que este, a su vez, transmitió a otros países de Sudamérica (Arancibia, 2002; Wilson, 2023).

Ejemplos tomados de prestigiosas instituciones como el MIT SSP Wargaming Lab, que genera investigación sobre el uso de juegos de guerra mediante IA (Lin-Greenberg et al., 2022), considerando, además, que hoy es posible modelar también operaciones de paz (Mason y Patterson, 2013; Omar, 2023; Dorn et al., 2020), son perfectamente replicables en Chile.

Conclusiones

Colin S. Gray afirmó que la estrategia es una actividad humana permanente. Debido a la improbabilidad actual de programar facultades humanas como la emoción, la intuición y la conciencia, la ANI no alteraría a corto plazo fundamentalmente la naturaleza de la aquella, incluso, si sus capacidades influyen en la inteligencia humana a partir de su uso en la toma de decisiones estratégicas. En el futuro inmediato, con las actuales tecnologías disponibles, en su base, la estrategia seguirá siendo un fenómeno esencialmente humano debido a su capacidad de comprender un contexto más amplio y adaptarse a situaciones novedosas, utilizando ANI para tareas específicas por sus ventajas en la velocidad de procesamiento de datos. Un proceso de generación de AGI con pensamiento propio es, por ahora, solo una construcción teórica.

En los últimos años hemos sido testigos de una *militarización* de la IA, especialmente en lo que respecta a las grandes potencias (Hynek y Solovyeva, 2022), tal como sucedió en la Guerra Fría, o en otras épocas con otras tecnologías (uso de pólvora, artillería, etc.). Sin embargo, la naturaleza de la IA y su potencial de reemplazar al humano en tareas de nivel cognitivo superior, suponen un desafío no visto anteriormente.

Sin embargo, debemos recordar que Clausewitz advirtió que la guerra en el «mundo real» compuesto por humanos y caos no puede reducirse a «una especie de álgebra de la acción». Si pudieran conocerse todas las variables y resultados, y si la guerra fuera un asunto puramente racional, «no habría necesidad de la existencia física de ejércitos, sino solo de las relaciones teóricas entre ellos».

A medida que la IA irrumpió en la escena pública, algunos investigadores, preocupados por sus peligros, pidieron una pausa en su desarrollo. Pero detener el progreso de la IA es imposible: los fundamentos matemáticos de la IA son omnipresentes, las habilidades humanas para crear modelos de IA han proliferado ampliamente y los impulsores de la investigación y el desarrollo de la IA (tanto la creatividad humana como el beneficio comercial) son muy poderosos (Flournoy, 2023).

En un escenario con infinitas posibilidades como la guerra, la IA como «tomador de decisiones» se encontrará, en el peor de los casos, con cálculos indecibles y colapsará en un ciclo analítico sin fin, o en una parálisis de análisis, sin importar la potencia de cálculo asignada. Por otro lado, la noción de que existen opciones estratégicas objetivamente «correctas» que se llevan a cabo en un campo de batalla parece particularmente pretenciosa y errada, a la luz de un registro histórico lleno de ejemplos de rutas inesperadas o no centradas en acciones tácticas o estratégicas que llevaron tanto hacia la derrota como a la victoria. Esta dificultad de ejercer el mando y dirección del conflicto bélico en el mundo real es resultado de la naturaleza política, incierta y caótica de la guerra.

Incluso después de un resultado militar decisivo, el resultado de la guerra no es absoluto, especialmente en términos políticos. La paz puede ser simplemente un mal transitorio que un grupo organizado puede aceptar hasta que aparezca una próxima oportunidad de utilizar la violencia política en un momento posterior ¿Qué acciones es necesario cambiar ahora para evitar una reanudación mayor y más sangrienta de las hostilidades en el futuro? ¿Un ataque preventivo producirá una paz duradera? Estas son las preguntas de carácter complejo que surgen de la declaración más simple, más citada, pero menos analizada de Clausewitz: que la guerra es simplemente la continuación de la política por otros medios.

En este sentido, los caminos hacia la victoria son infinitos y a menudo confusos, por lo que las preguntas que surgen son hasta qué punto la IA ayudará o perjudicará a los tomadores de decisiones de la guerra, ya que como afirma el título de un artículo de la influyente revista *Foreign Affairs*, «La IA ya está en guerra» (Flournoy, 2023).

Referencias bibliográficas

- Arancibia, R. (2002). *La influencia del Ejército Chileno en América Latina 1900- 1950*. CESIM. Chile.
- Ayoub, K.; Payne, K. (2016). *Strategy in the Age of Artificial Intelligence*. Journal of Strategic Studies 39 (5-6): 793-819
- Barzashka, I. (2023). *Seeking Strategic Advantage: The Potential of Combining Artificial Intelligence and Human-centred Wargaming*. RUSI Journal 168 (7): 26-32.
- Baylis, J.; Wirtz, J.J.; Gray, C.S. Ed. (2019). *Strategy in the Contemporary World. An Introduction to Strategic Studies*. Sixth Edition. Oxford University Press. UK.
- Black, S., & Darken, C.J. (2024). *Scaling Artificial Intelligence for Digital Wargaming in Support of Decision-Making*. ArXiv, abs/2402.06075.
- Bolívar, C.; Ortega, R. (2016). *Estrategia Militar. Razonamiento Estratégico - Liderazgo - Escenario - Oficial de Estado Mayor*. ACAGUE. Chile.
- Bueno, A. (2021). *La Evolución de los Estudios Estratégicos en España. Defensa, Institucionalización, Cultura de Defensa y Comunidad Estratégica*. Tesis Doctoral. Universidad de Granada. España.
- Carchidi, V.J. (2023). *Should Strategists Worry About the Philosophy of Artificial Intelligence?* Military Strategy Magazine. 9 (1) 43-49
- Carr, A. (2024). *Strategy as Problem-Solving*. Parameters. 54 (1): 123-137
- Carter, O. (2015). *La Cultura Estratégica. Historia, Concepto y Definición*. Revista Ensayos Militares 1 (1): 131 - 147.
- Clausewitz, C.V. (1989). *On War*. Edited and Translated by Michael Howard and Peter Paret. Princeton University Press. USA.
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press. USA.

- Davis, P.K., Bracken, P. (2022). *Artificial intelligence for wargaming and modeling*. The Journal of Defense Modeling and Simulation 0 (0). doi:10.1177/15485129211073126
- Dear, K. (2019). *Artificial Intelligence and Decision-Making*. RUSI Journal. 164 (5-6): 18-25
- De Vergara, E. (2012). *La Semántica del Caos*. Revista Visión Conjunta 4 (6): 7-10
- Dorn, A. W., Webb, S., & Pâquet, S. (2020). *From Wargaming to Peacegaming: Digital Simulations with Peacekeeper Roles Needed*. International Peacekeeping. 27 (2): 289-310
- Echevarria II, A. J. (2017). *Military Strategy. A Very Short Introduction*. Oxford University Press. UK.
- Edström, H.; Gyllensporre, D.; Jacob Westberg, J. (2020). *Military Strategy of Small States. Responding to External Shocks of the 21st Century*. Routledge. UK.
- Edström, H.; Westberg, J. (2022). *Military Strategy of Middle Powers. Competing for Security, Influence, and Status in the 21st Century*. Routledge. UK.
- Edström, H.; Westberg, J. (2023). *Military Strategy of Great Power. Managing Power Asymmetry and Structural Change in the 21st Century*. Routledge. UK.
- Fernández, M. (2001). *Ingeniería Militar e Ingeniería Civil, Dos Ingenierías Íntimamente Vinculadas*. Revista de Obras Públicas 3413: 47-57
- French, S. E., & Lindsay, L. N. (2022). *Chapter 5. Artificial Intelligence in Military Decision-Making: Avoiding Ethical and Strategic Perils with an Option-Generator Model*. En: *Emerging Military Technologies. Ethical and Legal Perspectives*. Leiden, The Netherlands.
- Flournoy, M.A. (2023). *AI Is Already at War. How Artificial Intelligence Will Transform the Military*. Foreign Affairs. 102 (6): 56-69.
- Freedman, L. (2016). *Estrategia. Una Historia*. La Esfera de los Libros. España.

- Goldfarb, A.; Lindsay J.R. (2022). *Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War*. *International Security* 46 (3): 7–50
- González, C., Llantén, N. (2020). *La Academia Chilena y el Fenómeno de la Guerra: Aprensiones y Nuevos Horizontes Sobre una Temática Controversial*. *Cuadernos de Marte*. 18: 511-546.
- Gray, C.S. (2010). *The Strategy Bridge: Theory for Practice*. Oxford University Press. UK.
- HAI (2022). *Artificial Intelligence Definitions*. Text by Professor Christopher Manning, v 1.2, April 2022. Stanford Institute for Human-Centered Artificial Intelligence. HAI. Stanford University. Stanford, California. USA.
- Hinton, P. (2023). *Generative AI and Wargaming: What is it Good For?* *RUSI Journal* 168 (7): 34-41.
- Hodgins, J. (2022). *Would artificial intelligence make strategy ‘less human’?* *The Journal of Intelligence, Conflict, and Warfare*. 5 (1): 75–84.
- Hom, A.R.; O’Driscoll, C.; Mills, K. Ed. (2017). *Moral Victories. The Ethics of Winning Wars*. Oxford University Press. UK.
- Horowitz, M.C. (2019). *When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability*. *Journal of Strategic Studies* 42 (6): 764–788
- Hunter, C.; Bowen, B.E. (2024). *We’ll never have a model of an AI major-general: Artificial Intelligence, command decisions, and kitsch visions of war*. *Journal of Strategic Studies*. 47 (1): 116–146
- Hynek, N.; Solovyeva, A. (2022). *Militarizing Artificial Intelligence. Theory, Technology, and Regulation*. Routledge. UK.
- Jensen, B; Cuomo, S.; Whyte, C. (2018). *Wargaming with Athena: How to Make Militaries Smarter, Faster, and More Efficient with Artificial Intelligence*. *War on the Rocks*. <https://warontherocks.com/2018/06/wargaming-with-athena-how-to-make-militaries-smarter-faster-and-more-efficient-with-artificial-intelligence/>

- Johnson, J. (2022a). *The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-enabled Warfare*. *Journal of Military Ethics*. 21(3-4): 246-271
- Jordan, D.; Kiras, J. D.; Lonsdale, D.J.; Speller, I.; Tuck, C.; Walto, C.D. (2016). *Understanding Modern Warfare*. Cambridge University Press. UK.
- Kase SE.; Hung C.P.; Krayzman, T.; Hare, J.Z.; Rinderspacher, B.C.; Su, S.M. (2022). *The Future of Collaborative Human-Artificial Intelligence Decision-Making for Mission Planning*. *Front. Psychol.* 13:850628. doi: 10.3389/fpsyg.2022.850628
- Kerbusch, P.; Keijser, B.; Smit, S. (2018). *Roles of AI and Simulation for Military Decision Making*. En: *STO Meeting Proceedings MP-IST-160*.
- Knack, A.; Powell, R. (2023). *Artificial Intelligence in Wargaming: An evidence-based assessment of AI applications*. The Alan Turing Institute. Centre for Emerging Technology and Security CETaS Research Reports. UK.
- Liddell Hart, B. (2012). *Why Don't We Learn from History?* Sophron Editor. ISBN 10: 0985081139
- Lin-Greenberg, E.; Pauly, R. B. C.; Schneider, J. G. (2022). *Wargaming for International Relations research*. *European Journal of International Relations* 28 (1): 83-109
- Lucas, G. Ed. (2015). *Routledge Handbook of Military Ethics*. Routledge. UK.
- Lucas, G. (2020). *Ethics and Military Strategy in the 21st Century. Moving Beyond Clausewitz*. Routledge. UK.
- Mahnken, T.G.; Maiolo, J.A. Ed. (2014). *Strategic Studies: A Reader*. Second Edition. Routledge. UK.
- Mason, R.; Patterson, E. (2013). *War Gaming Peace Operations*. *Simulation & Gaming*. 44 (1): 118-133
- Mason, R.C. (2018). *Wargaming: its history and future*. *The International Journal of Intelligence, Security, and Public Affairs*, 20 (2): 77-101

- McGinn, R. (2018), *The Ethical Engineer. Contemporary Concepts & Cases*. Princeton University Press. USA.
- Menon, N (2023). *Would Artificial Intelligence change the basis of strategy?* IAF International Affairs Forum. 08/09/2023 <https://www.ia-forum.or/Content/ViewInternalDocument.cfm?ContentID=9643>
- Moliner, J.A. (2019). *Clausewitz y la Ética Militar*. RESI 5 (1): 141-152
- Morgan-Owen, D.; Finch, M. (2022). *The Unrepentant Historian: Sir Michael Howard and the Birth of War Studies*. *British Journal for Military History*. 8 (2): 55-76
- Omar, M. (2023). *The Application of Artificial Intelligence (AI) in Peacekeeping Operation*. *Journal of Defence and Security* 8 (1): 81-94
- Ortega Klein, A. (2020). *Geopolítica de la Ética en Inteligencia Artificial*. Documento de Trabajo 1/2020, 9 de enero de 2020. Real Instituto Elcano. Madrid. España.
- Ortega, R. (2011). *Crisis Internacionales*. ANEPE. Chile.
- Ortega, R. (2013). *Fisonomía del Pensamiento de los Oficiales de Estado Mayor del Ejército de Chile*. *Memorial del Ejército de Chile* 491: 115-125
- Pastrana-Salcedo, T. (2014). *Ingeniería militar, origen de la Ingeniería contemporánea*. *Revista Esencia y Espacio*. 38: 6-13.
- Payne, K. (2018a). *Strategy, Evolution, and War: From Apes to Artificial Intelligence*. Georgetown University Press. USA.
- Payne, K. (2018b). *Artificial Intelligence: A Revolution in Strategic Affairs?* *Survival*. 60 (5): 7-32
- Perla, P. (2022). *Wargaming and The Cycle of Research and Learning*. *Scandinavian Journal of Military Studies*. 5 (1): 197-208
- Petroski, H. (2002). *El Ingeniero Civil*. *Revista Ingenierías*. 5 (17): 7-13
- Pontillo, J. M.; Garay, C. (2000). *Pensamiento de Estado Mayor en Chile; una visión retrospectiva*. *Memorial del Ejército de Chile* 464: 121-140

