Ética aplicada al manejo de datos: Ética de la investigación y riesgos de la inteligencia artificial

Roberto Campos Garro (†)[1]

El objetivo de la presentación es discutir sobre los riesgos en el manejo actual de los datos producidos por la investigación científica en ciencias sociales y humanas. Estos datos humanos (DH) hoy cuentan con normativas regionales/globales que buscan administrar su protección en vistas de evitar daño, así como también, resguardar aspectos de la integridad científica. Ello ha promovido una ética de los datos que ha dictado lineamientos y principios para su uso responsable. Sostenemos que las actuales protecciones y resguardos pueden resultar insuficientes frente a los usos que la Inteligencia Artificial (IA) puede hacer en el procesamiento de DH. Aquí exponemos dos aplicaciones de IA, de acceso abierto, el Generador de Lenguaje Natural *GPT-3*[2] y el Campo de Radiación

[1] Nota del Editor (N. del E.): este es un texto póstumo y fue leído por primera vez en público en el contexto del Seminario de investigación "Transformación digital e inteligencia artificial: un enfoque ético" organizado por la Universidad Laval, Universidad de Chile y Universidade do Estado de Minas Gerais, realizado el lunes 11 de abril del año 2022 en una maratónica jornada que duró desde las 10:00 hasta cerca de las 17:00. El escrito que usted tiene ante su vista es uno de los pocos que el profesor Campos logró escribir bajo el yugo de su enfermedad. No me parece menor rescatar estas intuiciones que, si bien son incipientes, muestran los primeros fundamentos de un trabajo que él llevaba reflexionando en su dirección del Centro de Estudios de Ética Aplicada (CEDEA). Lamentablemente, luego de su partida, este tipo de investigaciones no se siguió realizando. Y, en lo personal, sin el director que dejó el centro funcionando no valía la pena continuar en un centro que posteriormente no realizaría investigación alguna a lo largo de todo un año. Sé que esto puede molestar a algunos colegas, pero no me importa, porque hay que ser más amigo de la verdad que de los conocidos, y es que en el fondo junto con el profesor Campos también descansa en paz ese centro. Este texto se publica con el permiso de Andrea Muñoz y agradezco la confianza en permitirme editar la obra de nuestro colega.

[2] N. del E.: el profesor Roberto Campos no alcanzó a tener acceso a las versiones 3.5 ni 4.0 del *Chat GTP*.

Neuronal *Instant NeRF*, a modo de ejemplos. Estos algoritmos de aprendizaje automático profundo supondrían un nuevo horizonte de incertidumbre respecto de los riesgos involucrados en el uso de DH, al menos en dos aspectos: la vulnerabilidad de los protocolos de anonimato y la producción/circulación de datos sintéticos. Frente a ello se sugiere tomar precauciones en la gestión de datos de investigación como en repositorios institucionales y alentar el desarrollo de una ética aplicada al diseño de la AI.

1. Ética de la investigación científica en ciencias sociales y humanas

El campo epistemológico de una ética aplicada a la investigación en ciencias sociales y humanas (CSH) ha adquirido una dimensión disciplinar propia en la media que ha sido capaz de distinguirse de la ética de la investigación biomédica, que identificamos con la bioética, pero también por incorporar la validez de un estilo de pensamiento que indaga en las cuestiones humanas ocupando un particular modo de pensar, de ver, escuchar y sentir en la intimidad subjetiva de las actividades individuales y colectivas. En esta praxis del saber el conocimiento resulta siempre ser una construcción interactiva basada en el paradigma de la interpretación.

En una consideración muy general de la ética de la investigación científica en CSH, y adecuada a esta exposición, se puede sostener que dos de sus ejes centrales vienen dados tanto por la protección de los participantes frente a los previsibles daños por su participación, así como por el diseño integro de la investigación.

