INTELIGENCIA ARTIFICIAL

Teorías, aplicaciones, futuro

Juan David Gutiérrez
y Rubén Manrique
(edición académica)

INTELIGENCIA ARTIFICIAL

Juan David Gutiérrez
y Rubén Manrique
(edición académica)

INTELIGENCIA ARTIFICIAL

Teorías, aplicaciones, futuro

Universidad de los Andes
Escuela de Gobierno Alberto Lleras Camargo
Ediciones Uniandes

Nombre: Gutiérrez, Juan David. edición académica. | Manrique, Rubén, edición académica.

Título: Inteligencia artificial: teorías, aplicaciones, futuro / Juan David Gutiérrez y Rubén Manrique (edición académica)

Descripción: Bogotá: Universidad de los Andes, Escuela de Gobierno Alberto Lleras Camargo, Ediciones Uniandes, 2025. | 329 páginas: ilustraciones; 17 × 24 cm.

Identificadores: ISBN 978-958-798-844-4 (rústica) | 978-958-798-845-1 (e-book) | 978-958-798-846-8 (e-pub)

Materias: Inteligencia artificial

Clasificación: CDD 006.3–dc23 SBUA

Primera edición: octubre del 2025

Ediciones Uniandes
Carrera 1.ª n.º 18A-12, bloque Tm
Bogotá, D. C., Colombia
Teléfono: 601 339 4949, ext. 2133
http://ediciones.uniandes.edu.co
ediciones@uniandes.edu.co

ISBN: 978-958-798-844-4
ISBN *e-book*: 978-958-798-845-1
ISBN epub: 978-958-798-846-8
DOI: https://doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468

Corrección de estilo: Camilo Sierra Sepúlveda
Diagramación: María Victoria Mora
Diseño de cubierta: Boga Visual
Imagen de cubierta: adaptada por Boga Visual, con base en Yutong Liu & Kingston School of Art / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Impresión:
DGP Editores S. A. S.
Calle 63 n.º 70D-34
Teléfonos: 601 721 7641 - 601 721 7756
Bogotá, D. C., Colombia

Impreso en Colombia – *Printed in Colombia*

Este libro cuenta con el aval de la Escuela de Gobierno Alberto Lleras Camargo y fue sometido a evaluación de pares académicos.

Universidad de los Andes | Vigilada Mineducación. Reconocimiento como universidad: Decreto 1297 del 30 de mayo de 1964. Reconocimiento de personería jurídica: Resolución 28 del 23 de febrero de 1949, Minjusticia. Acreditación institucional de alta calidad, 10 años: Resolución 000194 del 16 de enero del 2025, Mineducación.

CONTENIDO

RECURSOS GRÁFICOS

USO DEL APRENDIZAJE POR REFUERZO PARA EL MANEJO DE COMPORTAMIENTO DESCONOCIDO EN SISTEMAS DE *SOFTWARE* DINÁMICOS

POTENCIAL DE LA INTELIGENCIA ARTIFICIAL EN TELEDETECCIÓN PARA EL DESARROLLO SOSTENIBLE Y LA GESTIÓN AMBIENTAL

INNOVACIONES DE LA WEB SEMÁNTICA PARA LA EDUCACIÓN SUPERIOR

INTELIGENCIA ARTIFICIAL PARA LA PREVENCIÓN DEL ABUSO SEXUAL EN LÍNEA DE LA INFANCIA Y LA ADOLESCENCIA EN COLOMBIA

SISTEMAS DE INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO DE AMÉRICA LATINA Y EL CARIBE

IMAGINARIOS SOCIOTÉCNICOS Y PRÁCTICAS ANTICIPATORIAS EN EL CUBRIMIENTO MEDIÁTICO DE LA INTELIGENCIA ARTIFICIAL Y SU RELACIÓN CON EL ESTADO EN COLOMBIA

INTRODUCCIÓN

Juan David Gutiérrez y Rubén Manrique

ESTE LIBRO REÚNE ESFUERZOS Y DOMINIOS TEMÁTICOS DE investigadores y profesores de la Universidad de los Andes que trabajan en el estudio y desarrollo de la inteligencia artificial (IA). De esta manera, busca aportar desde tres frentes: el avance y la crítica de la teoría sobre la IA; la creación y el análisis de herramientas y aplicaciones basadas en IA, y la reflexión sobre las implicaciones actuales y potenciales de estas tecnologías.

Esta obra colectiva está dividida en tres secciones y comprende diez capítulos, los cuales abordan el estudio de la IA a partir de disciplinas y áreas tan diversas como la ingeniería, la historia, la ciencia de datos, la filosofía, la arquitectura, el diseño, el desarrollo y gestión ambiental, la psicología, la administración, la política y la salud públicas.

La realización de este libro es una iniciativa de la Escuela de Gobierno Alberto Lleras Camargo, que financió su publicación y apoyó su gestión editorial, y de Ediciones Uniandes, que propuso la realización del libro y gestionó su edición. El proyecto apunta a fomentar la misión de la Escuela de Gobierno de articular esfuerzos de investigación interdisciplinarios para que la Universidad de los Andes contribuya al aprovechamiento de oportunidades colectivas y a la solución de problemas complejos de Colombia.

Para presentar este volumen comenzamos por exponer brevemente el objeto de estudio, la IA; luego, explicamos el papel de la Universidad de los Andes en los procesos de transformación digital del país, con énfasis en sus aportes a la teoría y la práctica de la IA; para terminar, en la última sección presentamos cada uno de los diez capítulos que componen este libro.

Qué es la inteligencia artificial

En 1950, el escritor Isaac Asimov publicó un libro de ciencia ficción que describía el desarrollo de "máquinas pensantes" por parte de una empresa denominada U. S. Robots & Mechanical Men Inc. Su obra, *Yo, robot*, es conocida por el planteamiento de las tres leyes de la robótica y por anticipar algunos retos técnicos, sociales, políticos y económicos asociados a la creación y adopción de lo que hoy conocemos como sistemas de IA.

Ese mismo año, Alan M. Turing publicó el ensayo titulado "Maquinaria computacional e inteligencia", en el cual propuso reemplazar el abordaje de la pregunta "¿pueden pensar las máquinas?" con un experimento mental que denominó *juego de la imitación* (luego conocido como *test de Turing*). En síntesis, el juego plantea una situación en la que un interrogador humano tiene la tarea de interactuar con dos interrogados, un humano y una máquina, y debe distinguir cuál es cuál —exclusivamente a partir de las respuestas de los interrogados (Turing, 1950)—.

De esta forma, Turing reemplazó la pregunta sobre si las máquinas pueden pensar por otro interrogante: "¿hay computadores digitales imaginables que tendrían un buen desempeño en el juego de la imitación?". Al final del ensayo, Turing reflexiona sobre cómo podría lograrse que una máquina imitara la mente de un humano: su respuesta inicial es partir por la simulación de procesos educativos, es decir, indagó sobre cómo desarrollar procesos de aprendizaje de máquinas, lo cual, como veremos más adelante, es una de las técnicas contemporáneas más exitosas en el desarrollo de la IA.

El término *inteligencia artificial* fue acuñado en 1955 por John McCarthy y Marvin L. Minsky, quienes plantearon que "el problema de la inteligencia artificial consiste en hacer que una máquina se comporte de un modo que se consideraría inteligente si lo hiciera un ser humano"; su propuesta de investigación partía de la conjetura de que "cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede hacer que una máquina lo simule" (McCarthy *et al*., 1955, p. 2).

Siete décadas después de que la IA fuera fundada como disciplina de investigación, no hay consenso acerca de la definición de su objeto de estudio. Sin embargo, para efectos de este texto introductorio, podemos entender las herramientas de IA como sistemas computacionales que operan a partir de datos y que, con diferentes grados de autonomía, pueden resolver problemas o alcanzar objetivos establecidos por seres humanos (Gutiérrez, 2024).

Los sistemas de IA se construyen a partir de diferentes técnicas y métodos, como la IA simbólica, el aprendizaje automático (p. ej., supervisado, no supervisado, por refuerzo, profundo), el procesamiento de lenguaje natural, entre

otros. Estas metodologías varían en sus principios y enfoques, pero persiguen un objetivo común: emular capacidades cognitivas humanas para resolver problemas complejos y realizar tareas de manera autónoma.

En los primeros días de la informática, los programadores creaban algoritmos para resolver problemas específicos, que luego codificaban en programas ejecutados por computadoras. Estos programas, aunque no poseían inteligencia propia (eran un reflejo de la inteligencia del programador), permitían resolver tareas de manera eficiente. A pesar de años de investigación entre las décadas de los sesenta y los noventa, los algoritmos desarrollados no fueron lo suficientemente buenos para muchas aplicaciones, como el entendimiento del mundo visual o del lenguaje humano.

La solución de los desarrolladores fue recopilar datos para estas tareas y emplear programas de aprendizaje automático —el área más importante de la IA en la actualidad— para crear algoritmos estocásticos a partir de estos. En un programa que "aprende", se especifica cómo utilizar los datos para actualizar los parámetros de un modelo matemático y mejorar su rendimiento en una labor particular (Alpaydin, 2020). El aumento en la disponibilidad de datos de alta calidad, junto con el rápido crecimiento del poder computacional, ha permitido en la actualidad desarrollar modelos capaces de manejar diversas tareas complejas, logrando un desempeño que con frecuencia supera al del ser humano.

Gracias a la IA, potenciada en las últimas dos décadas por el aprendizaje automático, se pueden construir soluciones a problemas tan variados como identificar y categorizar datos (p. ej., reconocimiento facial); detectar patrones, anomalías, valores atípicos (p. ej., detección de riesgo de fraude financiero); predecir futuros comportamientos a partir de hechos pasados y presentes (p. ej., predicción de la conducta de la población en medios de transporte masivos); desarrollar un perfil de un individuo y adaptarse a través del tiempo (p. ej., recomendación asistida en motores de búsqueda de internet); generar contenido a partir de interacción con seres humanos (p. ej., *chatbots*); encontrar soluciones óptimas a un problema (p. ej., optimización de operaciones logísticas); e inferir resultados a partir de modelamiento y simulación (p. ej. reclutamiento de talento humano) (Organización para la Cooperación y el Desarrollo Económico [OECD], 2022).

En este contexto, también se destaca la emergente y disruptiva área de la IA generativa. Se trata de sistemas capaces de crear contenido original a partir de datos existentes. Estos sistemas no solo identifican patrones en grandes volúmenes de información, sino que utilizan estos patrones para generar nuevos datos que imitan y combinan los existentes, innovando de forma constante sobre los datos originales.