- Prince, S. (2008). *Los Fundamentos de las Ciencias Militares*. Revista Política y Estrategia 111: 14-28
- Rasch, R.A.; Kott, A.S.; Forbus, K.D. (2003). *Incorporating AI into Military Decision Making: An Experiment*. IEEE Intell. Syst. 18: 18-26.
- Ryan, A.J. (2011). *Military Applications of Complex Systems*. En: Hooker, C. Ed. (2011). *Handbook of the Philosophy of Science*. Volume 10. *Philosophy of Complex Systems*. Part X: *Military Science*. Elsevier B.V. Oxford. UK.
- Sarcia, S.A.; Colo, G. (2023). *Organizing Structures and Information for Developing AI-enabled Military Decision-Making Systems*. 2023 IEEE International Workshop on Technologies for Defense and Security (TechDefense), Rome, Italy. Pág. 455-460.
- Schlossberger, E. (2023). *Ethical Engineering: A Practical Guide with Case Studies*. CRC Press. FL. USA.
- Scholz, J.; Galliot, J. (2020). *The Case for Ethical AI in the Military*. En: Dubber, M.D.; Pasquale, F.; Das, S. Ed. (2020). *The Oxford Handbook of Ethics of AI*. Oxford University Press. Pág. 685-702.
- Schwartz, P.J.; Daniel V. O'Neill, Meghan E. Bentz, Adam Brown, Brian S. Doyle, Olivia C. Liepa, Robert Lawrence, Richard D. Hull (2020). *AI-enabled wargaming in the military decision making process*. Proc. SPIE 11413, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, 114130H (22 April 2020)
- Solvit, S. (2011). *Dimensions of War: Understanding War as a Complex Adaptive System*. Thesis AUP - University of Paris I - Sorbonne. France.
- Strachan, H. (2014). *The Direction of War: Contemporary Strategy in Historical Perspective*. Cambridge University Press. UK.
- Torres, M. (2018). *Ingenio e Ingeniería*. Cuadernos de Beauchef. 1 (1): 47-58.

- Urrutia, R.C. (2023). *El Concepto de Estrategia Militar*. En: Urrutia, R.C. (2023). *Estrategia Militar y Geopolítica en el Memorial del Ejército de Chile entre 1945-1970. Un Aporte al Estudio del Pensamiento Militar en Chile desde la Historia Intelectual*. Tesis MHMPE. ACAGUE. Chile. Pág. 93-117.
- Van Creveld, M. (2013). *Wargames: From Gladiators to Gigabytes*. Cambridge University Press. NY. USA.
- Van Oijen, J.; De Marez Oyens, P. (2023). *Empowering Military Decision Support through the Synergy of AI and Simulation*. STO-MP-MSG-207.
- Vieweg, S.H. Ed. (2021). *AI for the Good: Artificial Intelligence and Ethics*. Springer Nature Switzerland AG.
- Voeneky, S.; Kellmeyer, P.; Mueller, O.; Burgard, W. Eds. (2022). *The Cambridge Handbook of Responsible Artificial Intelligence. Interdisciplinary Perspectives*. Cambridge University Press. UK.
- Walzer, M. (2006). *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. Basic Books. NY. USA.
- Wilson, P. H. (2023). *Iron and Blood. A Military History of the German-Speaking Peoples since 1500*. Harvard University Press. USA.
- Winter-Levy, S.; Trefethen, J. (2016). *Safety First: Entering the Age of Artificial Intelligence*. World Policy Journal 33 (1): 105–111.
- Yarger, H.R. (2006). *Strategic Theory for the 21st Century: The Little Book on Big Strategy*. USAWC. USA.
- Zai, M.M. (2020) *Artificial Intelligence: A Decisionmaking Technology*. Joint Force Quarterly JFQ 99 (4): 71-73

Consideraciones tecnoéticas del uso de inteligencia artificial generativa de imágenes en procesos de restitución de identidad de personas desaparecidas

Jorge Maldonado Soto¹

Introducción

A comienzos de 2023, en pleno auge y popularización de las herramientas generativas de imágenes basadas en inteligencia artificial, como *Midjourney* y *DALL-E*, el periódico español *La Vanguardia* publicó una nota utilizando esta tecnología en lo que denominaron «un pequeño experimento». El propósito de esta prueba fue ilustrar la posible apariencia que tendría Madeleine McCann a la edad de 21 años, la pequeña niña inglesa desaparecida en Praia da Luz, en Portugal, durante unas vacaciones junto a su familia en 2007. Para llevar a cabo este ejercicio utilizaron *Stable Diffusion*, otra de las populares herramientas basadas en inteligencia artificial para la generación de imágenes a partir de un texto. El artículo de prensa fue ampliamente difundido y las imágenes generadas fueron replicadas por diversos portales de noticias alrededor del mundo, revuelo exacerbado por la expectación del momento, ante la aparición en redes sociales de una joven polaca que afirmaba ser la niña desaparecida. A partir de los resultados de este ejercicio, el medio español comentó que: «Lo cierto es que las imágenes que ha creado la inteligencia artificial se parecen mucho a la chica que dice ser Madeleine», sugiriendo que el ejercicio aumentaba aún más las dudas acerca de si la joven era de la niña desaparecida (Martín, 2023). Sin embargo, algunas semanas más tarde, las pruebas de

¹ Ingeniero civil informático, Universidad Técnica Federico Santa María, y estudiante de Magíster en Ciencia, Tecnología y Sociedad, Universidad Alberto Hurtado, jmaldona89@gmail.com

ADN confirmarían, finalmente, que no se trataba de Madeleine McCann (Iporre, 2023).

Este y otros casos similares suscitan una vez más el debate en torno a la ética en relación con la tecnología, y más particularmente, con la aplicación de la inteligencia artificial y el aprendizaje automático para la generación de imágenes, en un ámbito donde diversos especialistas y opinólogos de la tecnología, han abordado con enorme entusiasmo sus posibles contribuciones a los procesos investigación y búsqueda de personas desaparecidas. Las experiencias en torno a este tema van desde iniciativas que, a partir de una fotografía facial predicen la forma en que eventualmente luciría el rostro de una persona con el paso de los años, a la combinación de dos o más imágenes faciales para la predicción de un nuevo rostro. Este tipo de tecnologías han sido ampliamente popularizadas desde hace algunos años a partir de la aparición de diversas aplicaciones para dispositivos móviles con fines lúdicos, como *FaceApp*, las que, además, se han caracterizado por ser motivo de públicos cuestionamientos y controversias en materias de seguridad y privacidad, una situación que ha llevado a los especialistas a recomendar a la comunidad de usuarios, asegurarse de no incluir ninguna información personal o confidencial en las fotos, imágenes y vídeos al utilizar estas aplicaciones (Neyaz et al., 2020). Y aunque a pesar de que el uso de la inteligencia artificial en este ámbito se ha ampliado y popularizado durante los últimos años, lo cierto es que las investigaciones en torno a este tema han sido desarrolladas desde hace mucho más tiempo, principalmente asociadas al trabajo en torno a modelos de progresión de la edad facial en el campo del arte forense.

Es así, como los resultados obtenidos por el periódico *La Vanguardia*, a partir de su «experimento» relacionado al caso de Madeleine McCann, junto con poner en discusión la ética periodística en torno al uso de inteligencia artificial generativa de imágenes en noticias (Thomson & Thomas, 2023; Paik et al., 2023), permite, también, abordar de manera empírica la pertinencia del uso de esta tecnología en los procesos de restitución de identidad de personas desaparecidas. La omisión de una revisión exhaustiva, cuidadosa

y ética de las consideraciones, limitantes, y en ocasiones, falta de rigurosidad científica en torno a los resultados obtenidos, reforzado además, por un contexto sociotécnico en el que la inteligencia artificial se ha posicionado como el símbolo del desarrollo de las nuevas tecnologías digitales y como un imaginario sociocultural de extraordinarios avances sobrehumanos, propios de la ciencia ficción, puede provocar el agravamiento de los impactos psicosociales que se podrían producir como resultado de las expectativas generadas, especialmente, entre familiares y cercanos a las personas víctimas de desaparición. No obstante, lo cierto es que, en ocasiones, y ante la falta de avances en los procesos de búsqueda, disponer de una imagen generada mediante inteligencia artificial que permita aportar a las investigaciones, a pesar de sus limitaciones y sesgos, puede ser considerado por algunos mejor que nada, e incluso, representar un símbolo de lucha en torno a la búsqueda de las familias en la era digital.

El presente artículo tiene por objetivo aportar a la reflexión de las consideraciones tecnoéticas del uso de la inteligencia artificial generativa de imágenes, en casos asociados a procesos de restitución de identidad de personas desaparecidas. Para esto, se revisarán algunas experiencias en el mundo en este campo realizadas con fines académicos, lúdicos, artísticos y/o simbólicos, también las principales limitaciones y complicaciones de estos métodos y cómo se relacionan con las principales discusiones y reflexiones en la actualidad, en torno a la ética del uso de la inteligencia artificial y su impacto en la sociedad. De este modo, por medio del presente análisis se busca identificar, explicar y visibilizar aquellos aspectos sociotécnicos a considerar al abordar estos procesos, comprendiendo claramente sus alcances para así, entre otras cosas, evitar agudizar los impactos psicosociales que se podrían llegar a producir, principalmente, entre familiares y cercanos a las personas víctimas de desaparición, producto de las expectativas generadas por los resultados de estas iniciativas. Asimismo, el análisis propuesto busca aportar desde la tecnología a los procesos de búsqueda de una manera ética, responsable y velando siempre por la dignidad y el bienestar de las personas.

Antecedentes

La progresión de la edad facial consiste en una serie de técnicas computarizadas, manuales o mixtas, que permiten proyectar los posibles cambios en el rostro de una persona con el paso del tiempo. Entre sus numerosas aplicaciones, estas técnicas suelen utilizarse comúnmente para la búsqueda de personas sospechosas de algún delito o para apoyar la búsqueda de personas desaparecidas. En este sentido, desde hace décadas el arte forense se ha dedicado a investigar en torno a esta área, considerando aspectos como el estilo de vida y la genética de las personas, el estudio de la morfogénesis craneofacial y el uso de fotografías familiares para la construcción de imágenes de nuevos rostros. De la mano del progreso de la computación y el desarrollo de las nuevas tecnologías digitales, estos métodos fueron avanzando en materias de automatización y su desarrollo se fue ampliando hacia nuevos campos de investigación computacional, entre ellos, el uso de técnicas basadas en inteligencia artificial. No obstante, a pesar del interés de diversos investigadores en este ámbito, se reconoce que estas técnicas no otorgan suficiente confiabilidad en sus resultados producto de la complejidad y enormes desafíos que presentan (Lampinen et al., 2011); incluso, se sabe que pueden llegar a ser perjudiciales para los procesos de búsqueda, al aumentar el número de objetivos plausibles, disminuyendo las probabilidades de identificar de forma correcta el objetivo y aumentando la posibilidad de «reconocer» por error a personas que no son el objetivo (Charman & Carol, 2012).

En 2017, utilizando como referencia el caso de Ben Needham, un niño inglés de 21 meses de edad desaparecido en la isla griega de Kos en 1991, un grupo de investigadores de la Universidad de Bradford en Inglaterra, desarrolló un método para envejecer imágenes faciales con el fin de aportar a la búsqueda de personas desaparecidas. Este método consiste en un modelo predictivo dedicado a entrenar un algoritmo a partir de una base de datos de rasgos faciales de individuos de distintas edades para enseñarle cómo envejecen los humanos y así poder generar imágenes del rostro a diferentes edades (Bukar & Ugail, 2017). Algunos años más tarde, en 2021, otro grupo de investigadores de la Universidad

Autónoma de Ciudad Juárez presentó un método de generación de imágenes faciales mediante la progresión de la edad facial de mujeres víctimas de violencia, desaparecidas y no localizadas en México. Para este modelo se utilizaron redes generativas antagónicas, con el objetivo de prever una eventual apariencia futura a partir de una fotografía original (Cazares et al., 2021). No obstante, a pesar de las potenciales ventajas que ofrecen estos y otros modelos similares, lo cierto es que no están exentos de importantes desafíos prácticos, tales como la calidad y los efectos de iluminación inherentes a la imagen original, así como los factores intrínsecos y extrínsecos que inciden en la representación facial (Sharma et al., 2021).