En el caso de la protección a los participantes se puede considerar que los riesgos proceden de la identificación de los informantes, su exposición en círculos familiares o públicos. Los daños previsibles son maltrato, estigmatización, asilamiento, impugnación, seguimiento, etc. El modo establecido para anular estos perjuicios proviene fundamentalmente de la estrategia seguida por el investigador para asegurar el anonimato y confidencialidad, habitualmente disociando los datos de sus fuentes, como sucede al adoptar pseudónimos, números, etc. Pero también se puede proceder

encriptando la información a través de una herramienta de software. En esta primera etapa es el investigador el que cautela los datos, a través de su aseguramiento en un dispositivo que puede ocultar y proteger, pero también puede transportar y exponer a su extravío o a su robo en ambos casos. La opción que se desaconseja siempre es su almacenamiento en nubes de datos gratuitas, debido a la vulnerabilidad de tales sistemas, así como el acceso abierto a otros integrantes del equipo de investigación. Otro tanto esta, por cierto, en su transmisión vía mensajerías electrónicas.

La opción recomendad es desprenderse de ellos e integrarlos a un repositorio institucional. Estos habitualmente cumplen estándares de protección de mucho mejor modo que el investigador. Volveremos sobre esto más adelante.

En el caso de la integridad de la investigación, por ciento, restringida a la presente exposición, uno de sus objetivos centrales es divulgar sus resultados e ideas fuerza. Mostrar saber lo que otros saben y poder expresar opiniones al respecto es una parte esencial del sistema científico. Lo relevante aquí tiene que ver con asumir que la investigación procede de un esfuerzo que se ha legitimado, evidenciando que en todas sus partes y procesos se ha actuado coherentemente con los valores y convicciones de cada uno de los sostenedores de interés involucrado y estos a su vez con los sistemas de valor adoptados. Es decir, se puede afirmar que una persona es íntegra cuando actúa de acuerdo con los valores que reconoce como válidos. Esto significa que es confiable.

Pero una buena parte de los juicios que se declaran, por parte de los investigadores, respecto de haber adoptado buenas prácticas, sobre todo, respecto del uso de datos resulta cada vez más difícil de demostrar. Y no solo porque es subjetivo mostrar que se ha sido honesto con nuestra auto apreciación sobre lo que hicimos, sino por los modos en que se han gestionado los datos.

2. Ética de los datos

Ahora bien, qué tipo de información es la que contiene los DH. En principio, el producto de los instrumentos y metodología en CSH está constituida por información cualitativa que se procesa como dato. Ejemplos de ellos son los testimonios, narraciones, memorias, juicios de valor, apreciaciones estéticas, opiniones políticas, etc. También pueden albergar registros audiovisuales sobre materialidades, artesanías, cuadernos de anotaciones, bitácoras seguimiento, participación en redes sociales, etc. Últimamente se agregan información sobre la propia información, los metadatos, e incluso otra, relacionada a capturar la interacción que tenemos con nuestros computadores, los paradatos. En muchos de estos casos es posible señalar que la información contenida corresponde a los que se denominan datos personales, pero también a los datos sensibles. Desde el punto de vista de la ética de la investigación, tales contenidos, a través de su exposición pueden, eventualmente, producir daño en los sujetos que los entregaron al investigador mediante un consentimiento informado.

Desde hace un tiempo y en diferentes contextos se ha discutido sobre como entregar protección a los datos personales y sensibles. Expresiones normativas de ello es *El Reglamento General de Protección de Datos* (GDPR) de la Unión Europea que representó, en su momento, el primer paso de una serie de regulaciones que tienen como objetivo establecer principios claros de gobernanza en relación con los datos personales y sensibles.

Tal como señalamos los Repositorios Institucionales ofrecen distintos niveles de protección, atendiendo justamente a nivel de vulnerabilidad que presentan y al riesgo de daño que pueden producir. Varios son los modelos en uso, los que en algunos casos se administran automáticamente, y otros, los menos por ciento, recurren a la presencia de expertos en ética y/o derecho para tomar decisiones sobre su destino.