Un ejemplo emblemático de la IA generativa son los modelos de lenguaje de gran tamaño (LLM, por sus siglas en inglés) que pueden generar ensayos, programar o mantener conversaciones con un humano. Esta capacidad de producir contenido sintético útil abre nuevas posibilidades en campos como el entretenimiento, la educación, el diseño, entre otros, llevándonos un paso más allá en el desarrollo del potencial de la IA en nuestra vida cotidiana.

En la última década y a nivel global, ha aumentado el porcentaje de organizaciones privadas y públicas que han adoptado diferentes sistemas de IA para apoyar o realizar sus procesos productivos o sus flujos de trabajo. De manera similar, cada vez más individuos utilizan tecnologías basadas en IA en su día a día con fines laborales y personales. Estas tendencias se han acentuado desde finales del 2022, debido al crecimiento de la oferta de herramientas de IA generativa como ChatGPT, Copilot, Gemini, Claude, Meta AI, Grok, DeepSeek, entre otros, que tienen acceso a través de aplicaciones móviles o plataformas web sin que medie contraprestación monetaria.

Recientemente, ha cobrado relevancia el concepto de *inteligencia artificial agente* (*agentic AI*), que hace referencia a sistemas capaces no solo de generar contenido o responder instrucciones, sino también de planear, razonar, tomar decisiones y ejecutar acciones de manera autónoma, para alcanzar objetivos definidos. Estos agentes inteligentes pueden descomponer una tarea compleja en subtareas, buscar información relevante, adaptar sus estrategias en tiempo real y coordinarse con otros agentes o usuarios. Su funcionamiento se apoya en LLM, que proporcionan las capacidades cognitivas necesarias para el procesamiento del lenguaje natural, el razonamiento contextual y la toma de decisiones en entornos dinámicos. Esta evolución expande las fronteras de la IA, más allá de la generación pasiva de contenidos, hacia sistemas proactivos que interactúan con el entorno de forma autónoma y efectiva.

En paralelo, la IA continúa expandiéndose hacia el mundo físico, lo que dio origen a lo que en la actualidad se conoce como *physical AI*. Esta área integra modelos inteligentes con sensores, actuadores y plataformas físicas, permitiendo que los sistemas operen directamente en entornos reales. La IA física habilita a sistemas autónomos —como robots humanoides, vehículos sin conductor, drones o espacios inteligentes—, para percibir su entorno, interpretar situaciones complejas y ejecutar acciones físicas con un alto grado de precisión y adaptabilidad. A diferencia de los modelos digitales que operan de forma exclusiva en entornos virtuales, esta vertiente exige una integración coordinada entre percepción sensorial, razonamiento computacional y control motor. Aunque plantea retos técnicos considerables, como la sincronización en tiempo real o la robustez frente a entornos cambiantes, también abre un campo de

aplicaciones transformadoras en sectores como la robótica de servicio, la automatización industrial y la movilidad inteligente.

El auge de la IA supone nuevas oportunidades y retos. Oportunidades para impulsar el desarrollo humano y retos asociados a usos que exacerban problemas públicos, o que producen nuevos problemas. Como ocurre con toda tecnología, las aplicaciones de diversas herramientas basadas en IA generan tanto efectos negativos como positivos, y dichos impactos tienden a estar heterogéneamente distribuidos en la sociedad (Postman, 2011)[1].

La maximización de los beneficios, la prevención de vulneración a los derechos fundamentales y la debida gestión de los riesgos vinculados con los usos de los sistemas de IA es el resultado de las decisiones que toman múltiples actores a largo del ciclo de vida de estas tecnologías: desarrolladores, comercializadores, responsables del despliegue, usuarios y quienes participan en la gobernanza de las herramientas basadas en IA, entre otros (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [Unesco], 2024).

Por último, vale la pena resaltar que el campo de estudio de estas tecnologías es muy dinámico, no solo por los constantes avances técnicos y las nuevas aplicaciones de los sistemas de IA, sino también por las variadas implicaciones sociales y las consecuencias políticas y económicas asociadas a su implementación en los sectores públicos y privados. En el futuro cercano, lo que entendemos por IA evolucionará a medida que estos sistemas ejecuten nuevas funciones con un mayor grado de autonomía y que estos cambios tecnológicos generen nuevos retos públicos.

Contribuciones de la Universidad de los Andes a la teoría y práctica de la inteligencia artificial

Los aportes de la Universidad de los Andes a la transformación digital en Colombia datan de comienzos de la década de los sesenta. En 1963, la Facultad de Ingeniería instaló el primer computador en una universidad colombiana, el IBM 650, y lo inauguró con un curso sobre "aplicación de computadores al diseño de carreteras" (Aristizábal, 2004, p. 105).

1 Esta idea la expresó el filósofo Neil Postman en la introducción de *Technopoly* (2011), a propósito de su reflexión sobre el intercambio descrito en los *Diálogos* de Platón entre el rey Thamus y el dios Theuth, en los siguientes términos: "Es un error suponer que toda innovación tecnológica tiene un efecto unidireccional. Toda tecnología es a la vez una carga y una bendición; no es una cosa o la otra, sino ambas a la vez" ["it is a mistake to suppose that any technological innovation has one-sided effect. Every technology is both a burden and a blessing; not either-or, but this-and-that"] (pp. 11-12).

Posteriormente, en 1980, Manuel Dávila, un ingeniero de sistemas de la Universidad de los Andes, cofundó la primera empresa de importación de microcomputadores del país y fue pionero en el desarrollo de *software* para empresas (Montes, 2004).

En la segunda mitad de los años ochenta la Universidad realizó los primeros proyectos para conectarse a redes internacionales de computadores, las precursoras de lo que hoy conocemos como la World Wide Web. Luego, en la década de los noventa, la Universidad contribuyó a la difusión del acceso a internet en Colombia (Borja Gómez, 2022).

La Universidad de los Andes también ha sido precursora en el campo de la IA. En noviembre del 2020 fue inaugurado el Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA), el primer centro de investigación académico en esta materia fundado en América Latina. Dentro de los muchos proyectos liderados por CinfonIA se destaca Guacamaya, una colaboración internacional con Microsoft, el Instituto Alexander von Humboldt y el Instituto Amazónico de Investigaciones Científicas SINCHI, que aplica IA para el monitoreo y la conservación de la Amazonía. En este proyecto se han desarrollado algoritmos avanzados capaces de procesar grandes volúmenes de datos bioacústicos, imágenes de cámaras trampa y fotografías satelitales, facilitando un análisis acelerado y preciso de la biodiversidad y la deforestación en la región amazónica (Universidad de los Andes, 2023).

Además, CinfonIA ha establecido una alianza significativa con Google DeepMind, una de las empresas líderes en IA, para fortalecer y diversificar la comunidad que está alrededor de estas investigaciones. A través del programa de becas de Google DeepMind, se han financiado estudios de maestría de estudiantes en la Universidad. Esta iniciativa no solo promueve la excelencia académica, sino que también proporciona recursos y mentoría especializada, apoyando a los becarios en su desarrollo personal y profesional en este campo.

Desde CinfonIA y otras iniciativas lideradas por profesores e investigadores de diferentes facultades, la Universidad de los Andes ha contribuido al estudio y desarrollo de la IA a través de procesos de investigación e innovación, enseñanza, desarrollo institucional y contribución a la agenda pública.

En el 2023, la Universidad obtuvo el registro calificado para la Maestría en Inteligencia Artificial, actualmente ofrecida en modalidad virtual en alianza con Coursera. Esta maestría, la primera en español impartida de forma virtual en la plataforma, ha sido un éxito desde su lanzamiento. Para el primer semestre del 2025, cuenta con cerca de 340 estudiantes inscritos, consolidándose como uno de los programas de posgrado más exitosos de la Universidad.

Además, en el primer semestre del 2024, la Escuela de Gobierno de la Universidad de los Andes entrenó a través de un curso en línea de Educación Continua a 1400 magistrados, jueces y servidores judiciales en fundamentos de IA para la administración de justicia. Este curso en línea de 50 horas (35 sincrónicas, 15 asincrónicas) fue pionero a nivel global tanto por las temáticas desarrolladas como por el número de estudiantes que lo tomaron.

En relación con los aportes de la Universidad a la agenda pública, por ejemplo, desde finales del 2023 la Escuela de Gobierno ha organizado junto con la Universidad Externado de Colombia ocho sesiones de la mesa de trabajo multiactor sobre regulación de IA en Colombia. En dichas sesiones se han abordado temáticas variadas como las implicaciones de la IA para los derechos de propiedad intelectual, la administración de justicia, la protección de los datos personales, la democracia y los derechos humanos.

Además, las investigaciones y aportes de los profesores de la Escuela de Gobierno han sido citados en documentos Conpes de política pública sobre IA, hojas de ruta para el desarrollo y la aplicación de la IA publicadas por ministerios, y sentencias de la Corte Constitucional, como la T-323 del 2024, que resolvió sobre el uso de IA en la administración de justicia, y la T-067 del 2025, sobre transparencia algorítmica en el sector público.

Más recientemente, en marzo del 2025, la Escuela de Gobierno lanzó el proyecto Sistemas de Algoritmos Públicos, que busca contribuir al conocimiento sobre los sistemas algorítmicos utilizados en el sector público, así como a la gobernanza de estas herramientas. En la plataforma en línea del proyecto están disponibles repositorios que documentan casi ochocientas herramientas de IA piloteadas o adoptadas por entidades públicas de América Latina y el Caribe, y casi seiscientos regulaciones y proyectos de regulación relacionados con estas tecnologías en la región (Sistemas de Algoritmos Públicos, 2025)[2].

Por otra parte, en octubre del 2024, la Universidad publicó los *Lineamientos para el uso de inteligencia artificial generativa (IAG) en la Universidad de los Andes*, un documento pionero en América Latina que orienta a estudiantes, profesores, investigadores y empleados administrativos sobre el uso responsable de este tipo de herramientas en actividades pedagógicas, investigativas y administrativas (Universidad de los Andes, 2024).

Por último, en noviembre del 2024, la Universidad importó el primer computador cuántico a Colombia, con la finalidad de que sus estudiantes tengan la oportunidad de profundizar en el aprendizaje e investigación sobre física y computación cuántica (Laguna Cardozo, 2024).

2 Los repositorios pueden consultarse en la plataforma: https://algoritmos.uniandes.edu.co/

Contribuciones a las conversaciones globales sobre la inteligencia artificial

Las contribuciones que componen este libro buscan responder preguntas teóricas y prácticas sobre el conjunto de tecnologías que denominamos IA y sus implicaciones sobre las personas, organizaciones y sociedades. El contenido de este volumen participa en conversaciones globales sobre el desarrollo y las aplicaciones de estas tecnologías, las interacciones entre tecnología y humanidad, y la gobernanza de la IA.