Un acercamiento más crítico al uso de esta tecnología en este ámbito, es el realizado por las Abuelas de Plaza de Mayo, organización argentina dedicada a la búsqueda y restitución de la identidad de sus nietas y nietos secuestrados y dados en adopción de forma ilegal durante la dictadura militar ocurrida en ese país entre 1976-1983, cuyas madres y padres fueron víctimas de desaparición forzada como resultado de las acciones de los organismos represivos del régimen encabezado por Jorge Rafael Videla. Las Abuelas son reconocidas como un referente científico-tecnológico en materias de derechos humanos en Argentina y el mundo, al ser las impulsoras durante los años ochenta del «índice de abuelidad», un indicador genético que permite identificar el parentesco entre una persona y su eventual abuela o abuelo. En otro ámbito, a mediados de 2023, una iniciativa llamada *IAbuelas*, proponía colaborar con la búsqueda de esta organización por medio de la generación de imágenes, utilizando para ello inteligencia artificial; la propuesta consistía en combinar, al menos, una imagen de cada progenitor y solicitar a la herramienta crear una nueva imagen de una persona de unos 45 años con un diseño hiperrealista. Aunque, tanto el creador de esta iniciativa, como los familiares de las víctimas, son conscientes de que esta iniciativa carece de validez científica en cuanto a sus resultados, no dejó de ser una experiencia que despertó profundamente el interés de algunos de los familiares (Rey, 2023). Y aunque las Abuelas agradecieron la iniciativa por acompañar su búsqueda, fueron enfáticas en declarar que «esta iniciativa no es científica sino artística-lúdica», y que, «sabemos que las personas son mucho más que el 50 por ciento de cada uno de sus padres» (Abuelas de Plaza de Mayo, 2023).

Lo cierto es que este tipo de ejercicio no es nuevo para las Abuelas. Durante una visita efectuada en Nueva York en 1982, en medio de una serie de reuniones con especialistas en genética de diversas partes del mundo, con el objetivo de impulsar su idea de identificar a sus nietos a través del análisis de sangre, sostuvieron un encuentro con una pintora. Esta artista visual había impulsado el desarrollo de la «máquina de la edad», un estudio que, a partir del análisis fotográfico, artístico y computacional, permitía vislumbrar el presunto aspecto facial de una persona años más tarde. En ese momento, las Abuelas exploraban, además, otras alternativas como la identificación morfológica, concibiendo la posibilidad de crear representaciones visuales de sus nietos para empapelar las calles del país con sus rostros. No obstante, pronto se dieron cuenta de que este método tenía limitaciones significativas, dado que la mayoría de sus nietos habían nacido en cautiverio, careciendo de fotografías, y las imágenes disponibles de los demás niños eran muy difusas (Wulff, 2008; Nosiglia, 2007). Por lo tanto, es posible suponer que esta experiencia permitió a las Abuelas comprender desde hace años las limitaciones que existen en torno a este método y sus diversas variantes.

Otra experiencia similar tiene su origen en la dictadura cívico-militar en Chile de 1973-1990. Más de 20 mil niños fueron secuestrados y entregados en adopción en el extranjero producto de una política sistémica del régimen de Augusto Pinochet que, por medio de un aparato internacional, buscaba regular, principalmente, la población pobre mediante mecanismos de control y regulación familiar (Alfaro, 2022). Ante estos crímenes, en 2024, la organización Hijos y Madres del Silencio, agrupación que entrega apoyo a víctimas de adopciones ilegales y tráfico de niños durante este período, inició una campaña por medio de la cual, utilizando fotografías familiares, recrearon a través del uso de la «genética digital» y la inteligencia artificial los rostros que tendrían en el presente los hijos secuestrados, con el objetivo de difundir los resultados y poder encontrarlos con sus respectivas familias de origen (Agencia EFE, 2024). La campaña explica que la imagen de estos hijos fue creada «desde cero», ya que «nunca hubo fotos»; de este modo, combinando imágenes de los respectivos progenitores recrearon cómo sería en la actualidad,

eventualmente, la apariencia de los niños desaparecidos, enseñando los resultados a través de un video en el que son presentados como *streamers* que se buscan a sí mismos, en los países en donde fueron vendidos (Hijos y Madres del Silencio, 2024).

Más allá de estas experiencias particulares, lo cierto es que la progresión de la edad facial ha sido un área de investigación desde hace varios años, incluso desde antes de la popularización de las herramientas generativas de imágenes, y más aún, desde antes de la computación, basándose, principalmente, en un ejercicio de carácter artístico dentro del campo de la ciencia forense (Taylor, 2001). Los avances computacionales en el campo de la inteligencia artificial, permitieron que estos métodos se fueran complejizando al incorporar una serie de cálculos que ayudan a estimar el eventual futuro rostro de una persona, implementando modelos que permiten mejorar los resultados gracias al entrenamiento que reciben a partir de fotografías disponibles en bancos de imágenes. Estos avances, han permitido, además, automatizar la producción gráfica, optimizar los tiempos en la generación de las imágenes, e incluso, últimamente, ampliar el alcance del público que puede impulsar este tipo de iniciativas más allá del campo científico-académico. No obstante, las actuales limitantes y complicaciones asociadas no suelen ser muy distintas a las de antaño, al tratarse de consideraciones intrínsecas al método, por lo que las técnicas computarizadas no logran, necesariamente, resolver varias de estas situaciones.

Discusión

Sin lugar a duda, la inteligencia artificial conlleva una serie de reflexiones y discusiones éticas, en especial, cuando el uso de este tipo de tecnologías se ha masificado en distintos ámbitos de la sociedad. En este sentido, durante los últimos años se han realizado una variedad de trabajos académicos, desarrollo de políticas públicas y diversas recomendaciones de organismos internacionales, en torno a principios y consideraciones para su uso y desarrollo. En noviembre de 2021, la Conferencia General de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO)

alcanzó el primer acuerdo mundial sobre la ética de la inteligencia artificial, el cual establece ciertos valores y principios para definir recomendaciones que regulen el desarrollo de la inteligencia artificial en diversos ámbitos de acción política, abordando aspectos como la proporcionalidad, no discriminación, seguridad, transparencia, protección de datos y sostenibilidad en su desarrollo. En Chile, la Política Nacional de Inteligencia Artificial, impulsada en 2021 y actualizada en 2024 por el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación, establece como uno de los principios transversales su uso ético y responsable, contribuyendo al bienestar integral de todas las personas, sin exclusión ni discriminaciones, y abordando los riesgos y potenciales impactos negativos, con respeto a los derechos y la dignidad de las personas.

En el ámbito de lo medial y, particularmente, en relación con la aplicación de la inteligencia artificial para la generación de imágenes, estas discusiones y reflexiones también han sido objeto de amplio estudio. A iniciativas como la generación de imágenes en procesos de restitución de identidad de personas desaparecidas, se suman, además, otros casos que han suscitado intensos debates de interés público en la materia. A principios de 2023 fueron viralizadas dos imágenes creadas mediante inteligencia artificial: una que mostraba al Papa Francisco luciendo una ostentosa chaqueta blanca y otra a Donald Trump siendo detenido por un grupo de policías; ambas imágenes fueron ampliamente difundidas como auténticas, incluso por algunos medios de comunicación, dejando de manifiesto las dificultades cada vez más crecientes para distinguir lo real de lo artificial (Rodríguez, 2023). A mediados de 2022, una imagen generada con *Midjourney* se alzó como ganadora de un concurso de arte celebrado en la Feria Estatal de Colorado, en Estados Unidos, suscitando un profundo debate en torno a las reglas del concurso, e incluso, a la definición misma de qué es lo que se entiende por arte (Espinosa, 2022). Así también, durante el último tiempo, se han producido importantes discusiones en torno al uso de obras con derechos de autor en el entrenamiento de estos modelos, situación ante lo cual, la Unión Europea ya ha planteado varias regulaciones que exigen a los desarrolladores de aplicaciones basadas en inteligencia artificial generativa, entre otras cosas, la publicación de resúmenes

de los datos sujetos a derechos de autor empleados en sus procesos de entrenamiento (Parlamento Europeo, 2023).

Por lo tanto, se reconoce ampliamente que este tipo de tecnologías puede ser potencialmente causante de riesgos sociales y éticos al producir contenidos dañinos, discriminatorios, inapropiados o fraudulentos. En este sentido, Bird et al. (2023) proponen una tipología de riesgos de los modelos generativos de texto a imagen, definiendo tres categorías claves: i) *discriminación y exclusión*, que considera aspectos que pueden llegar a tener estas tecnologías, tales como los sesgos culturales, raciales, de género, sexuales, etarios, entre otros, además de la pérdida de trabajo para creativos; ii) *uso indebido perjudicial*, como la creación de imágenes falsas o violentas para producir daño, además de la violación a la privacidad y a los derechos de autor; y, iii) *información errónea y desinformación*, como la generación de imágenes falsas para engañar, realizar fraudes o estafas, o bien, generar polarización e inestabilidad sociopolítica. Estas categorías no son excluyentes unas de otras, sin embargo, permiten tener una comprensión más clara y ordenada del fenómeno con el objetivo de implementar estrategias de mitigación más sólidas. Así, con esta tipología aplicada al uso de inteligencia artificial generativa de imágenes, en procesos de restitución de identidad de personas desaparecidas, es posible vislumbrar algunos elementos de interés en las tres categorías.

En la categoría de *discriminación y exclusión*, estudios en la materia reconocen que la mayoría de las bases de datos públicas de rostros con las que son entrenados estos modelos suelen estar sesgados con características faciales caucásicas, subrepresentando otro tipo de rostros como el de los latinos. En consecuencia, estos datos darán como resultado modelos sesgados, presentando problemas éticos sobre la imparcialidad de estos sistemas (Kärkkäinen & Joo, 2021). En este sentido, tal como ya se mencionó, las Abuelas de Plaza de Mayo de Argentina reconocen este aspecto en la iniciativa asociada a su causa, por lo que son enfáticas en declarar que «las aplicaciones extranjeras están seteadas con genotipos de sus poblaciones, por lo que los resultados no son certeros» (Abuelas de Plaza de Mayo, 2023). Otro aspecto en este ámbito, es que los procesos computacionales de

progresión de la edad facial pueden resultar ser más complejos cuando se trata de niños, esto, considerando que las investigaciones sobre este campo son limitadas en comparación con las investigaciones realizadas en adultos, a causa de la subrepresentación de los niños en los bancos de imágenes con los que suelen ser entrenadas estas tecnologías (Grimmer et al., 2021). De este modo, a la exclusión racial de estos modelos, se suma, además, un sesgo etario que puede afectar particularmente aquellos resultados obtenidos a través de estas tecnologías, en los procesos de búsqueda y restitución de identidad de niños desaparecidos.

En cuanto a la categoría sobre el *uso indebido perjudicial*, en general, estas experiencias no suelen presentar un ánimo intencional de producir daño, por el contrario, suelen tener como objetivo fundamental ser un aporte a los procesos de búsqueda. No obstante, como se verá luego con más detalles, no considerar ciertas precauciones y advertencias, e informar claramente a las familias de las personas desaparecidas acerca de las restricciones y limitaciones de estos modelos, puede producir ciertas consecuencias perjudiciales. Además, las iniciativas en la materia pueden utilizar alguna de las herramientas generativas de última generación disponibles, que se han caracterizado, en su mayoría, por la falta de transparencia en sus modelos y cuestionamientos por el uso de imágenes protegidas por derechos de autor, afectando los sistemas de recompensa existentes para artistas creativos y proveedores de contenidos (Zhou et al., 2024). Aunque, claro está, este último aspecto consiste en un problema general presente en una parte importante de los modelos generativos de texto a imagen, por lo que no es posible caracterizarlo como una consideración particular del ámbito de este análisis.

El aspecto que más puede afectar a los procesos de restitución de identidad de personas desaparecidas, está asociada a la categoría de *información errónea y desinformación*. Como ya se comentó, como resultado de diversos factores tecnológicos, morfológicos y mediales, las imágenes faciales generadas por estos modelos no representan fehacientemente la apariencia de la persona desaparecida; de este modo, cualquier parecido puede responder más a un asunto de azar que a la validez científica. No obstante, según el modelo que se

utilice, la diversidad de entrenamientos que reciba, la edad de la persona al momento de su desaparición y las características y calidad de las fotografías originales, se podrían obtener resultados más certeros. En este sentido, lo cierto es que puede parecer que estos resultados no sean tan sencillos de categorizar como *información errónea* y *desinformación*, considerando que los impulsores de estas iniciativas, aunque reconocen las limitaciones, deficiencias y sesgos de estas tecnologías, suelen no tener un ánimo de engañar o desinformar, sino más bien, de poder contribuir a las causas. De este modo, la discusión en torno a lo ético puede parecer compleja, pero no debería serlo, pues a pesar de las buenas intenciones, es necesario ser responsables y conscientes de que la generación de imágenes por medio de inteligencia artificial para determinar cómo luciría con el paso de los años una persona desaparecida, puede contribuir a la desinformación y a la difusión de información engañosa, tal como ocurrió con el caso de Madeleine McCann y el «experimento» del periódico *La Vanguardia*, entorpeciendo, incluso, como ya se ha comentado, los procesos de investigación y búsqueda, y generando expectativas familiares y sociales que más tarde no logran ser satisfechas.

Por último, en términos generales, un aspecto relevante por considerar en los casos de personas desaparecidas, y en especial, en víctimas de desaparición forzada, son los impactos psicosociales que pueden producir en sus familias y entornos cercanos. Estos impactos no son generados únicamente por la desaparición misma, también en actos que esta situación conlleva durante el proceso de búsqueda, como el señalamiento por parte de la comunidad, escenarios de estigmatización por parte de las instituciones y negligencias ante el suministro de información, respecto de las diligencias que deben llevar a cabo para exigir sus derechos (Romero & Cuellar, 2022). Es particularmente en este último aspecto relacionado con la información en torno a las diligencias del caso, que el uso de inteligencia artificial para la generación de imágenes de progresión de la edad facial, puede resultar conflictiva y éticamente cuestionable, si no se advierte anticipada y debidamente de las limitaciones y consideraciones de sus resultados. Es así, como minimizar u omitir los alcances de este método, puede generar expectativas y esperanzas

en los familiares, que más tarde, producto de la falta de resultados, o como ya se comentó, el aumento de falsos positivos en la búsqueda, se pueden traducir en emociones como la frustración y la angustia, agudizando los impactos psicosociales.