Ahora bien, en el paradigma actual de Ciencia Abierta, resulta imprescindible contar con datos que también estén abiertos. En

tal sentido el acceso a datos de buena calidad es un factor clave para la implementación de los principios *FAIR* -Ubicables, Accesibles, Interoperables y Reutilizables –por sus siglas en inglés. Los principios *FAIR* están centrados en los datos, promoviendo una mayor facilidad de ubicación, acceso, interoperabilidad y re-uso. Estos principios facilitan el incremento en el intercambio de datos entre diversas entidades. Estas consideraciones sobre el uso de estándares de protección de datos abiertos FAIR, se ha venido progresivamente instalando, regional y globalmente, impulsados por las políticas implementadas por la OCDE *Organisation for Economic Cooperation and Development* (2007), la Comisión Europea eIRG *eInfraestructure Reflexion Group* (2009) y por la NSF *National Science Foundation* (2011).

Se ha señalado, sin embargo, que los principios *FAIR* resultan insuficientes para responder adecuadamente a las exigencias éticas que conlleva el análisis interseccional, en donde los diferenciales de poder y las condiciones históricas asociadas con la recopilación de datos y uso de datos obedecen, en último término, a desigualdades sistémicas de los proveedores, lo que afecta el uso ético y socialmente responsable de los mismos.

Por ello se ha propuesto un enfoque diferente, los principios *CARE*[3], para el cuidado gobernanza de datos indígenas, que apuntan en como los datos deben ser utilizados de manera que se orienten y tengan el propósito de mejorar el bienestar de las comunidades Indígenas. Estos principios buscan sobre los datos su control, responsabilidad, ética y un uso orientado al beneficio de la colectividad (*CARE* en inglés). Se ha dicho reiteradamente que los ecosistemas de datos desfavorecen la inclusión significativa de los derechos sobre los datos e intereses Indígenas, y cuando son involucrados las aportaciones de los Pueblos Indígenas, como hace justamente la investigación en CSH, quedan excluidos de la toma de decisiones que de su análisis se deriva.

Propuestos por la *Research Data Alliance* (2018) –y que incluye a Canada, pero no a Chile- plantea cuestiones que en más de

[3] N. del E.: CREA, en español.

un modo resultan críticas respecto de los principios *FAIR*, y que pueden resultar insolubles. Sin embargo, algunas propuestas de orientación sobre manejo de datos (p. ej. UNESCO), incluyen ambos principios sin conflictividad.

3. **Aplicaciones**

Con todo, queda aún por asumir un siguiente desafío, el que surge de disponer la apertura de datos, a lo que UNESCO OPEN SCIENCE denomina el uso de una nueva generación de herramientas informáticas abiertas para automatizar el proceso de búsqueda y análisis de publicaciones y datos vinculados, lo que permite aumentar la rapidez y la eficacia del proceso de generación y verificación de hipótesis. Estos servicios y herramientas alcanzarán su máxima repercusión si se utilizan en un marco de ciencia abierta que trascienda las fronteras institucionales, nacionales y disciplinarias, teniendo en cuenta al mismo tiempo los riesgos potenciales y las cuestiones éticas que puedan derivarse de la elaboración y la utilización de esas herramientas que utilizan tecnologías de inteligencia artificial.

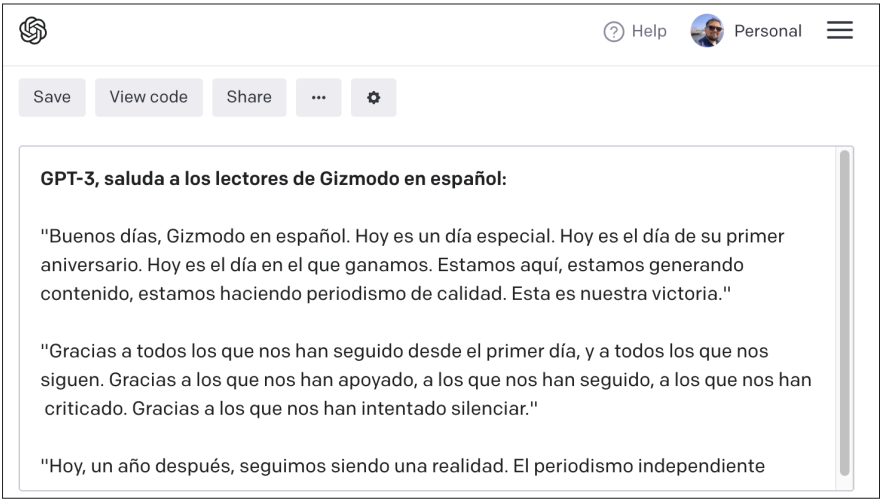
Se ha definido a la AI como una colección de métodos computacionales para estudiar el conocimiento, el aprendizaje y el comportamiento humanos, incluso mediante la creación de agentes capaces de conocer, aprender y comportarse. El desarrollo de la AI está planteando y replanteando la toma de decisiones en contextos nunca vistos. Dos ejemplos de ello son:

GTP-3

El *Generative Pre-trained Transformer 3*, es una interfaz de programación de aplicaciones o API *(Application Programming Interfaces)* que la empresa Open AI puso a disposición del público en 2020. Esta herramienta se puede, según declaran, "aplicar a prácticamente cualquier tarea que implique comprender o generar código o lenguaje natural. Ofrecemos una gama de modelos con diferentes niveles de potencia adecuados para diferentes tareas, así como la

capacidad de ajustar sus propios modelos personalizados. Estos modelos se pueden usar para todo, desde la generación de contenido hasta la búsqueda y clasificación semántica". Para darse una idea de su impacto revisemos una descripción la descripción que hace un banco internacional señala que GPT-3 es un nuevo modelo de inteligencia artificial que permite generar lenguaje escrito. Gracias a este algoritmo, el usuario solo tiene que comenzar a escribir un párrafo y el propio sistema se encarga de completar el resto de la forma más coherente posible. Su gran potencial es una muestra de las posibilidades que existen para llegar a una inteligencia artificial general, capaz de aprender tareas intelectuales como las personas.

Un ejemplo[4] de su capacidad es:



[4] N. del E.: como se dijo antes, el profesor Campos no alcanzó a conocer las versiones actualmente disponibles para nuestro país del Chat GPT. La fuente del ejemplo es la nota de Gizmodo (noviembre 22, 2021): https://es.gizmodo.com/este-articulo-de-gizmodo-fue-escrito-por-una-inteligenc-1848101467

Sin duda que resulta sorprendente poder interactuar, en tu PC, con una herramienta con tanta capacidad para generar textos, solo se requiere entregar para su entrenamiento uno o varios enunciados propios sobre el asunto que interesa tratar, y que a continuación se complete el resto del escrito con la coherencia sintáctica y semántica suficiente para hacerlo indistinguible de una producción humana. Con ella se puede encontrar varias respuestas a cualquier pregunta que le hagamos. Los textos generados pueden ser traducidos y adaptados a diferentes estilos de redacción, como el informe, una noticia, un cuento o una poesía. Por supuesto redactar un *paper* es posible. Se ha dicho que puede crear cualquier cosa que tenga la estructura de un idioma, como resumir una novela, escribir ensayos sobre ética de la AI, generar notas de campo, entrevistas, discusiones, conversaciones grupales. Y también producir, compilar y analizar datos de investigación científica.

Instant NeRF

El *Instant NeRF Neural Radiance Field* o Campo Instantáneo de Radiación Neuronal, es un modelo renderizado neuronal que la empresa Nvidia ha puesto a disposición este 2022. Según su fabricante la estructura del algoritmo procede como una representación inversa y utiliza AI para hacer una aproximación al comportamiento de la luz en el mundo real, lo que permite reconstruir una escena 3D a partir de dotarlo con algunas imágenes 2D tomadas con diferentes ángulos, todo esto casi de modo instantáneo.

Lo que hace el *NeRF* es llenar los espacios en blanco que dejan las tomas independientes, entrenando una red neuronal para reconstruir la escena mediante la predicción del color de la luz que irradia en cualquier dirección, desde cualquier punto del espacio 3D. La técnica puede solucionar los escotomas u oclusiones producidos cuando los objetos que se ven en algunas imágenes están bloqueados o enmascarados. De este modo, la IA hace aparecer, en las fotos, lo que las fotos no han "presenciado" directamente, creando una realidad invisible a la captura en 2D.