Este libro es útil e interesante para múltiples audiencias. Se estructura en tres secciones. La primera, "Teoría y métodos avanzados de la inteligencia artificial", está dividida en cuatro capítulos. En el primero, Giraldo, Quijano y Manrique investigan la IA distribuida (DAI) y sus aplicaciones para resolver problemas complejos, mediante la interacción de múltiples agentes en entornos compartidos. Así, ofrecen un panorama general de la evolución de la DAI y destacan cómo ha integrado conceptos de diversas disciplinas para abordar la cooperación y la toma de decisiones en sistemas multiagente. Luego, se enfocan en mejorar la resiliencia de comunidades vulnerables a través de estrategias de cooperación financiera, por medio de simulaciones y modelos dinámicos. Finalmente, exponen el uso de modelos de lenguaje de gran tamaño (*large language model*, LLM) y agentes autónomos en la gestión de dilemas sociales, sugiriendo oportunidades para futuras investigaciones.

En el segundo capítulo, Cardozo y Dusparic abordan el desarrollo de sistemas de adaptación dinámica (SAS) y cómo estos pueden ajustarse proactivamente a entornos cambiantes. Analizan la estructura fundamental para construir SAS, que incluye la definición del comportamiento base, la identificación de condiciones para la adaptación y la especificación del comportamiento especializado. Sin embargo, reconocen que los SAS tradicionales tienen una capacidad limitada para adaptarse a situaciones no previstas durante el diseño. Para superar estas limitaciones, los autores proponen mecanismos de aprendizaje dinámico, como Auto-COP y ComInA, que permiten que los SAS generen y compongan adaptaciones en respuesta a situaciones completamente desconocidas, aumentando de manera significativa su flexibilidad y capacidad de respuesta.

Por su parte, en el tercer capítulo, Manrique Gómez y Borja Gómez exploran nuevas formas de hacer investigación histórica a partir de la integración de herramientas de aprendizaje automático. Con ese fin, presentan el estado del arte del procesamiento de archivos multimodales (textos, manuscritos, imágenes y otros), analizan la contribución del proyecto Arte Colonial Americano (ARCA) de la Universidad de los Andes y discuten las implicaciones futuras de la

multidimensionalidad de las fuentes históricas en relación con el quehacer de los historiadores (por ejemplo, nuevas preguntas de investigación y desafíos éticos).

En el último capítulo de la primera sección, Páez examina qué significa que un algoritmo sea transparente y por qué la opacidad (jurídica y epistémica) de los algoritmos puede implicar retos éticos para el desarrollo y uso ético de sistemas de IA. El texto reflexiona sobre los diferentes tipos de opacidad algorítmica, los retos de los métodos de explicabilidad para proveer un mayor grado de comprensión sobre el funcionamiento de algoritmos y sobre la relación entre transparencia y confianza en el contexto de las decisiones automatizadas basadas en sistemas de IA.

La segunda sección del libro, "Aplicaciones y efectos de la inteligencia artificial", está compuesta por cuatro capítulos. Esta inicia en el capítulo cinco, en el cual Núñez Castro, Calderón Romero, Díaz Meza, Sierra Ramírez y Vásquez Pachón investigan el uso de técnicas de IA para el análisis de imágenes satelitales, con el objetivo de mejorar la comprensión de los cambios ambientales y apoyar el desarrollo sostenible y la gestión de recursos naturales. Utilizando imágenes satelitales y técnicas avanzadas de aprendizaje automático, presentan dos aplicaciones de *software* diseñadas para apoyar la toma de decisiones en políticas medioambientales. La primera aplicación es un atlas interactivo para estimar el potencial energético de la biomasa residual poscosecha, mientras que la segunda se centra en la monitorización del crecimiento periférico de ciudades para promover un desarrollo urbano sostenible y planificado.

En el sexto capítulo, Mariño Drews, Paquette y Manrique Piramanrique investigan los beneficios de la web semántica en la educación superior. Discuten las principales ventajas de esta tecnología de IA simbólica en actividades educativas, como el diseño de cursos, la personalización del aprendizaje y el apoyo en la búsqueda de recursos. Además, presentan resultados de más de una década de investigación y desarrollo en este ámbito en la Télé-Université de Quebec y la Universidad de los Andes.

En el séptimo capítulo, Bustamante Duarte, Pajarito Grajales, Portela y Parra Agudelo reflexionan de manera crítica sobre los recientes desarrollos en IA enfocada en aspectos geoespaciales (GeoAI) aplicada al entendimiento de los territorios y comunidades urbanas y rurales del sur global. El capítulo plantea que estas sociotecnologías, pese a diversas ventajas en nuestros contextos de eficiencia y recursos, necesitan ser analizadas cuidadosamente, entendiendo que no siempre reflejan la realidad y necesidades a nivel socioespacial de ciertas comunidades, debido a temas de sus sistemas de clasificación y categorías. De esta manera, la discusión se centra en casos de GeoAI en los que el uso de datos geoespaciales implícitos y participativos ha permitido responder a dichos retos.

En el último capítulo de la segunda sección, Arbeláez, Saldarriaga, Quintero, Castillo, Puentes, Calderón, Osejo, Castañeda, Paz, Hernández y Agudelo investigan cómo la IA puede contribuir a combatir la explotación sexual de niñas, niños y adolescentes en línea (OCSEA). Para ello, desarrollaron dos modelos de IA. El primero se centró en optimizar el análisis de los reportes sobre sextorsión, *sexting*, *grooming* y ciberacoso sexual recibidos por la línea de reporte Te Protejo en Colombia, lo que contribuye a los analistas a clasificar los casos según su gravedad y tipo de daño y reduce el riesgo de exposición a material perjudicial. El segundo modelo se diseñó como un prototipo de sistema de alerta que analiza información en foros de la web profunda, identificando patrones de conversación que podrían indicar actividades de OCSEA y sus posibles efectos en las víctimas.

La tercera sección cierra el libro con dos capítulos sobre "La inteligencia artificial y el Estado". En el noveno capítulo, Gutiérrez y Muñoz-Cadena examinan cómo los sistemas de IA adoptados por entidades públicas en América Latina pueden contribuir con el agendamiento, formulación, implementación y evaluación de políticas públicas. El texto se fundamenta en una nueva base de datos que documenta más de 700 herramientas adoptadas por entidades públicas en veintitrés países de la región y en casos de estudio de Argentina, Brasil, Chile, Colombia, Guatemala, Honduras, México y Perú.

Por último, en el capítulo que cierra esta compilación, Díaz-Valderrama, Niño-Machado, Guerrero-C. y González-Uribe rastrean los imaginarios a futuro sobre la IA y su relación con el Estado, así como las acciones estatales que buscan responder de manera anticipada a la creciente incursión de estas tecnologías en la sociedad colombiana. El capítulo identifica en el discurso mediático cuatro campos de imaginación sociotécnica: (1) innovación vs. regulación, (2) cuarta revolución industrial, (3) primicia, liderazgo y modernización regional, y (4) desarrollo y presencia nacional amplificada.

En conjunto, los capítulos que componen este libro ofrecen nuevo conocimiento sobre la IA a quien esté interesado en aprender sobre sus conceptos teóricos, aplicaciones prácticas e implicaciones presentes y futuras. Esperamos además que cada lector encuentre aquí la chispa que estimule su curiosidad y lo motive a involucrarse de manera crítica y constructiva en el área. Que este libro inspire a pensar más allá, innovar y aplicar de manera ética y responsable los avances de la IA en beneficio de todos.

Referencias

Alpaydin, E. (2020). *Introduction to machine learning*. Massachusetts Institute of Technology Press.

Aristizábal, J. C. (2004). Del 650 al 360: Los primeros computadores de la Universidad de los Andes. *Revista de Ingeniería*, *1*(20), 105-107. https://doi.org/10.16924/revinge.20.14

Borja Gómez, J. H. (2022, 21 de junio). Internet y la web en Colombia: Una historia de sus primeros años. *Credencial Historia*. https://www.banrepcultural.org/biblioteca-virtual/credencial-historia/numero-388/internet-y-la-web-en-colombia-una-historia-de-sus

Gutiérrez, J. D. (2024, 15 de noviembre). De qué hablamos cuando hablamos de IA. *Foro Administración, Gestión y Política Pública* (blog). https://forogpp.com/2024/11/15/de-que-hablamos-cuando-hablamos-de-ia/

Laguna Cardozo, M. (2024, 27 de noviembre). El primer computador cuántico llega a Colombia. *Noticias Universidad de los Andes*. https://www.uniandes.edu.co/es/noticias/ciencias-aplicadas/el-primer-computador-cuantico-llega-a-colombia.

McCarthy, J., Minsky, M. L., Rochester, N. y Shannon, C. E. (1955). A proposal for the Dartmouth Summer Research Project on artificial intelligence. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Montes, A (2004, 30 de mayo). Marzo 3 de 1957: La máquina que cambió al país. *Semana*. https://www.semana.com/especiales/articulo/marzo-1957-brla-maquina-cambio-pais/65917-3/.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco). (2024). *Consultation paper on AI regulation: Emerging approaches across the world*. Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000390979

Organización para la Cooperación y el Desarrollo Económico (OECD). (2022). OECD framework for the classification of AI systems. OECD Digital Economy Papers. https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf

Postman, N. (2011). *Technopoly: The surrender of culture to technology*. Vintage.

Sistemas de Algoritmos Públicos. (2025). *Informe de los repositorios 2.0*. Escuela de Gobierno, Universidad de los Andes. https://sistemaspublicos.tech/wp-content/uploads/Informe-de-Repositorios-Proyecto-SAP-v2.0.pdf

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX* (236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Universidad de los Andes. (2023). "Guacamaya", para salvar la Amazonía con inteligencia artificial. https://uniandes.edu.co/es/noticias/ambiente-y-sostenibilidad/guacamaya-para-salvar-la-amazonia-con-inteligencia-artificial

Universidad de los Andes. (2024). Lineamientos de uso de inteligencia artificial generativa (IAG) en la Universidad de los Andes. https://secretariageneral.uniandes.edu.co/images/documents/lineamientos-uso-inteligencia-artificial-generativa-IAG-uniandes.pdf

PARTE I

TEORÍA Y MÉTODOS AVANZADOS DE LA INTELIGENCIA ARTIFICIAL

INTELIGENCIA ARTIFICIAL DISTRIBUIDA: EVOLUCIÓN Y APLICACIONES

Luis Felipe Giraldo, Rubén Manrique, Nicanor Quijano

Para citar este capítulo:
https://doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.01

Introducción

La inteligencia artificial distribuida (*distributed artificial intelligence*, DAI) involucra la interacción de múltiples agentes dentro de un entorno compartido, para alcanzar objetivos individuales y colectivos. Este paradigma permite modelar y simular dinámicas complejas, ofreciendo una perspectiva única para abordar problemas en una amplia gama de dominios, como la optimización de recursos, el transporte inteligente, la gestión de redes y el estudio de la cooperación y resiliencia en comunidades, lo que contribuye al desarrollo de políticas públicas. La interacción entre agentes facilita la resolución de problemas difíciles o imposibles de tratar mediante enfoques tradicionales, al mismo tiempo que promueve la emergencia de comportamientos colectivos y soluciones adaptables.