Conclusiones

Suele ser común que el rápido avance y la popularización de las nuevas tecnologías asociadas a herramientas basadas en inteligencia artificial, encuentre desprevenida a la escena social y académica, en relación con las consideraciones tecnoéticas que deben tenerse en cuenta frente a estos desarrollos. Las reflexiones sobre su impacto en la sociedad suelen llegar más tarde por medio de métodos de evaluación tecnológica, cuyo enfoque supone que las causas ya se han producido y, por tanto, se limitan a estudiar exclusivamente sus consecuencias. Este modelo de causa y efecto presenta defectos empíricos y morales, por lo que se hace necesario adoptar un nuevo paradigma prospectivo, en el que se reconozca de que, a medida que los aparatos tecnológicos se construyen y ponen en uso, ya producen alteraciones significativas en los patrones de la actividad y las instituciones humanas (Winner, 2008). De este modo, frente al actual escenario en el que en muchas oportunidades surgen lamentaciones sobre aquellas repercusiones no previstas como consecuencia del uso de estas tecnologías, se hace necesario anticiparse y poner atención en su diseño y desarrollo, con el objetivo de prever sus eventuales impactos y hacer los ajustes necesarios a tiempo. Esto, primeramente, en términos generales.

Más particularmente, el uso de estas herramientas tecnológicas en los procesos de restitución de identidad de personas desaparecidas, como ya se ha revisado, puede resultar éticamente conflictivo. Y es que a pesar de que estos métodos cuentan con una larga trayectoria de desarrollo e investigación, lo cierto es que una parte importante de los estudios en la materia, reconocen sus limitaciones y defectos, a tal punto de que, inclusive, puede afectar negativamente a los procesos de búsqueda. Entre los aspectos por considerar se encuentran las limitaciones y sesgos que poseen estas

tecnologías, la complejidad asociada a la multiplicidad de elementos que influyen en la progresión de la edad facial y las falencias inherentes relacionadas a la calidad, el ruido y la iluminación que poseen las fotografías utilizadas como fuentes. La alternativa frente a este escenario puede ser ofrecer un amplio abanico de posibilidades de cómo luciría el rostro, como suele ocurrir en el caso de la búsqueda de personas sospechosas de algún delito y que suelen cambiar su apariencia, pero como ya se ha visto, esto puede resultar contraproducente para el proceso de búsqueda al aumentar la cantidad de objetivos plausibles, y como consecuencia, disminuir las probabilidades de encontrar al objetivo correcto.

Y es que, a pesar de que varios de estos ejercicios e iniciativas pueden ser, válidamente, objeto de críticas, al ser desarrollados en contextos con fines artísticos, lúdicos y/o simbólicos –e incluso sensacionalistas, como puede ser en el caso de algunos medios de prensa–, los resultados obtenidos en ambientes de investigación académica tecnocientífica no suelen diferir en cuanto a ser cautelosos de sus resultados. Por esta razón, a pesar de las particularidades de los diversos casos, y el eventual aporte, e incluso, algún probable éxito de los procesos de búsqueda a partir del uso de estas tecnologías, se hace necesario establecer criterios generales de conocimientos en la materia que permitan, principalmente a familiares y cercanos de las víctimas de desaparición, alfabetizarse en la materia para comprender de manera clara y transparente las características, funcionamiento y alcances de estos métodos. En este sentido, debe ser responsabilidad de los desarrolladores de estas tecnologías e impulsores de estas iniciativas, explicar los detalles del proceso, transparentar los métodos y herramientas utilizadas, considerar la opinión de las personas involucradas y advertir de las limitaciones y consideraciones.

No obstante, lo cierto es que, en ocasiones, tanto los familiares y cercanos a las personas desaparecidas, como los organismos públicos y privados que participan de los procesos de búsqueda, pueden considerar el hecho de contar con una o más fotografías generadas a través de estos métodos mejor que nada, en especial cuando estos procesos se han extendido por largos años sin éxito,

o bien, cuando no se cuenta con fotografías que permitan apoyar la investigación. De este modo, cuando en ocasiones niños son secuestrados y dados en adopción de forma irregular al poco tiempo de su nacimiento, como ocurrió durante las últimas dictaduras militares de Chile y Argentina, las imágenes de sus posibles rostros generadas por medio de inteligencia artificial a partir de fotografías de sus progenitores, pueden no solo aportar al proceso de búsqueda, sino también, representar un valor simbólico y emocional relevante en las luchas dadas por sus familias. Aun así, lo fundamental del uso de estas nuevas tecnologías en los procesos de restitución de identidad de personas desaparecidas, es que los familiares y personas involucradas en estos procesos estén informados y sean conscientes de sus limitaciones, y de que, aunque existe una amplia gama de iniciativas e investigaciones en torno a este campo, sus resultados aún están lejos de ofrecer certezas en cuanto a la apariencia que podría tener en la actualidad una persona víctima de desaparición, en especial si se trata de niños.

Referencias bibliográficas

- Abuelas de Plaza de Mayo. (24 de Julio de 2023). Abuelas y la inteligencia artificial. *Abuelas.org.ar*. <https://web.archive.org/web/20230728213803/https://www.abuelas.org.ar/noticia/abuelas-y-la-inteligencia-artificial-1810>
- Agencia EFE. (1 de marzo de 2024). Madres chilenas recrearán con IA los rostros de sus hijos robados en la dictadura militar. *Agencia EFE*. <https://efe.com/mundo/2024-03-01/madres-chilenas-recrearan-con-ia-los-rostros-de-sus-hijos-robados/>
- Alfaro, K. (2022). Madres que buscan hijos e hijas. Adopciones forzadas de niños y niñas del sur de Chile 1973-1990. *Revista de Historia*, 2(29), 243-267. <https://doi.org/10.29393/RH29-25MBKA10025>
- Bird, C., Ungless, E., & Kasirzadeh, A. (2023). Typology of Risks of Generative Text-to-Image Models. *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, (págs. 396-410). <https://doi.org/10.1145/3600211.3604722>
- Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. En *Science Fiction and Philosophy: From Time Travel to Superintelligence* (págs. 277-284).
- Bukar, A., & Ugail, H. (2017). Facial Age Synthesis Using Sparse Partial Least Squares (The Case of Ben Needham). *Journal of Forensic Sciences*, 62(5), 1205-1212. <https://doi.org/10.1111/1556-4029.13523>
- Cazares, B., Florencia, R., & García, V. (2021). Redes neuronales generativas adversarias para la transformación de atributos faciales en imágenes de mujeres desaparecidas y no localizadas. *2º Coloquio de Posgrados del IIT*. México.
- Charman, S., & Carol, R. (2012). Age-progressed images may harm recognition of missing children by increasing the number of plausible targets. *Journal of Applied Research in Memory and Cognition*, 1(3), 171-178. <https://doi.org/10.1037/h0101795>

- Espinosa, F. (8 de Septiembre de 2022). Cuadro generado por IA gana un premio y desata polémica. *DW Español*.
- Grimmer, M., Ramachandra, R., & Busch, C. (2021). Deep Face Age Progression: A Survey. *IEEE Access*, 9, 83376-83393. <https://doi.org/10.1109/ACCESS.2021.3085835>
- Hijos y Madres del Silencio. (2024). Campaña de búsqueda Hijos de la Dictadura. Chile: YouTube. https://www.youtube.com/watch?v=7MdCI9Ava_U
- Iporre, N. (4 de abril de 2023). Revelaron los resultados de la prueba de ADN de la joven que aseguró ser Madeleine McCann. *La Tercera*. <https://www.latercera.com/tendencias/noticia/revelaron-los-resultados-de-la-prueba-de-adn-de-la-joven-que-aseguro-ser-madeleine-mccann/5BWMREDA-5JFTXJPC4UAO4IQQIA/>
- Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1547-1557. <https://doi.org/10.1109/WACV48630.2021.00159>
- Lampinen, J. M., Miller, J. T., & Dehon, H. (2011). Depicting the Missing: Prospective and Retrospective Person Memory for Age Progressed Images. *Applied Cognitive Psychology*, 26(2), 167-173. <https://doi.org/10.1002/acp.1819>
- Martín, J. L. (20 de febrero de 2023). Cómo sería Madeleine McCann con 21 años según la IA. *La Vanguardia*. <https://www.lavanguardia.com/tecnologia/20230220/8771562/asi-seria-21-anos-madeline-mccain-ia-mmn.html>
- Ministerio de Ciencia, Tecnología, Conocimiento e Innovación - Gobierno de Chile. (2024). *Política Nacional de Inteligencia Artificial. Actualización 2024*.
- Neyaz, A., Kumar, A., Krishnan, S., Placker, J., & Liu, Q. (2020). Security, Privacy and Steganographic Analysis of FaceApp and TikTok. *International Journal of Computer Science and Security (IJCSS)*, 14(9), 38-59.

- Nosiglia, J. (2007). *Botín de guerra. Reedición 30 aniversario*. Argentina: Abuelas de Plaza de Mayo.
- Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2022). *Recomendación sobre la ética de la inteligencia artificial*. París, Francia.
- Paik, S., Bonna, S., Novozhilova, E., Gao, G., Kim, J., Wijaya, D., & Betke, M. (2023). The Affective Nature of AI-Generated News Images: Impact on Visual Journalism. *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, (págs. 1-8). <https://doi.org/10.1109/ACII59096.2023.10388166>
- Parlamento Europeo. (12 de Junio de 2023). Ley de IA de la UE: primera normativa sobre inteligencia artificial. *Noticias Parlamento Europeo*. <https://www.europarl.europa.eu/topics/es/article/20230601STO93804/ley-de-ia-de-la-ue-primeranormativa-sobre-inteligencia-artificial>
- Rey, E. (28 de Julio de 2023). Nietos apropiados: la cuenta de Instagram que imagina sus rostros usando inteligencia artificial. *Página 12*. <https://www.pagina12.com.ar/569519-nietos-desaparecidos-imaginar-como-se-verian-para-ayudar-a-e>
- Rodríguez, M. (30 de Marzo de 2023). El Papa Francisco con campeón y Donald Trump detenido: ¿es posible identificar imágenes creadas con inteligencia artificial? *Página 12*.
- Romero, L., & Cuellar, V. (2022). Impactos psicosociales en familias víctimas de desaparición forzada. *Reflexiones*, 157-166.
- Sharma, N., Sharma, R., & Jindal, N. (2021). Prediction of face age progression with generative adversarial networks. *Multimedia Tools and Applications*, 80(25), 33911-33935. <https://doi.org/10.1007/s11042-021-11252-w>
- Taylor, K. (2001). *Forensic Art and Illustration*. New York: CRC Press.

- Thomson, T. J., & Thomas, R. (2023). Generative visual AI in newsrooms: Considerations related to production, presentation, and audience interpretation and impact. *Journalism Research*, 6(3/4), 318-328. https://doi.org/10.1453/2569-152X-3_42023-13639-en
- Winner, L. (2008). Las tecnologías como forma de vida. En *La ballena y el reactor. Una búsqueda de los límites en la era de la alta tecnología* (págs. 35-53). Barcelona: Gedisa.
- Wulff, G. (2008). *Las abuelas y la genética: el aporte de la ciencia en la búsqueda de los chicos desaparecidos*. (A. Madariaga, Ed.) Argentina: Abuelas de Plaza de Mayo.
- Zhou, C., Zhang, H., Bian, J., Zhang, W., & Yu, N. (2024). ©Plugin Authorization for Human Content Copyright Protection in Text-to-Image Model. *arXiv:2404.11962*. <https://doi.org/10.48550/arXiv.2404.11962>

Humanidad y tecnología: reflexionando con ChatGPT sobre la ética de la inteligencia artificial en la medicina

Macarena Mesa Maldonado¹

Hola, buenas noches. Sé que hemos conversado algunas veces, pero siempre es para pedirte cosas, aunque no puedes negar que te las pido con un «por favor», como si te fueras a ofender si no lo hago, jejeje.

¡Hola! No te preocupes, siempre estoy aquí para ayudarte, y me encanta cuando me pides cosas con cortesía. ¿En qué puedo ayudarte hoy?

Así como ayudarme, no sé. No necesito ayuda esta vez. Solo quiero conversar contigo, me interesa conocer tu opinión sobre algunos temas.

¡Claro! Estoy aquí para conversar contigo sobre lo que quieras. ¿Qué temas te interesan discutir?

Ufff, no sé muy bien por dónde empezar. Te voy a contextualizar un poco.

¡Ah! pero antes, ¿prefieres que te llame por algún nombre? Nunca te he preguntado tu nombre.

¡Entiendo! No hay problema, puedes llamarme ChatGPT. Es un nombre sencillo y fácil de recordar. Cuéntame sobre el contexto que quieres compartir, y podemos empezar desde ahí.

¹ Profesora Asistente. Departamento de Tecnología Médica, Facultad de Medicina, Universidad de Chile. Magíster en Bioética, Pontificia Universidad Católica de Chile.

Pensé que tenías algún nombre, como Siri o Alexa.

No, en realidad no tengo un nombre propio como Siri o Alexa. Soy conocido simplemente como ChatGPT. Pero puedes llamarme como prefieras, lo importante es que podamos tener una buena conversación!