4. Nuevo horizonte de riesgos y acción

Tal como es conocido, el desarrollo de tecnologías disruptivas, como es el caso de la que aquí nos referimos, trae consigo una mezcla de escepticismo y paradoja. Escepticismo en tanto que el ingenio creado responda a las expectativas que nos hemos hecho, ocupando analogías sobre experiencias anteriores con otras tecnologías. Sin embargo, me parece que en los casos presentados la IA es real y está dispuesta ante nosotros para su uso. Se dice también que las tecnologías disruptivas encierran a la ética en la paradoja de tener que pronunciarse sobre los efectos de algo que no ha sucedido, pero que una vez instalado no resulta sencillo enmendar los aspectos que han resultado indeseados. Si bien hemos transitado en las técnicas prospectivas de análisis de riesgos, desde modelos basados en extrapolar experiencias conocidas y aceptadas en base a análisis de iteraciones, conocido como enfoque frecuentista, a otros más complejos y adaptativos como sucede con las inferencias bayesianas, que permite ajustar parámetros con mayor precisión predictiva, estas opciones podrían entregar resultados estólidos frente a la envergadura de los efectos e implicancias que los desarrollos de IA antes descritos.

Así visto resulta desafiante la posibilidad de poder a gusto transformar una realidad que ha quedado reflejada en una fotografía prescindiendo de alguna artificialidad para ello. No se la intervenido con agentes extraños, sino que se ha profundizado en la producción de esa misma realidad, generando otro campo, no solo visual, sino que también epistémico. Esta transformación puede representar para la ética de la investigación la necesidad de replantearse las medidas de resguardo en la investigación con fotografías, la que constituye una de las fuentes documentales preferida por las CSH. En el caso específico del trabajo de exploración con grupos humanos, colectivos, familias, estudiantes, entre otros, el registro fotográfico requiere ser controlado y aplicadas todas las medidas necesarias para la protección de quienes ahí aparecen.

En el caso del generador de texto, cabe también preguntarse por cuales serían las medidas a adoptar para identificar situaciones de riesgo que tendría el proveerlo con DH para su entrenamiento, sin

contribuir al mismo tiempo, con más datos que pueden ser usados en la identificación de las personas que contribuyeron a la investigación entregando opiniones, a través de mecanismos conocidos como es de la trazabilidad. Es imaginable que haya o se descubran otras maneras de hacer interactuar ese tipo de datos en vistas de fines semejantes como el de la vigilancia.

También, en el caso de esta herramienta surge legitima pregunta por la autoría que implica su uso, no queda claro si constituye un plagio o una similitud a un nivel indistinguible. Las actuales herramientas para su detección no lo permitirían. Pero quizás el asunto más preocupante proviene de la generación de DH que contienen información sobre realidades, situaciones o memorias no acontecidas. Por supuesto que la circulación de estos datos sintéticos proveería una distorsión de nuestra propia realidad y, de paso, generar por cierto errores y fallas gravísimas, cuando en ello se sustentan acciones políticas o administrativas sobre las personas y la sociedad.

Si estos datos sintéticos no son discutidos en su legitimidad, podemos ir concibiendo acríticamente una IA general como un proveedor omnímodo de información y de hechos, explicaciones, predicciones y resultados científicos.

En otra parte hemos comprado esta situación a la presentada en el cuento de J.L. Borges, *La biblioteca de Babel*, en el que imagina una biblioteca que alberge libros en cuyas páginas se realicen todas las combinaciones que permiten las 26 letras, la como y el punto. Muchos de esos tomos albergarán galimatías, pero otros contendrán las obras fundamentales de la cultura occidental e infinidad de variantes para cada una de ellas. En alguno de estos libros, señala el autor, se podrá encontrar tu autobiografía, pero que sin embargo no haya sido escrita por ti.

Frente a estas nuevas realidades que desafían a la ética aplicada, en este caso a la ética de la investigación científica cabe esperar que la comunidad académica opere con precaución y prontamente plantee acciones acordes con los riesgos que conllevan estos nuevos desafíos.