En este capítulo se presentan algunos de los desafíos y soluciones propuestas por investigadores de la Universidad de los Andes en este campo, abarcando aplicaciones de la inteligencia artificial (IA) que van desde la teoría de juegos evolutiva y el aprendizaje por refuerzo multiagente, hasta agentes cooperativos basados en los hoy populares modelos de lenguaje de gran tamaño (*large language model*, LLM).

Primero, se ofrece un panorama general de la evolución de la DAI, destacando cómo esta disciplina ha integrado conceptos de la sociología, el control, la economía y la teoría de juegos para abordar problemas de cooperación y toma de decisiones en sistemas multiagente. Esta sección presenta aquellos avances y aplicaciones principalmente lideradas por el profesor Nicanor Quijano, que sustentan el desarrollo de sistemas donde múltiples agentes interactúan en un entorno compartido, toman decisiones autónomas y optimizan su desempeño colectivo.

A continuación, el texto se enfoca en la aplicación de estas nociones para mejorar la resiliencia de comunidades vulnerables, mediante estrategias de cooperación financiera. Se examinan los desafíos que enfrentan las comunidades, como la incertidumbre en los ingresos y la falta de acceso a servicios financieros, y cómo las herramientas basadas en DAI pueden ofrecer soluciones innovadoras. Basados en la investigación liderada por el profesor Luis Felipe Giraldo, se analizan casos específicos de asociaciones de ahorro y crédito, y se evalúa su efectividad por medio de simulaciones de Monte Carlo y modelos dinámicos que permiten optimizar las estrategias de cooperación, según las características de cada comunidad.

Finalmente, se aborda el uso de LLM y agentes autónomos basados en estos modelos, destacando su capacidad para tomar decisiones y cooperar en entornos complejos. A través de experimentos en entornos simulados, liderados por el profesor Ruben Manrique, se evalúa la eficacia de estos agentes en la gestión de dilemas sociales y se resaltan tanto sus capacidades actuales como las áreas que requieren mejoras. Esta sección presenta un análisis de los desafíos y las oportunidades que presentan estas tecnologías, así como caminos para futuras investigaciones que podrían mejorar la cooperación y la resiliencia en una variedad de contextos, desde aplicaciones comunitarias hasta entornos más amplios, como la ciberseguridad y la gestión de infraestructuras críticas.

Evolución de la inteligencia artificial distribuida

La IA se ha inspirado y ha utilizado la psicología y el comportamiento no solo en el desarrollo teórico, sino también a nivel metafórico. Además de las aplicaciones tradicionales de reconocimiento de patrones, imágenes y características, la IA moderna tiene aplicaciones que buscan tomar decisiones basadas en conocimientos adquiridos. Estas aplicaciones modernas de IA se están transformando desde el reconocimiento de características puras (por ejemplo, detectar un gato en una imagen) a la toma de decisiones (conducir vehículos autónomos a través de un cruce de tráfico de manera segura), donde emerge la interacción entre diferentes actores (Yang y Wang, 2020). La aproximación moderna a la IA está centrada alrededor del concepto de *agente racional*. Un agente es un ente computacional (programa) o físico (robot), el cual puede percibir (sensar) y actuar en un ambiente; es autónomo y por lo general su comportamiento dependerá en parte de su experiencia y de la retroalimentación que le brinde el lugar. Esta flexibilidad y racionalidad se logran sobre la base de procesos clave, como la toma de decisiones, la planificación y el aprendizaje. Como este agente interactúa con el ambiente, su comportamiento se verá afectado por otros

agentes o entes humanos, y esta interacción puede ser de índole cooperativa o competitiva (Russell y Norvig, 2016). Un agente que siempre trata de optimizar una medida de desempeño determinada se denomina *agente racional*. Al ser tan general esta definición, se podría ligar a agentes humanos (poseen ojos como sensores y manos como actuadores), agentes robóticos (cámaras y llantas) o agentes de *software* (interfaz gráfica que cumple ambas labores de sensado y actuación). En otras palabras, desde esta perspectiva, la IA se concibe como el estudio de las bases y el diseño de agentes artificiales racionales. Como en la mayoría de los casos estos agentes no están solos e interactúan con otros agentes, estaríamos hablando de sistemas de múltiples agentes (*multi-agent system*, MAS) y de ahí que se derive el subcampo de investigación de la DAI (Vlassis, 2022).

La DAI se podría definir como el estudio, construcción y aplicación de MAS, por ejemplo, sistemas en los que diversos agentes inteligentes interactúan, en busca de un propósito común o lograr una serie de tareas definidas (Weiss, 1999). Estos múltiples agentes por lo general son heterogéneos (p. ej., han sido diseñados y concebidos en *hardware* de diferentes formas); interactúan en ambientes dinámicos (en su mayoría, un solo agente suele ser entrenado en un ambiente estático, pero al interactuar con otros su comportamiento varía con el tiempo y, por ende, se requiere un mejor desarrollo de la parte matemática [Gómez *et al.*, 2022]); la información que se tiene de los sensores es distribuida (p. ej., a nivel espacial o temporal), lo que hace que el mundo percibido por cada agente sea parcialmente observable; el control de los agentes es descentralizado (p. ej., cada agente toma su propia decisión y no está atado a un ente central, debido a los problemas de robustez y tolerancia a fallos; en este caso, los problemas de toma de decisiones pueden resolverse mediante la teoría de juegos); y la interacción entre agentes depende en gran medida de la comunicación que puede facilitar la coordinación y la cooperación entre individuos, como se haría en la naturaleza (p. ej., la forma en la que las abejas se comunican entre sí a la hora de buscar alimento [Quijano y Passino, 2010; Giraldo *et al.*, 2015]) (Vlassis, 2022).

Las primeras investigaciones en DAI datan de los años setenta y ochenta, cuando se emplearon modelos computacionales para simular MAS, combinando elementos de la teoría de juegos, la simulación de Monte Carlo, la programación evolutiva y las teorías de la emergencia y los sistemas complejos. La DAI es un área que toma ideas, conceptos y resultados de disciplinas como la IA, las ciencias de la computación, la sociología, la economía, la filosofía, el manejo de las organizaciones, los sistemas de control, entre otros. Así, se podría decir que la IA tradicional utiliza la psicología y el comportamiento para sus ideas, su inspiración y como metáfora; la DAI utiliza la sociología, la economía y el control, por lo que se percibe como una generalización de la IA. Hay dos razones para

tratar con DAI: (1) los MAS desempeñan un papel fundamental en los sistemas tecnológicos de hoy en día. Los sistemas de cómputo y control que encontramos desplegados en diferentes ámbitos tienen como características ser distribuidos y heterogéneos. Por lo tanto, se requiere ver a estos elementos como agentes en lugar de partes, con el fin de interactuar de una manera más adecuada, sobre todo en esta era de la cuarta revolución industrial; y (2) los MAS tienen la capacidad de interactuar dentro de lo que llamaríamos *cyber-physical human systems* (CPHS), con lo cual hay que pensar en el desarrollo de sistemas autónomos que interactúen no solo en su entorno, sino que tengan en cuenta al humano o los seres vivos con los que lo hace (Annaswamy *et al*., 2023).

Como se puede observar, hay varios conceptos teóricos que forman parte de la idea general de DAI. En primera instancia, podríamos hablar de *cooperación*. Los problemas de cooperación, en los que los agentes tienen oportunidades de mejorar su bienestar común, pero son notablemente capaces de hacerlo, son omnipresentes e importantes. Es posible encontrarlos en todas las escalas, desde nuestras rutinas diarias, como conducir, programar reuniones y trabajar en colaboración, hasta nuestros retos globales, como el experimentado recientemente a la hora de prepararnos para pandemias (Dafoe *et al*., 2020). Otros trabajos, como el del politólogo Robert Axelrod (1984), muestran cómo esta cooperación emerge en áreas como la política, lo cual fue uno de los elementos que nos salvó de tener conflictos de mayor envergadura durante la época de la Guerra Fría. La investigación de la IA relacionada con la cooperación se ha llevado a cabo en muchas áreas diferentes, dentro de las que se incluyen los MAS, la teoría de juegos y la elección social, la interacción y alineación entre el ser humano y la máquina, el procesamiento del lenguaje natural y la construcción de herramientas y plataformas sociales.

La teoría de juegos es el nombre que se le da a la metodología que utiliza herramientas matemáticas para modelar y analizar situaciones en las que se toman decisiones de forma interactiva. Estas nociones se habían comenzado a discutir a finales del siglo XIX, principios del siglo XX, pero es hasta el desarrollo del teorema del minimax, por parte de John von Neumann y el libro que publica con Oskar Morgenstern (Von Neumann y Morgenstern, 2007), que se concretan las ideas iniciales. La diseminación del concepto toma varios años, y es gracias a los aportes de la corporación RAND que estas nociones logran permear diferentes áreas del conocimiento (Bhattacharya, 2021). La investigación en *machine learning* (ML) y aprendizaje por refuerzo (*reinforcement learning*, RL) se ha centrado en casos de conflicto de intereses y, en particular, en entornos con dos jugadores de suma cero (Dafoe *et al*., 2020). En el ámbito del aprendizaje por refuerzo basado en juegos, en los últimos años se han visto enormes avances

en los juegos de suma cero de dos jugadores, como el ajedrez, Go, StarCraft II y póquer de dos jugadores. Este tipo de juegos fueron un dominio productivo para la investigación inicial de múltiples agentes, ya que son especialmente tratables: la solución minimax coincide con el equilibrio de Nash y se puede calcular en tiempo polinomial a través de un programa lineal, sus soluciones son intercambiables y tienen garantías de peor caso (Von Neumann y Morgenstern, 2007). Esta tratabilidad puede explicar por qué estos juegos han recibido una considerable atención de investigación, a pesar de ser relativamente raros en el mundo real y en el espacio de posibles juegos.