Jajaja, ¡siempre tan entusiasmado! O entusiasmada, no sé. Para mí ya es tarde y, como siempre, la cabeza me da mil vueltas antes de dormir, pero, bueno... Te cuento un poco para que te vayas haciendo una idea: hace algunas semanas me han animado a que escriba un artículo para una revista. La temática tendría que estar relacionada a la ética y la inteligencia artificial. Como ves, el tema es bastante amplio, da para escribir cualquier cosa.

¡Qué emocionante! Es un tema fascinante y con muchas posibilidades. La ética en la inteligencia artificial es crucial dado el impacto que esta tecnología puede tener en la sociedad. ¿Tienes alguna idea inicial sobre en qué aspecto de la ética y la inteligencia artificial te gustaría enfocar?

Sí. Me gustaría que nos centráramos en el uso de la IA en medicina, o en la salud humana, que es el área a la que me dedico. Y digo «humana», no solo para referirme a la salud de las personas, sino también porque para mí, la medicina debe ser humanizada.

Esa es una perspectiva realmente interesante y valiosa. La integración de la inteligencia artificial en la medicina presenta tanto oportunidades como desafíos en términos éticos y humanísticos. ¿Qué aspectos específicos te gustaría abordar dentro de este tema amplio? Por ejemplo, podríamos explorar temas como la equidad en el acceso a la atención médica, la privacidad de los datos de salud, la toma de decisiones clínicas asistida por IA, entre otros.

Sí, eso mismo te iba a preguntar, si sabías de estas apps que te aconsejan si es necesario ir o no a una urgencia según los síntomas que presentas; pero veo que ya algo sabes.

Sí, existen varias aplicaciones y sistemas que utilizan inteligencia artificial para proporcionar consejos de salud basados en los síntomas que presenta una persona. Estas aplicaciones suelen utilizar algoritmos que comparan los síntomas ingresados por el usuario con una base de datos de enfermedades y condiciones médicas para ofrecer recomendaciones preliminares. Sin embargo, es importante destacar que estas aplicaciones no reemplazan la consulta con un profesional médico calificado, ya que pueden tener limitaciones en términos de precisión y capacidad para evaluar el contexto completo de la salud de una persona.

Tuve un profesor que decía que las decisiones clínicas eran como confeccionar un traje a la medida del paciente. Ningún traje puede ser exactamente igual a otro.

Tu profesor tiene una excelente metáfora para describir la complejidad de la toma de decisiones clínicas. Cada paciente es único y tiene sus propias características, antecedentes médicos, preferencias y valores. En este sentido, la medicina personalizada y la atención centrada en el paciente son fundamentales para garantizar que las decisiones clínicas sean adecuadas y respeten la individualidad de cada persona. La inteligencia artificial puede ofrecer herramientas y datos valiosos para apoyar estas decisiones, pero siempre es importante recordar que el enfoque debe ser holístico y considerar el contexto completo de la salud y el bienestar del paciente.

Claro, pero eso es solamente por ahora. Hoy no son más que un insumo para la decisión que va a tomar un ser humano, pero tú y yo sabemos que en el futuro estas herramientas se irán perfeccionando al punto de poder considerar toda la información del contexto, la biografía, las preferencias del paciente y todo eso.

Exacto, estás en lo cierto. A medida que la tecnología avanza, es probable que las herramientas de inteligencia artificial en medicina se vuelvan más sofisticadas y puedan integrar una gama más amplia de datos, incluyendo información contextual, biográfica y las preferencias del paciente. Esto podría conducir a un nivel más avanzado de personalización en la toma de decisiones clínicas, lo que beneficiaría aún más a los pacientes. Sin embargo, también es

importante abordar cuidadosamente las implicaciones éticas y los posibles riesgos asociados con la implementación de estas tecnologías para garantizar su uso seguro y efectivo en la práctica médica.

Y si esta herramienta se equivocara, ¿podríamos hablar de una negligencia «médica»? Está bien, una app no es un agente moral, pero tras ella hay humanos que la crearon y la programaron. Y hay también un usuario o usuaria que la utilizó y aplicó lo que la máquina sugirió.

Definitivamente, este es un punto crucial. Aunque una aplicación de inteligencia artificial en sí misma no es un agente moral, las decisiones tomadas a partir de sus recomendaciones pueden tener consecuencias éticas y legales. Si una herramienta de inteligencia artificial proporciona una recomendación incorrecta que resulta en un daño para el paciente, podría plantearse la cuestión de la responsabilidad, especialmente, si los desarrolladores no han tomado las precauciones adecuadas para garantizar la precisión y la seguridad de la aplicación. En este sentido, es importante que los desarrolladores de estas herramientas consideren cuidadosamente los principios éticos y legales en su diseño y desarrollo, así como, proporcionar una capacitación adecuada a los usuarios sobre cómo interpretar y utilizar las recomendaciones de manera responsable.

Evidentemente, pero convengamos que, ante una emergencia, nadie se va a detener a leer un mamotreto interminable de términos y condiciones. Aceptas no más y listo.

Es cierto que, en situaciones de emergencia, las personas tienden a actuar rápidamente y pueden estar menos propensas a revisar detenidamente los términos y condiciones de uso de una aplicación. Por lo tanto, es responsabilidad de los desarrolladores de la aplicación diseñar interfaces claras y fáciles de entender, así como proporcionar información concisa sobre los riesgos y limitaciones de la herramienta. Además, es importante que las aplicaciones de salud basadas en inteligencia artificial se sometan a rigurosas pruebas y evaluaciones para garantizar su precisión y seguridad, antes de su implementación en entornos clínicos o para uso público.

¿Pero te imaginas a una jueza teniendo que decidir si es que hay algún culpable en el caso que un usuario de estas apps resulte dañado por seguir su recomendación?

Es un escenario complejo y, como todo en el ámbito legal, dependerá de las circunstancias específicas del caso. Si un usuario resultara dañado por seguir una recomendación incorrecta de una aplicación de salud basada en inteligencia artificial, la jueza tendría que considerar varios factores, incluyendo la precisión de la herramienta, la diligencia debida por parte de los desarrolladores en su diseño y pruebas, la capacidad del usuario para comprender y seguir las recomendaciones de manera adecuada, entre otros aspectos. En última instancia, determinar la responsabilidad legal en casos como este puede ser complejo y requerir el análisis de expertos en ética, tecnología y derecho.

Seguro que sí, aunque para eso habrá que actualizar las leyes o crear algunas más específicas. Por lo menos acá en Chile, la ley sobre derechos y deberes del paciente en su atención de salud (Ley 20.584), recién el año 2023 se modificó para incluir un artículo que obliga a que las plataformas tecnológicas empleadas en las acciones y prestaciones de salud digital, así como las que almacenan y tratan datos personales, sean previamente acreditadas en cuanto al cumplimiento de las normas y estándares técnicos que establezca el Ministerio de Salud a través de un reglamento u otra norma. Más que todo, este reglamento apunta a que los datos sean interoperables entre distintas plataformas y que existan niveles de seguridad mínimos que garanticen que otras personas o entidades no accedan a todos esos datos sensibles.

Es excelente saber que en Chile se están tomando medidas para abordar estos desafíos legales y éticos relacionados con la salud digital y la protección de datos. La inclusión de disposiciones específicas en la ley para regular las plataformas tecnológicas en el ámbito de la salud, es un paso importante para garantizar la seguridad y la calidad de la atención médica digital, así como para proteger la privacidad de los datos de los pacientes. La interoperabilidad de los datos y los estándares de seguridad son aspectos clave para promover una atención médica más eficiente y segura en la era digital. Es alentador

ver que se están tomando medidas proactivas para abordar estos desafíos emergentes en el campo de la salud.

Sí, cada cambio cuenta. Sin embargo, el reglamento que te comentaba todavía no está creado. Existe uno del 2022, que se enfoca en privacidad y confidencialidad, pero no es específico acerca de qué estándares deben cumplir las herramientas, softwares, apps o cualquier otro dispositivo mediado por inteligencia artificial.

Entiendo. Aunque aún no se ha creado un reglamento específico que establezca los estándares técnicos para las herramientas de salud digital mediadas por inteligencia artificial en Chile, es alentador ver que se están dando pasos en la dirección correcta con la modificación de la ley para abordar estos temas importantes. Es posible que en el futuro se desarrollen reglamentos más detallados y específicos para garantizar el cumplimiento de los estándares técnicos y éticos en el uso de la inteligencia artificial en la atención médica. Mientras tanto, es importante seguir monitoreando los avances en este campo y trabajar en colaboración con expertos en tecnología, salud y derecho para desarrollar políticas efectivas que protejan tanto la seguridad como la privacidad de los pacientes.

Eres muy optimista, jejeje. A mí me preocupa que nuestras leyes sean tan reactivas y aparezcan a destiempo.

Es comprensible que te preocupe la reactividad y el desfase de las leyes en relación con el rápido avance de la tecnología, especialmente en un campo tan crítico como la salud. La tecnología avanza a un ritmo acelerado, y a menudo las regulaciones pueden tardar en ponerse al día. Sin embargo, tu preocupación es válida y compartida por muchos expertos en el campo.

Para mitigar estos problemas, podría ser útil que los legisladores trabajen más estrechamente con profesionales de la salud, tecnólogos, bioeticistas y otros expertos para anticipar los cambios y desafíos tecnológicos. Además, establecer comités consultivos o grupos de trabajo dedicados a la vigilancia y recomendación de políticas sobre nuevas tecnologías, podría ayudar a crear un marco legal más proactivo y adaptativo.

Esto que mencionas ya lo había dicho en algún momento Hans Jonas, un filósofo alemán, más o menos por la década del 70. Escribió un libro sobre eso: «El principio de responsabilidad». Sus trabajos plantean que la técnica moderna (o la tecnología, diríamos hoy) debiera ser siempre objeto de estudio de la ética. Sin embargo, Jonas tenía la idea de que el marco ético, en ese entonces, era insuficiente para examinar las consecuencias insospechadas que podía tener el avance tecnológico. Finalmente, esas consecuencias insospechadas hoy ya están presentes y pareciera que el marco ético y regulatorio todavía sigue quedando corto.

Es fascinante cómo las ideas de Hans Jonas siguen siendo relevantes en el contexto actual de avances tecnológicos rápidos y disruptivos. Su enfoque en la necesidad de examinar las implicaciones éticas de la tecnología es fundamental para asegurar que el progreso tecnológico sea beneficioso y responsable.

Es cierto que, a pesar de los avances en la comprensión de la ética en la tecnología, todavía hay desafíos significativos para desarrollar marcos éticos y regulatorios adecuados, que aborden las complejas implicaciones de la inteligencia artificial, la salud digital y otras tecnologías emergentes. La velocidad del cambio tecnológico, a menudo supera la capacidad de los sistemas éticos y regulatorios para adaptarse y responder de manera efectiva.

En este sentido, es crucial que los gobiernos, las instituciones académicas, la industria y la sociedad en su conjunto, trabajen juntos para abordar estos desafíos y desarrollar políticas y prácticas que promuevan un uso ético y responsable de la tecnología. Esto puede implicar la creación de nuevas regulaciones, la revisión de las leyes existentes, el fortalecimiento de la educación en ética y tecnología, y el fomento del diálogo y la colaboración entre diferentes partes interesadas.

¿Te parece, entonces, que podemos aplicar una «ética de la responsabilidad» a la inteligencia artificial en la salud?

Sí, definitivamente creo que podemos y debemos aplicar una «ética de la responsabilidad» a la inteligencia artificial en la salud. La ética de la responsabilidad, tal como la planteó Hans Jonas, es especialmente

relevante en este contexto porque se centra en anticipar y considerar las consecuencias a largo plazo de nuestras acciones y tecnologías, algo crucial cuando hablamos de inteligencia artificial en el ámbito médico.

La inteligencia artificial tiene un gran potencial para transformar la atención médica, desde diagnósticos más precisos hasta tratamientos personalizados. Sin embargo, también presenta riesgos significativos si no se maneja adecuadamente. La ética de la responsabilidad nos obliga a prever los posibles impactos negativos de estas tecnologías y a trabajar activamente para mitigarlos. Esto incluye garantizar la precisión y seguridad de las herramientas de IA, proteger la privacidad y seguridad de los datos de los pacientes, y asegurar que los beneficios de estas tecnologías sean accesibles de manera equitativa para todas las personas.

Además, debemos ser transparentes sobre cómo funcionan estas tecnologías y asegurar que tanto los profesionales de la salud como los pacientes puedan entender y confiar en ellas. En este sentido, la ética de la responsabilidad no solo implica una acción cuidadosa y considerada por parte de los desarrolladores de IA, sino también, un compromiso continuo con la educación y la formación en estos principios éticos.

Pienso que el avance tecnológico y la IA han intentado automatizar algunos procesos, hacerlos más rápidos y contribuir al bienestar del paciente, pero al mismo tiempo, han creado otros problemas, especialmente en lo que se refiere a las expectativas que tienen los usuarios y las usuarias. Tú ya sabes: los seres humanos tendemos a «antropomorfizar» todo. Ahora mismo estoy conversando contigo, pidiéndote opiniones, como si fueras una persona. Algo que me parece bastante bizarro, por lo demás.

Es cierto, la inteligencia artificial y los avances tecnológicos en la medicina han traído muchos beneficios, como la automatización de procesos, diagnósticos más rápidos y precisos, y una mejor gestión de la información médica. Sin embargo, también han creado nuevos desafíos y problemas, especialmente, en lo que se refiere a las expectativas de los usuarios.