Otro de los aspectos relevantes de los DAI es el problema de toma de decisiones. Este problema está ligado al área de control óptimo[1]; en este punto es donde la programación dinámica desempeña un papel fundamental (Vlassis, 2022). Varios ejemplos se dan hoy en cuanto a problemas asociados a toma de decisiones. Por ejemplo, en vehículos autónomos (VA) se tienen dos grandes problemas: (1) en cada instante de tiempo, durante la toma de decisión, el agente no solo debe considerar sus acciones presentes, sino también las consecuencias de las acciones futuras; y (2) para determinar las decisiones correctas y seguras, se deben tener en cuenta las acciones y los comportamientos de los demás agentes y elementos en el ambiente en el que se desenvuelven (Yang y Wang, 2020; Detjen-Leal *et al*., 2023). La necesidad de tener un marco adaptativo de toma de decisiones, además de la complejidad de tener interacción con múltiples aprendices, lleva al desarrollo del aprendizaje por refuerzo de múltiples agentes (*multi-agent reinforcement learning*, MARL). Esta noción busca resolver problemas de toma de decisiones secuenciales con agentes inteligentes, que operan en un ambiente compartido con otros agentes y con cierta incertidumbre y estocasticidad, donde cada ente buscará maximizar sus recompensas mediante la interacción con el ambiente y otros agentes. Históricamente, el mecanismo de RL se desarrolló basado en el estudio del comportamiento de los gatos en una caja (Thorndike, 1898). En 1954, Minsky propuso por primera vez el modelo computacional de RL en su tesis de doctorado y nombró su máquina analógica resultante la calculadora estocástica de refuerzo neural-analógico. En 1961, Minsky sugirió por primera vez la conexión entre la programación dinámica (Bellman, 1952) y el RL (Minsky, 1961). Más adelante, Bertsekas y Tsitsiklis (1996) propusieron métodos aproximados de programación dinámica

1 R. Bellman es el creador de la programación dinámica, mientras que Pontryagin es uno de los pioneros en el cálculo de variaciones. Estas nociones son la base del control óptimo, en cuya área autores como D. Bertsekas y J. Tsitsiklis han realizado aportes fundamentales.

fundamentados en redes neuronales, lo cual muestra de forma clara la relación entre las nociones de aprendizaje y control. En el dominio de control clásico se han estudiado extensamente los enfoques basados en modelos en los que el agente de aprendizaje primero construirá un "modelo" explícito de espacio de estado, para comprender cómo funciona el entorno en términos de dinámica de transición de estado y función de recompensa, y luego aprenderá del "modelo". La ventaja de los algoritmos basados en modelos reside en el hecho de que a menudo requieren muchas menos muestras de datos del entorno. En principio, la comunidad MARL ha trabajado con enfoques basados en modelos, por ejemplo, el famoso algoritmo R-MAX, de hace casi dos décadas. Sorprendentemente, los desarrollos en la línea de sistemas basados en modelos se detuvieron desde entonces; teniendo en cuenta los impresionantes resultados que estos enfoques han demostrado en las tareas de RL de un solo agente, los métodos de MARL merecen más atención de la comunidad (Yang y Wang, 2020).

Por otro lado, la toma óptima de decisiones por parte de múltiples agentes que interactúan entre sí por medio de una red (sea física o de comunicaciones) ha sido uno de los temas que más auge ha tenido en la comunidad de sistemas dinámicos y de control en las últimas décadas (Obando *et al*., 2024). ¿La razón? Estos problemas emergen en ingeniería, ciencias sociales y económicas, sistemas urbanos o inteligencia artificial, donde se encuentran aplicaciones como el análisis de redes sociales (Jackson, 2008), el manejo o control de redes inteligentes (Mojica-Nava *et al*., 2013; Ananduta *et al*., 2018), las redes inalámbricas (Han *et al*., 2019), la ciberseguridad (Pawlick y Zhu, 2021), la infraestructura crítica (Rass *et al*., 2020) o los sistemas ciberfísicos (Groot *et al*., 2014).

Una de las maneras de modelar estos sistemas complejos de gran escala con múltiples decisiones y acciones, en la que los diferentes entes/controladores interactúan entre sí, es mediante el uso de la teoría de juegos (Muros, 2021; Carrasco Martínez *et al*., 2023). Dentro de los trabajos desarrollados, es pertinente mencionar el de Bacci *et al.* (2016), el cual muestra la relación entre los juegos, la optimización y el aprendizaje para el procesamiento de señales en red. Otros ejemplos similares incluyen la carga de vehículos eléctricos (Grammatico *et al*., 2016), la coordinación de redes de robots (Jaleel y Shamma, 2020; Park y Barreiro-Gómez, 2023), los problemas de congestión vehicular (Kara *et al*., 2022), el control de pandemias y epidemias (Martins *et al*., 2023), las técnicas de aprendizaje por refuerzo (Gao y Pavel, 2021), la respuesta a la demanda (Genis-Mendoza *et al*., 2022) y el manejo de recursos y regulación de sistemas de agua (Lu *et al*., 2022). En estos trabajos, los autores abordan los problemas desde diferentes ángulos de la teoría de juegos. Algunos parten de los *juegos matriciales*, lo cuales son conocidos por su forma normal, en la que la interacción simultánea entre jugadores

se produce de manera estática y cada jugador se entiende como un ente individual. Por otra parte, en *juegos continuos* los jugadores pueden elegir entre una amplia gama de estrategias, las cuales cambian con el tiempo.

Hay otros llamados *juegos dinámicos*, que usan un mecanismo de aprendizaje que permite ajustar las acciones basadas en eventos previos. Estos juegos se distinguen por tres problemas principales: (1) modelar el ambiente en el que interactúan los jugadores; (2) modelar los objetivos que persiguen los jugadores; y (3) describir el orden en el que los jugadores toman decisiones y la cantidad de información que tienen. En este caso, se asume que la interacción ocurre entre un gran número (desconocido) de jugadores, y lo que nos interesa estudiar es la proporción de individuos que utiliza una estrategia u otra. Los *juegos evolutivos*, que fueron creados con base en el comportamiento ecológico, se clasifican como juegos dinámicos. Las nociones de las estrategias evolutivamente estables fueron desarrolladas por Maynard-Smith y Price (1973), quienes fueron los pioneros de este concepto. Después, Taylor y Jonker (1978) crearon el modelo de replicadores (*replicator dynamics*), que se utiliza ampliamente en aplicaciones de ingeniería, para examinar el comportamiento dinámico y su relación con la parte genética. De igual manera, mediante la introducción de protocolos de revisión y *mean dynamics*, se han desarrollado aproximaciones de juegos evolutivos desde el punto de vista de sistemas económicos (Sandholm, 2010).

El trabajo de Obando *et al.* (2024) presenta algunos ejemplos de asignación dinámica de recursos a través de juegos poblacionales y modelos dinámicos de pago (Park *et al.*, 2019). Esta investigación destaca la utilidad y la idoneidad de estas técnicas para modelar dinámicas de sistemas complejos de ingeniería, así como para diseñar estrategias de gestión y control, de acuerdo con políticas particulares y restricciones físicas y operativas, tanto locales como globales. Las estrategias desarrolladas por medio de este paradigma son de fácil implementación física, como se observa en distintos trabajos (Martínez-Piazuelo *et al.*, 2022a; J. Barreiro-Gómez *et al.*, 2021), en los que se muestran criterios para seleccionar parámetros y su implementación. Por otra parte, este nuevo paradigma también es capaz de abarcar problemáticas como los retrasos (Obando *et al.*, 2016; Park y Leonard, 2021), lo que ofrece una nueva alternativa respecto a otras técnicas desarrolladas, cuya implementación tiene sus dificultades. Algunas ideas que se presentan en ese artículo se derivan de la evolución de trabajos anteriores; se puede encontrar un resumen de los trabajos que se presentaron hasta el 2017 en Quijano *et al.* (2017). Desde entonces, se han hecho contribuciones en términos de dinámicas distribuidas en tiempo continuo (Barreiro-Gómez *et al.*, 2017a) y en tiempo discreto (Martínez-Piazuelo *et al.*, 2022a), así como en la combinación de técnicas en sistemas híbridos (Ochoa *et al.*, 2021). De igual modo, se ha trabajado

en temas como aplicaciones para vehículos autónomos no tripulados (Barreiro-Gómez *et al*., 2021) o la combinación de técnicas de control para aplicaciones en redes de agua (Barreiro-Gómez *et al*., 2017b; Obando *et al*., 2022). Además, se han reportado recientemente resultados significativos sobre la relación entre estas poblaciones dinámicas y los equilibrios generalizados de Nash (Martínez-Piazuelo *et al*., 2022b, 2022c, 2022d, 2023; Sánchez-Amores *et al*., 2023).

Inteligencia artificial distribuida para resiliencia comunitaria

Las personas con ingresos muy bajos que pertenecen a comunidades desfavorecidas enfrentan un desafío significativo al gestionar su vida financiera. Muchas de ellas son trabajadores por cuenta propia y forman parte del 10,7 % de la población mundial que sobrevive con menos de USD 1,90 al día. Estas personas enfrentan continuamente eventos negativos e impredecibles, llamados *shocks*, como problemas de salud o condiciones climáticas adversas en los cultivos. Además, se ven afectadas por la falta de servicios de salud y financieros, y por una escasa formación educativa. Los desafíos de la gestión financiera para individuos de bajos ingresos y las herramientas utilizadas en la actualidad para enfrentar la escasez giran en torno a tres aspectos principales: manejar ingresos inciertos, resistir *shocks* financieros y encontrar estrategias efectivas para ahorrar dinero (Collins, 2009).

El alcance de las instituciones de microfinanzas, que podrían ayudar a mitigar algunos de estos desafíos, sigue siendo limitado, al igual que el uso de la tecnología para brindar asesoramiento financiero. En respuesta a esta situación han surgido estrategias de cooperación financiera informal en diversas partes del mundo, lo que ha fortalecido la resiliencia de las comunidades de bajos ingresos, al fomentar el apoyo mutuo entre sus miembros. Estas estrategias, basadas en planes de ahorro y crédito, son relativamente fáciles de implementar y ofrecen a los participantes una medida de estabilidad financiera. Ejemplos de estos esquemas de cooperación incluyen la asociación rotativa de ahorro y crédito (*rotating savings and credit associations*, ROSCA) y la asociación de ahorro y crédito acumulativo (*accumulating savings and credit associations*, ASCA) (Bouman, 1995; Zambrano *et al*., 2023). En una ROSCA, un grupo de personas acuerda reunirse de manera periódica para contribuir con una cantidad fija de dinero a un fondo común. En cada reunión, uno de los miembros recibe la totalidad del fondo, lo que le proporciona una suma significativa de dinero en ese momento. Este proceso continúa hasta que todos los miembros hayan recibido el fondo en alguna de las rondas. En algunas regiones de Colombia, esta estrategia de cooperación es llamada *cadena* o *natillera* (Salas Bahamón, 2022). Por

otro lado, una ASCA es una forma de ahorro y crédito más estructurada y compleja que una ROSCA; a diferencia de la rotación simple de fondos que ocurre en una ROSCA, en una ASCA los fondos se acumulan y se utilizan para otorgar préstamos con intereses a los miembros del grupo.

En comunidades de bajos ingresos, donde los servicios bancarios son escasos o inaccesibles, las ROSCA y las ASCA ofrecen una manera de participar en actividades financieras sin necesidad de cumplir con requisitos complicados, como historial crediticio, garantías o altos depósitos iniciales. "No me gusta tener que lidiar con otras personas por dinero, pero si eres pobre, no hay alternativa. Tenemos que hacerlo para sobrevivir" (Collins, 2009, p. 13), señala uno de los participantes de estas estrategias de cooperación, resaltando la importancia de estos esquemas en comunidades vulnerables.

Aunque estos esquemas de cooperación son útiles para promover el desarrollo socioeconómico de la comunidad, aún tienen limitaciones para prevenir fallos de gestión cuando varios miembros se ven afectados por *shocks* o cuando alguien decide dejar de participar una vez la asociación está implementada. El diseño de nuevas estrategias de cooperación que minimicen estas limitaciones implicaría varios meses de observación *in situ*, retroalimentación y un proceso de rediseño. Ante esta necesidad, se han propuesto herramientas computacionales basadas en DAI para proporcionar, en un periodo relativamente corto de tiempo, una mejor comprensión del comportamiento de una comunidad que implementa estas estrategias de cooperación y facilitar el diseño de nuevas estrategias o variaciones de estas, minimizando tales desventajas en una amplia variedad de escenarios y tipos de comunidades.

González Villasanti *et al*. (2018) presentan el resultado de una investigación conjunta entre investigadores de la Universidad de los Andes, en Colombia, y la Universidad Estatal de Ohio, en Estados Unidos. Este es un trabajo seminal que marca una línea de estudio en el área, donde se propone un modelo basado en la teoría de control óptimo, que caracteriza la vida financiera de individuos y su participación en asociaciones informales de ahorro y crédito, con el fin de estudiar estrategias optimizadas de gestión de recursos a través de simulaciones computacionales. Primero, se introduce un modelo de toma de decisiones que describe el comportamiento de un individuo en términos de riqueza, salud y educación. Este modelo se basa en la gestión financiera personal y asume que los individuos tienen la capacidad de tomar decisiones que dependen de sus intereses y prioridades, su capacidad para generar riqueza, su condición de salud, su educación actual y eventos inesperados, como *shocks* financieros. Esta toma de decisiones se realiza mediante proyecciones, las cuales se asume que el individuo puede realizar de acuerdo con un horizonte de tiempo. Luego, los

individuos se interconectan para crear una comunidad heterogénea, donde interactúan según una estrategia cooperativa.

En este trabajo se estudiaron dos estrategias de cooperación financiera: la ASCA y una nueva estrategia basada en donaciones. Así, se introdujeron indicadores como la tasa de fracaso en la gestión y el índice de desarrollo comunitario, para cuantificar el rendimiento de la comunidad en términos de las tres dimensiones que describen el desarrollo humano (riqueza, salud y educación). A través de simulaciones de Monte Carlo y estos indicadores de desempeño y resiliencia comunitaria, se evaluaron una gran cantidad de escenarios y poblaciones. Las simulaciones muestran que la ASCA presenta un mejor desempeño que la estrategia de donaciones en comunidades con baja desigualdad en activos y habilidades, mientras que la estrategia basada en donaciones muestra mejores resultados en comunidades desiguales. Estos resultados sugieren que estas estrategias deberían adaptarse a cada grupo para maximizar los impactos positivos.

Esta investigación fue el punto de partida para estudios posteriores liderados por investigadores de la Universidad de los Andes y financiados por Google Research, debido a su pertinencia e impacto. El trabajo propone dos variaciones de las ROSCA convencionales, que potencialmente aumentan la resiliencia de la comunidad (Zambrano *et al*., 2022). Estas estrategias fueron evaluadas por medio de herramientas computacionales. Basado en el trabajo de González Villasanti *et al*. (2018), se emplearon modelos dinámicos para simular escenarios donde se implementan estas estrategias de cooperación financiera. Por un lado, se formula una estrategia que involucra una reserva de efectivo, similar a la usada en instituciones financieras formales, para reducir el impacto de los individuos que abandonan la asociación. Como resultado, se muestra que la ausencia de confianza entre los miembros de la asociación debe ser compensada en igual medida por la cantidad de reserva de efectivo, con el fin de reducir la probabilidad de fracaso en el esquema de cooperación. Por otro lado, se evalúa el desempeño de comunidades en las que sus miembros participan en varias ROSCA al mismo tiempo, visto como una estrategia de participación descentralizada. Se demostró que los individuos deben cooperar en varias asociaciones con un bajo número de miembros, con el objetivo de reducir el tiempo necesario para recibir un retorno de su cooperación y disminuir la probabilidad de perder su dinero, debido a individuos no confiables. Estos resultados son de particular interés en países como Colombia, donde los esquemas financieros de este tipo se consideran ilegales cuando la cantidad de participantes supera un umbral.

Este cuerpo de investigación ha sido fundamental para entender y abordar los desafíos financieros a los que se enfrentan las comunidades de bajos ingresos. Al utilizar herramientas computacionales avanzadas y modelos dinámicos de

simulación, no solo se profundiza en la comprensión de las estrategias de cooperación financiera existentes, sino que también se proponen nuevas soluciones que pueden fortalecer la resiliencia comunitaria ante *shocks* económicos. La relevancia de este trabajo radica en su potencial para influir en políticas públicas y en el diseño de intervenciones más efectivas, las cuales mejoren la estabilidad financiera y el desarrollo humano en contextos vulnerables. La investigación en curso, respaldada por entidades como Google, promete avances significativos en la teoría y práctica de la cooperación financiera, y tiene el potencial de transformar la vida de millones de personas que dependen de estos esquemas para su subsistencia y bienestar.

En la actualidad, investigaciones en curso en la Universidad de los Andes, financiadas por Google a través del Google Research Award y el Google DeepMind Scholar Programme, buscan avanzar en este trabajo mediante el uso de MAS basados en DAI. Estas investigaciones implementan estrategias de aprendizaje avanzadas, como el MARL y el LLM, dentro de un marco de IA cooperativa. El objetivo es comprender cómo se pueden construir comunidades resilientes en escenarios complejos e inciertos, donde no solo interactúan agentes humanos entre sí, sino también humanos con máquinas y máquinas con otras máquinas.

El impacto de la DAI en políticas públicas está en desarrollo y se espera que sus efectos se perciban a medida que se comprendan mejor estas propuestas basadas en modelos computacionales. Los resultados preliminares de esta investigación han contribuido al diseño de intervenciones en comunidades de bajos ingresos, donde estrategias de ahorro y cooperación, como las ROSCA, han mostrado potencial para mejorar la estabilidad financiera. Así, se anticipa que estas estrategias de inclusión financiera generen recomendaciones de política que fortalezcan la resiliencia comunitaria a nivel local e internacional en el futuro.

Inteligencia artificial distribuida basada en LLM

Los LLM son algoritmos de IA entrenados en grandes corpus de texto para predecir, generar y manipular el lenguaje humano. Estos modelos se destacan por su capacidad para entender contextos y producir texto coherente. Un ejemplo destacado es GPT-4, que ha demostrado que puede generar respuestas detalladas y relativamente precisas a partir de un amplio rango de preguntas (OpenAI *et al.*, 2024). El desarrollo continuo de los LLM ha permitido que estos modelos se apliquen en tareas de procesamiento de lenguaje natural (PLN) y en dominios más complejos, que requieren comprensión semántica y generación de textos a partir de instrucciones abstractas (Broekhuizen *et al.*, 2023). Así, estas capacidades los posicionan como herramientas clave para aplicaciones que van desde *chatbots*

hasta la generación automática de contenido. Sin embargo, un reto inherente en el uso de LLM es que pueden generar respuestas basadas en información incorrecta o incompleta, lo que afecta su desempeño en tareas específicas. Esto ha llevado a investigaciones para mejorar la precisión y fiabilidad de estos modelos mediante la integración de módulos extra, como las memorias a corto y largo plazo (Schick *et al*., 2023; Yao *et al*., 2023).

Los agentes autónomos basados en LLM (*augmented autonomous agents*, LAA) son una extensión de los LLM que se emplean para ejecutar tareas de forma autónoma dentro de entornos específicos. Estos agentes no solo generan texto de manera coherente, sino que también toman decisiones basadas en entendimientos previos y observaciones del entorno (Park *et al*., 2023). Para funcionar eficazmente, los LAA requieren de arquitecturas que gestionen de manera efectiva la memoria y las capacidades cognitivas complejas. Estas arquitecturas suelen incluir módulos para la percepción, la planificación, la reflexión y la acción. Por ejemplo, el *framework* Generative Agents proporciona un robusto armazón que permite a los LAA almacenar recuerdos a largo plazo, reflexionar sobre experiencias pasadas y planificar acciones futuras con base en esa información (Park *et al*., 2023). Este enfoque ha sido útil para crear comportamientos humanos convincentes en entornos simulados y ha facilitado la navegación autónoma en videojuegos como Minecraft (Wang *et al*., 2023). A medida que se exploran más áreas de aplicación, se revela que los LAA pueden imitar comportamientos humanos realistas y manejar interacciones complejas; sin embargo, su capacidad para colaborar eficazmente con otros agentes aún está en fase de investigación (Du *et al*., 2023; Zhang *et al*., 2023).

La cooperatividad entre agentes es una habilidad crítica para resolver problemas en entornos complejos. Esta capacidad se deriva de la comprensión mutua y la toma de acciones colaborativas. En el ámbito de los LAA y la DAI, la cooperación se da cuando múltiples agentes colaboran para alcanzar un objetivo común, utilizando el conocimiento de su entorno y las habilidades adquiridas durante su entrenamiento (Gross *et al*., 2023). Para que la cooperación entre LAA sea efectiva, es fundamental que estos agentes puedan comunicarse y coordinarse entre sí. Esta habilidad está influenciada por sus arquitecturas, que deben incluir módulos para la comunicación y el entendimiento de los otros agentes. Por ejemplo, un agente necesita ser capaz de interpretar las intenciones y acciones de otros agentes, así como comunicarse con ellos para lograr objetivos comunes (Dafoe *et al*., 2020a). A pesar de los avances, los estudios muestran que aunque los LAA tienden a cooperar, sus acciones no siempre reflejan una comprensión clara de la colaboración efectiva dentro del entorno; esto subraya la necesidad de arquitecturas más robustas para mejorar su capacidad de colaboración (Agapiou *et al*., 2023).