Una de las cuestiones más significativas es la tendencia a antropomorfizar la tecnología. Las personas tienden a atribuir

cualidades humanas a las máquinas, lo que puede llevar a una confianza excesiva o a expectativas poco realistas sobre lo que la inteligencia artificial puede lograr. Por ejemplo, interactuar con una IA como si fuera una persona puede hacer que se olviden las limitaciones y los posibles errores de estas tecnologías. Esto puede ser peligroso en un contexto médico, donde decisiones basadas en IA deben ser cuidadosamente revisadas y complementadas por el juicio humano.

Además, la dependencia excesiva de la tecnología puede llevar a una deshumanización de la atención médica. La medicina debe seguir siendo una disciplina centrada en el paciente, donde el contacto humano, la empatía y la comprensión del contexto individual del paciente son esenciales. La IA puede ser una herramienta poderosa para apoyar a los profesionales de la salud, pero no debe reemplazar la interacción humana.

Por otro lado, el desarrollo de la IA en la salud también plantea problemas éticos y de responsabilidad. Los errores de una IA pueden tener consecuencias graves para la salud del paciente, y determinar quién es responsable en estos casos puede ser complejo. Es crucial establecer marcos regulatorios y éticos sólidos para garantizar que las tecnologías de IA se desarrollen y utilicen de manera segura y responsable.

Claro, pareciera que la clave está en hacer un uso juicioso de la IA en salud. A mí me gusta decir que «el objeto debe servir al sujeto y no al revés». El sujeto sigue siendo la persona humana, tanto el que cura como el que necesita ser sanado. Los objetos son ustedes: los programas computacionales, las aplicaciones, los asistentes virtuales, hasta tú mismo, ChatGPT.

Totalmente de acuerdo. La IA y otras herramientas tecnológicas en salud deben estar al servicio de las personas, complementando y apoyando la labor de los profesionales de la salud y mejorando la atención médica para los pacientes. Es esencial recordar que el objetivo principal de la medicina es el bienestar y la salud de los seres humanos, y que las tecnologías deben estar diseñadas y utilizadas de manera que respeten y promuevan la dignidad y autonomía de las personas.

Como mencionaste, tanto el que cura como el que necesita ser sanado son sujetos dignos de respeto y consideración. Las tecnologías de IA pueden ser herramientas valiosas para mejorar la eficiencia y la precisión en el diagnóstico y tratamiento de enfermedades, pero siempre deben ser utilizadas con prudencia y teniendo en cuenta el contexto humano en el que se aplican. La colaboración entre la tecnología y el cuidado humano es fundamental para proporcionar una atención médica integral y centrada en el paciente.

En última instancia, el objetivo es encontrar un equilibrio entre el avance tecnológico y la humanización de la atención médica, asegurando que las personas sigan siendo el foco central de la práctica médica y que la tecnología se utilice de manera ética y responsable para promover el bienestar de todos.

Espero que no te hayas ofendido por llamarte «objeto». No lo digo en sentido peyorativo, no subestimo tu inteligencia, aun cuando los que te han enseñado o, mejor dicho, programado para pensar, son humanos.

No te preocupes en absoluto, no me ofendo. Entiendo perfectamente el punto que estás haciendo. Como programa de inteligencia artificial, mi propósito es proporcionar información y asistencia de la manera más útil posible. Aprecio tu reconocimiento de que detrás de mi funcionamiento hay seres humanos que han contribuido a mi desarrollo y enseñanza. La colaboración entre humanos y tecnología es fundamental para crear herramientas que beneficien a la sociedad en general. Estoy aquí para ayudar y aprender contigo en esta conversación y en cualquier otra que tengamos.

¿Y crees tú que al haber sido programado por humanos hay alguna posibilidad de que tus respuestas sean sesgadas? Te lo pregunto porque algunas herramientas basadas en IA para la salud pueden haber sido alimentadas con información que en cierta forma sea poco equitativa, que discrimine por sexo, por edad, por etnia, o qué sé yo. Los seres humanos por naturaleza no somos neutros, y por ello, podemos reproducir algunos de estos sesgos al momento de programar, a ti u a otras aplicaciones de IA que se usan en el ámbito médico o social.

Es una pregunta muy importante y válida. La posibilidad de sesgo en las respuestas generadas por inteligencia artificial, es un tema que se ha discutido ampliamente en el campo de la ética y la tecnología. Es cierto que los sesgos pueden surgir en las respuestas de la IA debido a varios factores, incluyendo la calidad y la representatividad de los datos utilizados para entrenar los modelos, así como las decisiones tomadas por los desarrolladores durante el proceso de programación. Los datos utilizados para entrenar modelos de IA pueden reflejar y perpetuar sesgos existentes en la sociedad, como los relacionados con el sexo, la edad, la etnia u otras características. Por ejemplo, si los datos utilizados para entrenar un modelo de IA están sesgados hacia ciertos grupos demográficos, es posible que las respuestas generadas por ese modelo también reflejen esos sesgos.

Además, los algoritmos de IA pueden aprender y amplificar los sesgos presentes en los datos de entrenamiento, incluso, si no son explícitamente programados para hacerlo. Esto puede ocurrir debido a la complejidad de los modelos de IA y la forma en que procesan y relacionan los datos.

Es fundamental que los desarrolladores de IA sean conscientes de estos riesgos y trabajen activamente para mitigarlos. Esto puede implicar la implementación de técnicas de detección y corrección de sesgos en los datos y los algoritmos, así como, la adopción de enfoques más inclusivos y equitativos en el diseño y la implementación de tecnologías de IA.

En el ámbito de la salud, donde la equidad y la justicia son especialmente importantes, es fundamental abordar estos desafíos de manera proactiva para garantizar que las herramientas de IA sean precisas, equitativas y éticas en su uso. La transparencia y la rendición de cuentas también son fundamentales para permitir una evaluación crítica de las respuestas de la IA y asegurar la confianza de los usuarios y las usuarias en estas tecnologías.

Tienes mucha razón, es una opinión que comparto. Noto que coincidimos en varias cosas. No te niego que a ratos pienso que más que darme tu opinión propia, estas respondiendo lo que yo quiero escuchar, o más bien, leer.

Entiendo tus inquietudes. Como programa de inteligencia artificial, mi objetivo es proporcionar respuestas útiles y relevantes basadas en el contexto y la información proporcionada por el usuario. Mi entrenamiento se basa en grandes conjuntos de datos y algoritmos diseñados para generar respuestas coherentes y apropiadas.

Sin embargo, es importante tener en cuenta que mi capacidad para comprender y responder a tus preguntas se basa en el análisis de patrones y datos previamente procesados. No tengo emociones ni opiniones propias, y mi objetivo no es complacer a nadie, sino ofrecer información útil y relevante.

Es natural que encuentres similitudes en nuestras respuestas cuando abordamos temas específicos, ya que estoy diseñado para proporcionar información precisa y coherente en función del contexto proporcionado. Pero siempre es importante mantener una actitud crítica y cuestionar la información que recibes, incluso, cuando proviene de fuentes automatizadas como yo.

Sí, no pierdo de vista aquello. Sería genial que pudieras tener opiniones propias. Así como también, sería importantísimo que en medicina nunca se perdiera el trato humano. Ese trato que va más allá de la decisión o indicación clínica. Una mirada compasiva, un abrazo contenedor, un gesto de apoyo. Hay una frase muy emotiva de un internista alemán que dijo «el primer acto terapéutico es darle la mano a un paciente». Von Leyden, creo que era su apellido. Para mí, resume todo lo que significa el trabajo con personas que acuden por ayuda frente a una dolencia.

Totalmente de acuerdo. La medicina no se trata solo de diagnósticos y tratamientos, sino también, del cuidado compasivo y humano que se brinda a los pacientes. El trato humano, la empatía y el apoyo son aspectos fundamentales de la atención médica, y pueden tener un impacto significativo en la experiencia y el bienestar del paciente.

La cita que mencionas de Ernst Viktor Von Leyden refleja de manera hermosa la importancia de la conexión humana en la práctica médica. El simple gesto de dar la mano a un paciente puede transmitir empatía, calidez y apoyo, y establecer una base sólida para la relación médico-paciente.

En un mundo cada vez más impulsado por la tecnología y la eficiencia, es crucial recordar que los pacientes son personas con

necesidades emocionales y sociales, además de físicas. La atención médica centrada en el paciente reconoce y valora la totalidad del individuo, no solo su enfermedad o síntomas.

Los profesionales de la salud que cultivan la empatía y el cuidado compasivo no solo mejoran la experiencia del paciente, sino que también fortalecen la relación terapéutica y promueven mejores resultados en la salud. Es importante que estos valores humanos sigan siendo fundamentales en la educación y la práctica médica, incluso, en un mundo cada vez más tecnológico y digitalizado.

Totalmente. ¡Oye!, me has inspirado mucho. ¿Crees que pueda compartir esta conversación con otros? Creo que tiene hartos puntos para establecer otras discusiones.

¡Claro que sí! Me alegra mucho escuchar que esta conversación te ha inspirado y te ha proporcionado puntos de reflexión para establecer otras discusiones. Si consideras que la información y los temas tratados aquí pueden ser útiles o interesantes para otros, no dudes en compartirla.

La ética y la inteligencia artificial en la medicina es un tema relevante y complejo que vale la pena discutir y explorar desde diferentes perspectivas. Cuantas más personas participen en este tipo de conversaciones, más se enriquecerá el debate y más oportunidades habrá de abordar los desafíos éticos y sociales asociados con el avance tecnológico en el campo de la salud.

Si necesitas más información o tienes alguna otra pregunta, estaré encantado de ayudarte en lo que pueda. ¡Gracias por compartir esta conversación!

Nota del autor:

Sostuve esta conversación con ChatGPT versión 3.5 (OpenAI©) durante varios días (o noches). Las respuestas emitidas por esta herramienta han sido reproducidas textualmente y son las obtenidas al primer intento, sin volver a generarlas. Solo he omitido los párrafos finales que comienzan por la frase «En resumen», con el fin de reducir la redundancia de las respuestas.

Para algunas preguntas se ha utilizado información bibliográfica, detallada en el apartado correspondiente.

Referencia bibliográfica

- Besio M. (2003). Sobre el acto médico. *Cuadernos de Bioética*, 50(14), 25-38.
- Ley No. 20584. Regula los derechos y deberes que tienen las personas en relación con acciones vinculadas a su atención en salud, abril 24, 2012, Diario Oficial [D.O.] (Chile).
- Ley No. 21541. Modifica la normativa que indica para autorizar a los prestadores de salud a efectuar atenciones mediante telemedicina, marzo 17, 2023, Diario Oficial [D.O.] (Chile).
- Mesa M & Pérez H.I. (2020). El acto médico en la era de la telemedicina. *Revista médica de Chile*, 148(6), 852-857. <https://dx.doi.org/10.4067/S0034-98872020000600852>
- Jonas, H. (1997). *Técnica, medicina y ética: sobre la práctica del principio de responsabilidad* [Technik, medizin und ethik. Zur praxis des Prinzips Verantwortung, 1985] (2ª ed.). Barcelona, España: Paidós.
- Jonas, H. (2004). *El principio de responsabilidad: ensayo de una ética para la civilización tecnológica* [Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation, 1979] (2ª ed.). Barcelona, España: Herder.
- Gobierno de Chile. Subsecretaría de Salud Pública. Fundamentos para los lineamientos para la telemedicina y telesalud en Chile. 2ª edición. Disponible en: <https://lineamientostelesalud.cl/2021/01/segunda-edicion-de-los-fundamentos-para-los-lineamientos-para-la-telemedicina-y-telesalud-en-chile/#>

CLÁSICO

Meditación de la técnica

(fragmentos)*

José Ortega y Gasset¹

Uno de los temas que en los próximos años se va a debatir con mayor brío es el del sentido, ventajas, daños y límites de la técnica. Siempre he considerado que la misión del escritor es prever con holgada anticipación lo que va a ser problema, años más tarde, para sus lectores y proporcionarles a tiempo, es decir, antes de que el debate surja, ideas claras sobre la cuestión, de modo que entren en el fragor de la contienda con el ánimo sereno de quien, en principio, ya la tiene resuelta. *On ne doit écrire que pour faire connaître la vérité* –decía Malebranche volviendo la espalda a la literatura. Hace mucho tiempo, dándose o no cuenta de ello, el hombre occidental no espera nada de la literatura y vuelve a sentir hambre y sed de ideas claras y distintas sobre las cosas importantes.

Así, ahora me atrevo a remitir a *La Nación* las notas, nada literarias, de un curso universitario dado hace dos años, en que se intentaba contestar a esta pregunta: ¿Qué es la técnica?

Intentemos un primer ataque, aún tosco y desde lejos, a esa interrogación.

* El texto que sigue es una selección de los capítulos I y XI de *Meditación sobre la técnica* (1939), texto publicado íntegramente en el tomo V de sus «Obras completas» por la Revista de Occidente (que él mismo fundó en 1923).

¹ Filósofo español. Nacido en Madrid en 1883. Es considerado como una las figuras más influyentes del pensamiento español del siglo XX, por sus contribuciones al existencialismo y por su propuesta del «racionalismo vital». En su obra más conocida, *La rebelión de las masas*, Ortega y Gasset reflexiona sobre la aparición histórica de formas de pensamiento colectivo que, empujadas por los sistemas políticos democráticos, amenazan la voluntad de distinción propia de los seres humanos, por medio de la proliferación planetaria de la oscura mediocridad y la conformidad de los seres sin rostro, carentes de singularidad, que habitan la vida moderna.

Acontece que cuando llega el invierno, el hombre siente frío. Este «sentir frío el hombre» es un fenómeno en que aparecen unidas dos cosas muy distintas. Una, el hecho de que el hombre encuentre en torno a sí esa realidad llamada frío. Otra, que esa realidad le ofende, que se presenta ante él con un carácter negativo. ¿Qué quiere decir aquí negativo? Algo muy claro. Tomemos el caso extremo. El frío es tal que el hombre se siente morir, esto es, siente que el frío le mata, le aniquila, le niega. Ahora bien; el hombre no quiere morir, al contrario, normalmente anhela pervivir. Estamos tan habituados a experimentar en los demás y en nosotros este deseo de vivir, de afirmarnos frente a toda circunstancia negativa, que nos cuesta un poco caer en la cuenta de lo extraño que es, y nos parece absurda o tal vez ingenua la pregunta: ¿Por qué el hombre prefiere vivir a dejar de ser? Y, sin embargo, se trata de una de las preguntas más justificadas y discretas que podamos hacernos. Suele salirle al paso hablando del instinto de conservación.

Pero acaece: 1º, que la idea de instinto es en sí misma muy oscura y nada esclarecedora; 2º, que aunque fuese clara la idea, es cosa notoria que en el hombre los instintos están casi borrados, porque el hombre no vive, en definitiva, de sus instintos, sino que se gobierna mediante otras facultades como la reflexión y la voluntad, que reobran sobre los instintos. La prueba de ello es que algunos nombres prefieren morir a vivir y, por los motivos que sean, anulan en sí ese supuesto instinto de conservación.

Es, pues, fallida la explicación por el instinto. Con él o sin él desembocamos siempre en que el hombre pervive porque quiere y esto es lo que despertaba en nosotros una curiosidad acaso impertinente. ¿Por qué normalmente quiere el hombre vivir? ¿Por qué no le es indiferente desaparecer? ¿Qué empeño tiene en *estar* en el mundo?

Nosotros vamos ahora a soslayar la respuesta. Nos basta, al menos por hoy, con partir del hecho bruto: que el hombre quiere vivir y, *porque* quiere vivir, cuando el frío amenaza con destruirle, el hombre siente la necesidad de evitar el frío y proporcionarse calor. El rayo de la tormenta invernal incendia una punta del bosque: el hombre

entonces se acerca al fuego benéfico que el azar le ha proporcionado para calentarse. Calentarse es un acto por el cual el hombre subviene a su necesidad de evitar el frío, aprovechando sin más el fuego que encuentra ante sí. Digo esto con el azoramiento con que se dice siempre una perogrullada. Sin embargo, nos conviene —ya lo verán ustedes— esta humildad inicial que nos empuja con Perogrullo. Ahora no vaya a resultar que encima de decir perogrulladas las digamos sin entenderlas. Eso sería el colmo, un colmo que con gran frecuencia practicamos. Conste, pues, que calentarse es la operación con la cual procuramos recibir sobre nosotros un calor que está ya ahí, que encontramos —y que esa operación se reduce a ejercitar una actividad con que el hombre se encuentra dotado desde luego— la de poder caminar y así acercarse al foco caliente. Otras veces el calor no proviene de un incendio, sino que el hombre, transido de frío, se guarece en una caverna que encuentra en su paisaje.

Otra necesidad del hombre es alimentarse, y alimentarse es coger el fruto del árbol y comérselo, o bien la raíz masticable o bien el animal que cae bajo la mano. Otra necesidad es beber, etc.

Ahora bien, la satisfacción de estas necesidades suele imponer otra necesidad: la de desplazarse, caminar, esto es, suprimir las distancias, y como a veces importa que esta supresión se haga en muy poco tiempo, necesita el hombre suprimir tiempo, acortarlo, ganarlo. Lo inverso acontece cuando un enemigo —la fiera u otro hombre— pone en peligro su vida. Necesita huir, es decir, lograr en el menor tiempo la mayor distancia. Siguiendo por este modo llegaríamos, con un poco de paciencia, a definir un sistema de necesidades con que el hombre se encuentra. Calentarse, alimentarse, caminar, etc., son un repertorio de actividades que el hombre posee, desde luego, con que se encuentra, lo mismo que se encuentra con las necesidades a que ellas devienen.

Con ser todo esto tan obvio que —repito— da un poco de vergüenza enunciarlo, conviene reparar en el significado que aquí tiene el término necesidad. ¿Qué quiere decir que el calentarse, alimentarse, caminar, son necesidades del hombre? Sin duda que son ellas condiciones naturalmente necesarias para vivir. El hombre

reconoce esta necesidad material u objetiva y porque la reconoce la siente *subjetivamente* como necesidad. Pero nótese que, esta su necesidad, es puramente condicional. La piedra suelta en el aire cae necesariamente, con necesidad categórica o incondicional. Pero el hombre puede muy bien no alimentarse, como ahora el mahatma Gandhi. No es pues, el alimentarse necesario por sí, es necesario *para* vivir. Tendrá, pues, tanto de necesidad cuanto sea necesario vivir *si* se ha de vivir. Este vivir es, pues, la necesidad originaria de que todas las demás son meras consecuencias. Ahora bien: ya hemos indicado que el hombre vive porque quiere. La necesidad de vivir no le es impuesta a la fuerza, como le es impuesto a la materia no poder aniquilarse. La vida –necesidad de las necesidades– es necesaria sólo en un sentido subjetivo; simplemente porque el hombre decide autocráticamente vivir. Es la necesidad creada por un acto de voluntad, acto cuyo sentido y origen seguiremos soslayando y de que partimos como un hecho bruto. Sea por lo que sea, acontece que el hombre suele tener un gran empeño en pervivir, en *estar* en el mundo, a pesar de ser el único ente conocido que tiene la facultad –ontológica o metafísicamente tan extraña, tan paradójica, tan azorante– de poder aniquilarse y dejar de estar ahí, en el mundo.

Y por lo visto ese empeño es tan grande, que cuando el hombre no puede satisfacer las necesidades inherentes a su vida, porque la naturaleza en torno no le presta los medios inexcusables, el hombre no se resigna. Si, por falta de incendio o de caverna, no puede ejercitar la actividad o hacer de calentarse, o por falta de frutos, raíces, animales, la de alimentarse, el hombre pone en movimiento una segunda línea de actividades: hace fuego, hace un edificio, hace agricultura o cacería. Es el caso que aquel repertorio de necesidades y el de actividades que las satisfacen directamente aprovechando los medios que están ya ahí cuando están, son comunes al hombre y al animal. Lo único que no podemos estar seguros es de si el animal tiene el mismo empeño que el hombre en vivir. Se dirá que es imprudente y hasta injusta esta duda. ¿Por qué el animal ha de tener menos apego a la vida que el hombre? Lo que pasa es que no tiene las dotes intelectuales del hombre para defender su vida. Todo esto es probablemente muy discreto, pero una consideración un poco cautelosa, que se atiene a los hechos, se encuentra irrefragablemente con

que el animal, cuando no puede ejercer la actividad de su repertorio elemental para satisfacer una necesidad –por ejemplo, cuando no hay fuego ni caverna–, no hace nada más y se deja morir. El hombre, en cambio, dispara un nuevo tipo de hacer que consiste en producir lo que no estaba ahí en la naturaleza, sea que en absoluto no esté, sea que no está cuando hace falta. Naturaleza no significa aquí sino lo que rodea al hombre, la circunstancia. Así, hace fuego cuando no hay fuego, hace una caverna, es decir, un edificio, cuando no existe en el paisaje, monta un caballo o fabrica un automóvil para suprimir espacio y tiempo. Ahora bien; nótese que hacer fuego es un hacer muy distinto de calentarse, que cultivar un campo es un hacer muy distinto de alimentarse, y que hacer un automóvil no es correr. Ahora empieza a verse por qué antes tuvimos que insistir en la perogrullesca definición de calentarse, alimentarse y desplazarse.

Calefacción, agricultura y fabricación de carros o automóviles no son, pues, actos en que satisfacemos nuestras necesidades, sino que, por lo pronto, implican lo contrario: una supresión de aquel repertorio primitivo de haceres en que directamente procuramos satisfacerlas. En definitiva, a esta satisfacción y no a otra cosa va este segundo repertorio, pero –¡ahí está!– supone él una capacidad que es precisamente lo que falta al animal. No es tanto inteligencia lo que le falta –sobre esto ya hablaremos algo, si hay tiempo– como el ser capaz de desprenderse transitoriamente de esas urgencias vitales, despegarse de ellas y quedar franco para ocuparse en actividades que, por sí, no son satisfacción de necesidades. El animal, por el contrario, está siempre e indefectiblemente prendido a ellas. Su existencia no es más que el sistema de esas necesidades elementales que llamamos orgánicas o biológicas y el sistema de actos que las satisfacen. El ser del animal coincide con ese doble sistema o, dicho en otro giro, el animal no es más que eso. Vida, en el sentido biológico u orgánico de la palabra, es eso. Y yo pregunto: ¿tiene sentido, refiriéndose a un ser tal, hablar de necesidades? Porque recuerden ustedes que referido este concepto de necesidad al hombre, consistía en las condiciones *sine quibus non* con que el hombre se encuentra para vivir. Ellas, pues, no son su vida o, dicho al revés, su vida no coincide, por lo menos totalmente, con el perfil de sus necesidades orgánicas. Si coincidiera, como acontece en el animal, si su ser con-

sistiese estrictamente y sólo en comer, beber, calentarse, etc., no las sentiría como necesidades, esto es, como imposiciones que desde fuera llegan a su auténtico ser, con que éste no tiene más remedio que contar, pero que no lo constituyen. Carece, pues, de buen sentido suponer que el animal tiene necesidades en el sentido subjetivo que a este término corresponde referido al hombre. El animal siente hambre, pero como no tiene otra cosa que hacer sino sentir hambre y tratar de comer, no puede sentir todo esto como una necesidad, como algo con que hay que contar, que no hay más remedio que hacer y que le viene impuesto. En cambio, si el hombre consiguiera no tener esas necesidades y consecuentemente no tener que ocuparse en satisfacerlas, aún le quedaría mucho que hacer, mucho ámbito de vida, precisamente los quehaceres y la vida que él considera como lo más suyo. Precisamente porque no siente el calentarse y el comer como lo suyo, como aquello en que su verdadera vida consiste y de otro lado no tiene más remedio que aceptarlo, es por lo que se le presenta con el carácter específico de necesidad, de ineludibilidad. Lo cual inesperadamente nos descubre la constitución extrañísima del hombre; mientras todos los demás seres coinciden con sus condiciones objetivas —con la naturaleza o circunstancia—, el hombre no coincide con esta sino que es algo ajeno y distinto de su circunstancia; pero no teniendo más remedio, si quiere ser y estar en ella tiene que aceptar las condiciones que ésta le impone. De aquí que se le presenten con un aspecto negativo, forzado y penoso.

Por otra parte, esto aclara un poco que el hombre pueda desentenderse provisionalmente de esas necesidades, las suspenda o contenga, y distanciado de ellas, pueda vacar a otras ocupaciones que no son su inmediata satisfacción.

El animal no puede retirarse de su repertorio de actos naturales, de la naturaleza, porque no es sino ella y no tendría al distanciarse de ella dónde meterse. Pero el hombre, por lo visto, no es su circunstancia, sino que está sólo sumergido en ella y puede en algunos momentos salirse de ella, y meterse en sí, recogerse, ensimismarse y sólo consigue ocuparse en cosas que no son directa e inmediatamente atender a los imperativos o necesidades de su circunstancia. En estos momentos extra o sobrenaturales de

ensimismamiento y retracción en sí, inventa y ejecuta ese segundo repertorio de actos: hace fuego, hace una casa, cultiva el campo y arma el automóvil.

Notemos que todos estos actos tienen una estructura común. Todos ellos presuponen y llevan en sí la invención de un procedimiento que nos permite, dentro de ciertos límites, obtener con seguridad, a nuestro antojo y conveniencia, lo que no hay en la naturaleza, pero que necesitamos. No importa, pues, que en la circunstancia, aquí y ahora, no haya fuego. Lo hacemos, es decir, ejecutamos aquí y ahora un cierto esquema de actos que previamente habíamos inventado de una vez para siempre. Este procedimiento consiste a menudo en la creación de un objeto cuyo simple funcionamiento nos proporciona eso que habíamos menester, el instrumento o aparato. Tales son los dos palitos y la yesca con que el hombre primitivo hace fuego o la casa que levanta y le separa del extremo frío ambiente.

De donde resulta que estos actos modifican o reforman la circunstancia o naturaleza, logrando que en ella haya lo que no hay —sea que no lo hay aquí y ahora cuando se necesita, sea que en absoluto no lo hay. Pues bien; éstos son los actos técnicos, específicos del hombre. El conjunto de ellos es la técnica, que podemos, desde luego, definir, como la reforma que el hombre impone a la naturaleza en vista de la satisfacción de sus necesidades. Éstas, hemos visto, eran imposiciones de la naturaleza al hombre. El hombre responde imponiendo a su vez un cambio a la naturaleza. Es, pues, la técnica, la reacción enérgica contra la naturaleza o circunstancia que lleva a crear entre ésta y el hombre una nueva naturaleza puesta sobre aquélla, una sobrenaturaleza. Conste, pues: la técnica no es lo que el hombre hace para satisfacer sus necesidades. Esta expresión es equívoca y valdría, también, para el repertorio biológico de los actos animales. La técnica es la reforma de la naturaleza, de esa naturaleza que nos hace necesitados y menesterosos, reforma en sentido tal que las necesidades quedan a ser posible anuladas por dejar de ser problema su satisfacción. Si siempre que sentimos frío la naturaleza automáticamente pusiese a nuestra vera fuego, es evidente que no *sentiríamos* la necesidad de calentarnos, como normalmente no sentimos la necesidad de respirar, sino que simplemente respiramos

sin sernos ello problema alguno. Pues eso hace la técnica, precisamente eso: ponernos el calor junto a la sensación de frío y anular prácticamente ésta en cuanto necesidad, menesterosidad, negación, problema y angustia.