En el estudio "Can LLM-augmented autonomous agents cooperate? An evaluation of their cooperative capabilities through Melting Pot" (Mosquera *et al*., 2024), desarrollado por investigadores de la Universidad de los Andes, se utilizaron entornos del proyecto Melting Pot para evaluar la capacidad de cooperación de los LAA como avance fundamental de DAI. Estos entornos están diseñados para simular escenarios de dilemas sociales, donde agentes independientes deben tomar decisiones que afectan el bienestar individual y colectivo. El experimento clave utilizó el escenario Commons Harvest, que se desarrolla en una cuadrícula, donde los agentes deben recolectar manzanas. Los agentes que recolectan de manera insostenible pueden agotar los recursos, lo que refleja la "tragedia de los comunes" (Agapiou *et al*., 2023). Los LAA en este entorno recibieron acciones de alto nivel, como "moverse a una posición específica" o "explorar un área", y se evaluó su capacidad para manejarse en un entorno poblado tanto por otros agentes como por *bots* que recolectan de manera insostenible.

Los resultados del experimento destacaron que, aunque los LAA manifestaron una tendencia a la cooperación, tuvieron problemas para entender claramente cómo colaborar de manera efectiva en el entorno dado (Mosquera *et al*., 2024; Park *et al*., 2023). Los agentes presentaron comportamientos que incluyeron la coordinación en algunos aspectos, como evitar la recolección de la última manzana de un árbol para prevenir su agotamiento; no obstante, también se observó que los agentes fallaron en intercambiar información crítica entre sí y establecer estrategias comunes a largo plazo. Estos problemas resaltan las limitaciones actuales de las arquitecturas utilizadas, lo que sugiere que aspectos como la comunicación y la planificación conjunta necesitan mejorarse para alcanzar una cooperación más eficiente.

Los desafíos identificados incluyen la necesidad de mejorar las capacidades de comunicación y entendimiento entre los agentes. Las arquitecturas actuales, aunque efectivas en algunos contextos, demuestran limitaciones significativas cuando se enfrentan a dilemas sociales complejos. Un área prometedora de investigación es el desarrollo de módulos especializados que mejoren la interpretación y el uso de la intención detrás de las acciones de otros agentes (Dafoe *et al*., 2020a). El trabajo futuro podría centrarse en extender las capacidades de memoria y reflexión, para incluir evaluaciones críticas de las acciones pasadas y su impacto en el entorno. Esto implicaría la integración de sistemas de puntuación de reputación y módulos de compromiso que incentiven comportamientos cooperativos a largo plazo (Ni y Buehler, 2024). A pesar de que los LAA muestran potencial para la cooperación, es claro que se necesitan arquitecturas más robustas y especializadas para que estos puedan colaborar de manera eficaz y eficiente en una variedad de entornos complejos.

Conclusiones

Los avances presentados en este documento subrayan el compromiso de la Universidad de los Andes con la investigación de vanguardia en DAI. Los resultados obtenidos hasta ahora no solo han demostrado la relevancia de estas tecnologías en problemas de ingeniería, resiliencia comunitaria y optimización de recursos en entornos complejos, sino que también han abierto nuevas vías para explorar aplicaciones más amplias en diferentes contextos. A medida que la DAI sigue evolucionando, su combinación con modelos avanzados de IA, como los LLM, tiene el potencial de abrir nuevas oportunidades de investigación en problemas relevantes.

La toma de decisiones en entornos multiagente enfrenta retos críticos, entre ellos la necesidad de que los agentes interpreten correctamente las acciones de los demás. En escenarios dinámicos y descentralizados, los agentes deben coordinarse y comunicarse, así como entender las intenciones y respuestas de otros agentes para adaptarse en tiempo real. Esta capacidad es esencial para manejar la incertidumbre y los posibles fallos de comunicación, factores que afectan la estabilidad y el desempeño del sistema. El trabajo futuro debería centrarse en construir sistemas más robustos, donde los agentes autónomos puedan comprender y predecir mejor las acciones de humanos y máquinas. En algunos contextos esto se denomina teoría de la mente (*theory of mind*) (Strachan *et al*., 2024). Esto llevaría a un trabajo colectivo más efectivo en situaciones que requieren de un alto nivel de coordinación, control, adaptabilidad y resiliencia. Además, es necesario mejorar los sistemas actuales, en especial en lo que respecta a la comunicación, el control, el entendimiento, las estructuras sociales y el compromiso entre los agentes para cooperar. Al incorporar capacidades de memoria y toma de decisiones más avanzadas, los sistemas futuros podrían trabajar juntos de manera más efectiva y resistir desafíos de forma resiliente.

A medida que avanzamos, la integración de DAI con otras tecnologías emergentes, como el internet de las cosas y los sistemas ciberfísicos, probablemente dará lugar a nuevas aplicaciones en áreas como las ciudades inteligentes y un nexo para la gestión optimizada de infraestructuras críticas. También hay un esfuerzo importante por reducir limitaciones en cuanto a la colaboración efectiva entre agentes y humanos en entornos reales. Aunque los modelos actuales permiten cierto nivel de interacción, persisten desafíos en la comprensión mutua, la adaptabilidad y la comunicación en escenarios no controlados. La DAI aún tiene un largo camino por recorrer, con muchos desafíos y oportunidades por delante; sin embargo, con investigación y desarrollo continuos, tiene el potencial de transformar no solo la tecnología, sino también las comunidades que dependen de ella.

Agradecimientos

Los autores quieren agradecer el aporte de Google Research para el desarrollo de algunas de las ideas expuestas en este artículo.

Referencias

Agapiou, J. P., Vezhnevets, A., Duéñez-Guzmán, E., Matyas, J., Mao, Y., Sunehag, P., Köster, R., *et al.* (2023). Melting Pot 2.0. *arXiv*. https://arxiv.org/abs/2211.13746

Ananduta, W., Barreiro-Gómez, J., Ocampo-Martínez, C. y Quijano, N. (2018). Mitigation of communication failures in distributed model predictive control strategies. *IET Control Theory & Applications*, *12*(18), 2507-2515. https://doi.org/10.1049/iet-cta.2018.5044.

Annaswamy, A., Khargonekar, P., Spurgeon, S. y Lamnabhi-Lagarrigue, F. (2023). *Cyberphysical-human systems: Fundamentals and applications*. John Wiley & Sons.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Bacci, G., Lasaulce, S., Saad, W. t Sanguinetti, L. (2016). Game theory for networks: A tutorial on game-theoretic tools for emerging signal processing applications. *IEEE Signal Processing Magazine*, *33*(1), 94-119. https://doi.org/10.1109/MSP.2015.2451994

Barreiro-Gómez, J., Obando, G. y Quijano, N. (2017a). Distributed population dynamics: Optimization and control applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *47*(2), 304-314. https://doi.org/10.1109/TSMC.2016.2523934

Barreiro-Gómez, J., Ocampo-Martínez, C. y Quijano, N. (2017b). Dynamical tuning for multiobjective model predictive control based on population games. *ISA Transactions*, *69*(1), 175-186. https://doi.org/10.1016/j.isatra.2017.03.027

Barreiro-Gómez, J., Mas, I., Giribet, J., Moreno, P., Ocampo-Martínez, C., Sánchez-Peña, R. y Quijano, N. (2021). Distributed data-driven UAV-formation control via evolutionary games: Experimental results. *Journal of The Franklin Institute*, *358*(1), 5334-5352. https://doi.org/10.1016/j.jfranklin.2021.05.002.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*(8), 716-719.

Bertsekas, D. y Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bhattacharya, A. (2021). *The man from the future: The visionary life of John von Neumann*. Penguin.

Bouman, F. (1995). Rotating and accumulating savings and credit associations: A development perspective. *World Development*, *23*(3), 371-384.

Broekhuizen, T., Dekker, H., De Faria, P., Firk, S., Nguyen, D. y Sofka, W. (2023). AI for managing open innovation: Opportunities, challenges, and a research agenda. *Journal of Business Research*, *167*, 114196. https://doi.org/10.1016/j.jbusres.2023.114196

Carrasco Martínez, S., Gamboa Montero, J. J., Maroto Gómez, M., Alonso Martín, F. y Salichs, M. Á. (2023). Aplicación de estrategias psicológicas y sociales para incrementar el vínculo en interacción humano-robot. *Revista Iberoamericana de Automática e Informática Industrial*, *20*(2), 199-212. https://doi.org/10.4995/riai.2023.18739

Collins, D. (2009). *Portfolios of the poor: How the world's poor live on $2 a day*. Princeton University Press.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K., Leibo, J., Larson, K. y Graepel, T. (2020). Open problems in cooperative AI. *arXiv*. https://arxiv.org/abs/2012.08630

Detjen-Leal, D., Quijano, N. y Rodríguez, C. (2023). *Dynamic policy evaluation for ethical decision-making in autonomous vehicles* [ponencia]. IEEE 6th Colombian Conference on Automatic Control (CCAC).

Du, Y., Li, S., Torralba, A., Tenenbaum, J. y Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv*. https://arxiv.org/abs/2305.14325

Gao, B. y Pavel, L. (2021). On passivity, reinforcement learning, and higher order learning in multiagent finite games. *IEEE Transactions on Automatic Control*, *66*(1), 121-136. https://doi.org/10.1109/TAC.2020.2978037

Genis-Mendoza, F., Konstantopoulos, G. y Bauso, D. (2022). Online pricing for demand-side management in a low-voltage resistive micro-grid via a Stackelberg game with incentive strategies. *IET Smart Grid*, *5*(2), 76-89. https://doi.org/10.1049/stg2.12053

Giraldo, J., Quijano, N. y Passino, K. (2015). Honeybee social foraging algorithm for resource allocation. En *Springer handbook of computational intelligence* (pp. 1361-1376). Springer.

Gómez, D., Quijano, N. y Giraldo, L. F. (2022). Information optimization and transferable state abstractions in deep reinforcement learning.

ieee Transactions on Pattern Analysis and Machine Intelligence, *45*(4), 4782-4793.

González Villasanti, H., Giraldo, L. F. y Passino, K. (2018). Feedback control engineering for cooperative community development: Tools for financial management advice for low-income individuals. *ieee Control Systems Magazine*, *38*(3), 87-101.