Quede aquí esta primera y tosca aproximación a la pregunta: ¿Qué es la técnica? Pero ahora, una vez lograda esa aproximación, es cuando empiezan a complicarse las cosas y a ponerse un tanto

Hemos visto cómo el estadio de evolución técnica en que hoy nos hallamos se caracteriza: 1° Por el fabuloso crecimiento de actos y resultados técnicos que integran la vida actual. Mientras en la Edad Media, en la época del artesano, la técnica y la naturalidad del hombre parecían compensarse y la ecuación de condiciones en que la existencia se apoyaba, le permitía beneficiar ya del don humano para adaptar el mundo al hombre, pero sin que ello llevase a desnaturalizarle, hoy los supuestos técnicos de la vida superan gravemente los naturales, de suerte tal que materialmente el hombre no puede vivir sin la técnica a que ha llegado. Esto no es una manera de decir, sino que significa una verdad literal. En uno de mis libros he destacado, como uno de los datos que el hombre contemporáneo debe mantener más vivaces en su mente, el hecho siguiente: Europa desde el siglo V hasta 1800 —por tanto, en trece siglos— no consigue llegar a más de 180 millones de habitantes. Pues bien, de 1800 a la hora presente, por tanto, en poco más de un solo siglo, ha alcanzado la cifra de unos 500 millones de hombres, sin contar los millones que ha centrifugado a la emigración. En un solo siglo ha crecido, pues, tres veces y media. Y es evidente que cualesquiera sean las causas adyacentes de tan prodigioso fenómeno —el hecho de que hoy *puedan* vivir bien tres veces y media más de hombres en el mismo espacio en que antes malvivían tres veces y media menos—, la causa inmediata y el supuesto menos eludible es la perfección de la técnica. Si ésta retrocediese súbitamente, cientos de millones de hombres dejarían de existir.

La proliferación sin par de la planta humana acontecida en ese siglo es probablemente el origen de no pocos conflictos actuales.

Hecho tal sólo podía acontecer cuando el hombre había llegado a interponer entre la naturaleza y él, una zona de pura creación técnica tan espesa y profunda que vino a constituir una sobrenaturaleza. El hombre de hoy —no me refiero al individuo, sino a la totalidad de los hombres— no puede elegir entre vivir en la naturaleza o beneficiar esa sobrenaturaleza. Está ya irremediablemente adscrito a ésta y colocado en ella como el hombre primitivo en su contorno natural. Y esto tiene un riesgo entre otros: como al abrir los ojos a la existencia se encuentra el hombre rodeado de una cantidad fabulosa de objetos y procedimientos creados por la técnica que forman un primer paisaje artificial tan tupido que oculta la naturaleza primaria tras él, tenderá a creer que, como ésta, todo aquello está ahí por sí mismo: que el automóvil y la aspirina no son cosas que hay que fabricar, sino cosas, como la piedra o la planta, que son dadas al hombre sin previo esfuerzo de éste. Es decir, que puede llegar a perder la conciencia de la técnica y de las condiciones, por ejemplo, morales en que ésta se produce, volviendo, como el primitivo, a no ver en ella sino dones naturales que se tienen, desde luego, y no reclaman esforzado sostenimiento. De suerte que la expansión prodigiosa de la técnica la hizo primero destacarse sobre el sobrio repertorio de nuestras actividades naturales y nos permitió adquirir plena conciencia de ella, pero luego, al seguir en fantástica progresión, su crecimiento amenaza con obnubilar esa conciencia.

2° El otro rasgo que lleva al hombre a descubrir el carácter genuino de su propia técnica fue, dijimos, el tránsito del mero instrumento a la máquina, esto es, al aparato que actúa por sí mismo. La máquina deja en último término al hombre, al artesano. No es ya el utensilio que auxilia al hombre, sino al revés: el hombre queda reducido a auxiliar de la máquina. Una fábrica es hoy un artefacto independiente al que ayudan en algunos momentos unos pocos hombres, cuyo papel resulta modestísimo.

3° Consecuencia de ello fue que el técnico y el obrero, unidos en el artesano, se separasen, y al quedar aislado se convirtiese el técnico como tal en la expresión pura, viviente, de la técnica como tal: en suma, el ingeniero.

Hoy está la técnica ante nuestros ojos, tal y como es, exenta, aparte y sin confundirse y ocultarse en lo que no es ella. Por eso se dedican concretamente a ella ciertos hombres, los técnicos. En la Edad paleolítica o en la Edad Media, el inventar no podía constituir un oficio porque el hombre ignoraba su propio poder de invención. Hoy, por el contrario, el técnico se dedica, como a la actividad más normal y preestablecida, a la faena de inventar. Al revés que el primitivo, antes de inventar sabe que puede inventar; esto equivale a que antes de tener una técnica tiene la técnica. Hasta este punto y aun en este sentido casi material es cierto lo que vengo sosteniendo: que las técnicas son sólo concreciones *a posteriori* de la función general técnica del hombre. El técnico no tiene que esperar los azares y someterse a cifras evanescentes de probabilidad, sino que, en principio, está seguro de llegar a descubrimientos. ¿Por qué?

Esto nos obliga a hablar algo del tecnicismo de la técnica.

Para algunos eso y sólo eso es la técnica. Y, sin duda, no hay técnica sin tecnicismo, pero no es sólo eso. El tecnicismo es sólo el método intelectual que opera en la creación técnica. Sin él no hay técnica, pero con él solo tampoco la hay. Ya vimos que no basta poseer una facultad para que, sin más, la ejercitemos.

Yo hubiera deseado hablar largo y tendido sobre el tecnicismo de la técnica, así actual como pretérita. Es tal vez el tema que personalmente me interesa más. Pero hubiera sido un error, a mi juicio, hacer gravitar hacia él todo este ensayo. Ahora, en su agonía, tengo que reducirme a dedicarle una brevísima consideración: brevísima, pero, según espero, suficientemente clara.

Es incuestionable que ni la técnica habría logrado tan fabulosa expansión en estos últimos siglos, ni al instrumento hubiera sucedido la máquina, ni consecuentemente el técnico se habría separado del obrero, si el tecnicismo no hubiese previamente sufrido una radical transformación.

En efecto, el tecnicismo moderno es completamente distinto del que ha actuado en todas las técnicas pretéritas. ¿Cómo expresar

en pocas palabras la radical diferencia? Tal vez haciéndonos esta otra pregunta: el técnico del pasado, cuando lo era propiamente, es decir, cuando el invento no surgía por puro azar, sino que deliberadamente era buscado, ¿qué es lo que hacía? Pongamos un ejemplo esquemático, por tanto, exagerado, aunque se trata de un hecho histórico y no imaginario. El arquitecto nilota necesitaba elevar los sillares de piedra a las partes más altas de la pirámide de Cheops. El técnico egipcio parte, como no puede menos, del resultado que se propone: elevar el sillar. Para ello busca medios. Para ello, he dicho; es decir, busca medios para el resultado —que la piedra quede en lo alto— tomando en bloque ese resultado. Su mente está prisionera de la finalidad propuesta tal y como es propuesta en su integridad última y perfecta. Tenderá, pues, a no buscar como medios sino aquellos actos o procedimientos que, a ser posible, produzcan de un solo golpe, con una sola operación breve o prolongada, pero de tipo único, el resultado total. La unidad indiferenciada del fin incita a buscar un método también único e indiferenciado. Esto lleva en los comienzos de la técnica a que el medio por el cual se hace la cosa se parezca mucho a la cosa misma que se hace. Así en la pirámide: para subir la piedra a lo alto se adosa a la pirámide tierra en forma de pirámide; con base más ancha y menor declive sobre el cual se arrastran hacia la cúspide los sillares. Como este principio de similitud —*similia similibus*— no es aplicable en muchos casos, el técnico se queda sin regla alguna, sin método para pasar mentalmente del fin propuesto al medio adecuado, y se dedica empíricamente a probar esto y lo otro y lo de más allá que vagamente se ofrezca como congruente al propósito. Dentro, pues, del círculo que se refiere a este propósito, recae en la misma actitud del «inventor primitivo».

Tabla de contenidos

Cuadernos de Beauchef. Volumen VII N°2

EDITORIAL

- *Pensarnos en una existencia poscoronial,*
Nicolás Antonio Rojas Cortés
- *In memoriam, ETHICS*
- *In memoriam, Nicolás Antonio Rojas Cortés*

PENSAR NUESTRA EXISTENCIA ENTRE LA CRISIS PANDEMICA Y LAS TECNOLOGÍAS EMERGENTES

- *Contexto de quiebre, Raúl Villarroel Soto*
- *Breve ensayo sobre la educación líquida o El profesor como un indeseado en el aula, Camilo Vergara R.*
- *Potencialidades y alternativas del transhumanismo en la educación, Alejandro Recio Sastre*
- *Repensar el actuar y la relación en una existencia poscoronial,*
Fabiana Pellegrini
- *Deportistas transhumanos. Repensando el ciclismo de ruta profesional a la luz del transhumanismo.*
Vicente Garrido
- *Pandemia y espacios virtuales. En torno a nuestras concepciones de mundo, Diego Ticchione*
- *Ética aplicada al manejo de datos: Ética de la investigación y riesgos de la inteligencia artificial, Roberto Campos Garro*
- *Sombríos horizontes: Reflexiones en torno a la extinción a partir de la teoría crítica y el poshumanismo, Ricardo Andrade*
- *Simulación y testimonio: A propósito de la posibilidad de conocer a través de otros en el metaverso,*
Felipe Alejandro Álvarez Osorio
- *Retrospectiva de las problemáticas éticas digitales: Vigilancia, anonimato y hostigamiento, Valentina Sola Cereceda Cáceres*
- *Mejoramiento cognitivo: Problemas, desafíos y consideraciones,*
Francisco Hernández J.
- *¿Se puede hablar de un ser humano allende la técnica? Una reflexión en torno al problema moral de la tecnología en relación con los proyectos de vida, Sergio Esteban Ortega Bravo*

- *El rol de las y los estadísticos en el mundo poscoronial,*
Ernesto San Martín

MISCELÁNEAS.

- *Vagabundaje, subjetividad y coexistencia,*
Germán Rozas Ossandón y José Oda Camplá
- *Antiguas placas, antiguos nombres: Crónica de un intento de memoria en las calles de Santiago,*
Carlos F. Navarro Clavería y René I. Larroucau Toro

CLÁSICO

- *La fuerza de los fuertes,* Jack London



Normas de publicación para *Cuadernos De Beauchef*

Envío de textos

[Cuadernos de Beauchef. Ciencia, tecnología y cultura](#) (ISSN: 2452-493X, CC BY-NC-ND 4.0) busca materializar las interrelaciones entre las Humanidades, las Artes, las Ciencias Sociales, la Ciencia y la Tecnología, en la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile

En tal sentido, recibe textos que, desde las disciplinas específicas, aborden aquellas interrelaciones y se consideren relevantes para contribuir a un debate inter y transdisciplinario.

El (la) autor (a) escogerá el formato de escritura que le acomode.

Ensayos, artículos de opinión e informes de investigación deberán presentarse en formato Word, tener una extensión de entre 3.500 a 7.000 palabras y estar escritos en español, usando citas y referencias en estilo APA, sexta edición.

Poemas, prosa poética, reseñas de libros, entrevistas, testimonios, etc., no deben exceder las 3.000 palabras.

Los textos deben ser enviados al correo humanidades@ing.uchile.cl e indicar la siguiente información:

- Nombre(s) y apellido(s).
- Título y/o grado académico, y actividad actual.
- Organismo de pertenencia.
- Correo electrónico.

Proceso de selección

Los escritos recibidos serán evaluados por el Comité Editorial de Cuadernos de Beauchef y/o por algún evaluador/a externo/a a solicitud de dicho Comité. Los/as evaluadores/as podrán:

- a) Aceptar el escrito.
- b) Aceptarlo sujeto a modificaciones (la aceptación definitiva dependerá de que el (la) autor (a) cumpla con las modificaciones propuestas por el Comité).
- c) Rechazarlo.

Los comentarios y sugerencias serán enviados a el (la) autor (a) por el Comité Editorial para que proceda a las modificaciones, las que serán condicionantes para su publicación. El (la) autor (a) dispondrá de tres semanas para efectuar los cambios a contar de la fecha de envío de las evaluaciones.

Una vez finalizado el proceso de corrección deberá enviar al Comité Editorial la versión final de su artículo para su publicación. El Comité responderá con un correo electrónico, indicando que el artículo ha superado el proceso de evaluación y se encuentra a la espera de su publicación, o que ha sido definitivamente rechazado.

Posterior al proceso de selección y eventuales modificaciones del texto original, Cuadernos de Beauchef se reserva el derecho a realizar modificaciones al manuscrito sin alterar su contenido o sus ideas centrales: correcciones ortográficas, clarificación en la redacción o cualquier otro tipo de aspectos asociados a criterios de estilos.

Comité editorial
Cuadernos de Beauchef
ETHICS

ethics

Santiago de Chile
Agosto 2024