Grammatico, S., Parise, F., Colombino, M. y Lygeros, J. (2016). Decentralized convergence to nash equilibria in constrained deterministic mean field control. *ieee Transactions on Automatic Control*, *61*(11), 3315-3329. https://doi.org/10.1109/TAC.2015.2513368.

Groot, N., De Schutter, B. y Hellendoorn, H. (2014). Toward system-optimal routing in traffic networks: A reverse Stackelberg game approach. *ieee Transactions on Intelligent Transportation Systems*, *16*(1), 29-40. https://doi.org/10.1109/TITS.2014.2322312.

Gross, J., Méder, Z., De Dreu, C., Angelo Romano, A., Molenmaker, W. y Hoenig, L. (2023). The evolution of universal cooperation. *Science Advances*, *9*(7), eadd8289. https://www.science.org/doi/508abs/10.1126/sciadv.add8289

Han, Z., Niyato, D., Saad, W. y Bacsar, T. (2019). *Game theory for next generation wireless and communication networks: Modeling, analysis, and design*. Cambridge University Press.

Jackson, M. (2008). *Social and economic networks* (vol. 3). Princeton University Press.

Jaleel, H. y Shamma, J. (2020). Distributed optimization for robot networks: From real-time convex optimization to game-theoretic self-organization. *Proceedings of the ieee*, *108*(11), 1953-1967. https://doi.org/10.1109/JPROC.2020.3028295

Kara, S., Martins, N. y Arcak, M. (2022). *Population games with Erlang clocks: Convergence to Nash equilibria for pairwise comparison dynamics* [ponencia]. 2022 ieee Conference on Decision and Control, Cancún, México. https://doi.org/10.1109/CDC51059.2022.9993228

Lu, Z., Cai, F., Liu, J., Yang, J., Zhang, S. y Wu, S. (2022). Evolution of water resource allocation in the river basin between administrators and managers. *Hydrology Research*, *53*(5), 716-732. https://doi.org/10.2166/nh.2022.128

Martínez-Piazuelo, J., Díaz-García, G., Quijano, N. y Giraldo, L. F. (2022a). Discrete-time distributed population dynamics for optimization and control. *ieee Transactions on Systems, Man, and Cybernetics: Systems*, *52*(11), 7112-7122. https://doi.org/10.1109/TSMC.2022.3151042

Martínez-Piazuelo, J., Ocampo-Martínez, C. y Quijano, N. (2022b). Generalized Nash equilibrium seeking in population games under the Brown-von Neumann-Nash dynamics. En *2022 European Control Conference, London, UK* (pp. 2161-2166). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.23919/ECC55457.2022.9838437

Martínez-Piazuelo, J., Quijano, N. y Ocampo-Martínez, C. (2022c). A payoff dynamics model for generalized Nash equilibrium seeking in population games. *Automatica*, *140*(1), 110227. https://doi.org/10.1016/j.automatica.2022.110227

Martínez-Piazuelo, J., Quijano, N. y Ocampo-Martínez, C. (2022d). Nash equilibrium seeking in full-potential population games under capacity and migration constraints. *Automatica*, *141*(1), 110285. https://doi.org/10.1016/j.automatica.2022.110285.

Martínez-Piazuelo, J., Ananduta, W., Ocampo-Martínez, C., Grammatico, S. y Quijano, N. (2023). Population games with replicator dynamics under eventtriggered payoff provider and a demand response application. IEEE *Control Systems Letters*, *7*(1), 3417-3422. https://doi.org/10.1109/LCSYS.2023.3285532

Martins, N. C., Certorio, J. y La, R. J. (2023). Epidemic population games and evolutionary dynamics. *Automatica*, *153*(1), 111016. https://doi.org/10.1016/j.automatica.2023.111016.

Maynard-Smith, J. y Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15-18. https://doi.org/10.1038/246015a0

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the* IRE, *49*(1), 8-30.

Mojica-Nava, E., Macana, C. A. y Quijano, N. (2013). Dynamic population games for optimal dispatch on hierarchical microgrid control. IEEE *Transactions on Systems, Man, and Cybernetics: Systems*, *44*(3), 306-317. https://doi.org/10.1109/TSMCC.2013.2266117

Mosquera, M., Pinzón, J. S., Ríos, M., Fonseca, Y., Giraldo, L. F., Quijano, N. y Manrique, R. (2024). Can LLM-augmented autonomous agents cooperate? An evaluation of their cooperative capabilities through Melting Pot. *arXiv*. https://arxiv.org/abs/2403.11381.

Muros, F. J. (2021). El control coalicional en el marco de la teoría de juegos cooperativos. *Revista Iberoamericana de Automática e Informática Industrial*, *18*(2), 97-112. https://doi.org/10.4995/riai.2020.13456

Ni, B. y Buehler, M. J. (2024). MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and

integrate knowledge. *Extreme Mechanics Letters*, *67*, 102131. https://doi.org/10.1016/j.eml.2024.102131

Obando, G. Martínez-Piazuelo, J., Quijano, N. y Ocampo-Martínez, C. (2024). Juegos poblacionales y modelos dinámicos de pago: Un nuevo paradigma para control y optimización. *Revista Iberoamericana de Automática e Informática Industrial*, *21*(4), 287-305.

Obando, G., Poveda, J. I. y Quijano, N. (2016). Replicator dynamics under perturbations and time delays. *Mathematics of Control, Signals, and Systems*, *28*(3), 1-32. https://doi.org/10.1007/s00498-016-0170-9.

Obando, G., Quijano, N. y Ocampo-Martínez, C. (2022). Decentralized control for urban drainage systems using replicator dynamics. *IEEE Access*, *10*(1), 56740-56762. https://doi.org/10.1109/ACCESS.2022.3177631

Ochoa, D. E., Poveda, J. I., Uribe, C. y Quijano, N. (2021). Robust optimization over networks using distributed restarting of accelerated dynamics. *IEEE Control Systems Letters*, *5*(1), 301-306. https://doi.org/10.1109/LCSYS.2020.3001632.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., *et al.* (2024). GPT-4 technical report. *arXiv.* https://arxiv.org/abs/2303.08774

Park, S. y Barreiro-Gómez, J. (2023). Payoff mechanism design for coordination in multi-agent task allocation games. *arXiv*. https://arxiv.org/abs/*2306.02278*

Park, S. y Leonard, N. E. (2021). KL divergence regularized learning model for multi-agent decision making. En *2021 American Control Conference, New Orleans, US*. (pp. 4509-4514). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.23919/ACC50511.2021.9483414.

Park, S., Martins, N. C. y Shamma, J. S. (2019). From population games to payoff dynamics models: A passivity-based approach. En *2019 IEEE Conference on Decision and Control* (pp. 6584-6601). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/CDC40024.2019.9029756

Park, S., O'Brien, J., Cai, C., Morris, M., Liang, P. y Bernstein, M. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv*. https://arxiv.org/abs/2304.03442

Pawlick, J. y Zhu, Q. (2021). *Game theory for cyber deception: From theory to applications*. Springer-Verlag.

Quijano, N., Ocampo-Martínez, C., Barreiro-Gómez, J., Obando, G., Pantoja, A. y Mojica-Nava, E. (2017). The role of population games and evolutionary dynamics in distributed control systems. *IEEE*

Control Systems Magazine, *37*(1), 70-97. https://doi.org/10.1109/MCS.2016.2621479.

Quijano, N. y Passino, K. M. (2010). Honeybee social foraging algorithms for resource allocation: Theory and application. *Engineering Applications of Artificial Intelligence*, *23*(6), 845-861.

Rass, S., Schauer, S., Konig, S. y Zhu, Q. (2020). *Cyber-security in critical infrastructures: A game-theoretic approach*. Springer.

Russell, S. J. y Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.

Salas Bahamón, L. M. (2022). Inclusión financiera en Colombia. Evaluación de impacto del programa Grupos de Ahorro y Crédito Comunitario. *Cuadernos de Economía*, *41*(87), 747-782.

Sánchez-Amores, A., Martínez-Piazuelo, J., Maestre, J. M., Ocampo-Martínez, C., Camacho, E. F. y Quijano, N. (2023). Coalitional model predictive control of parabolic-trough solar collector fields with population-dynamics assistance. *Applied Energy*, *334*(1), 120740. https://doi.org/10.1016/j.apenergy.2023.120740

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. y Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv*. https://arxiv.org/abs/2302.04761

Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., *et al.*, (2024). Testing theory of mind in large language models and humans. *Nature Human Behavior*, *8*, 1285-1295.

Taylor, P. y Jonker, L. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, *40*(1-2), 145-156. https://doi.org/10.1016/0025-5564(78)90077-9

Thorndike, E. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i-109.

Vlassis, N. (2022). *A concise introduction to multiagent systems and distributed artificial intelligence*. Springer Nature.

von Neumann, J. y Morgenstern, O. (2007). *Theory of games and economic behavior: 60th anniversary commemorative edition*. Princeton University Press.

Wang, G. Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L. y Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv*. https://arxiv.org/abs/2305.16291

Weiss, G. (1999). Multiagent systems: A modern approach to distributed artificial intelligence. MIT Press.

Yang, Y. y Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv*. https://arxiv.org/abs/2011.00583

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. y Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *arXiv*. https://arxiv.org/abs/2210.03629

Zambrano, A. F., Giraldo, L. F., Perdomo, M. T., Hernández, I. D. y Godoy, J. M. (2022). Variations of rotating savings and credit associations forcommunity development. *IEEE Transactions on Computational Social Systems*, *10*(2), 614-622.

Zambrano, A. F., Giraldo, L. F., Perdomo, M. T., Hernández, I. D. y Godoy, J. M. (2023). Rotating savings and credit associations: A scoping review. *World Development Sustainability*, *3*, 100081.

Zhang, J., Xu, X. y Deng, S. (2023). Exploring collaboration mechanisms for LLM agents: A social psychology view. *arXiv*. https://arxiv.org/abs/2310.02124

USO DEL APRENDIZAJE POR REFUERZO PARA EL MANEJO DE COMPORTAMIENTO DESCONOCIDO EN SISTEMAS DE *SOFTWARE* DINÁMICOS

Nicolás Cardozo, Ivana Dusparic

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.02

Introducción

Los sistemas de *software* actuales están en continua interacción con sistemas externos o distintas fuentes de información; esta interacción permite que los sistemas sean más conscientes de su entorno de ejecución y, en respuesta, puedan adaptar su comportamiento para que este sea el más adecuado con respecto a dicho entorno. Estos sistemas, llamados *sistemas de adaptación dinámica* (*self-adaptive systems*, SAS) (Salehie y Tahvildari, 2009), continuamente adaptan su comportamiento de acuerdo con múltiples situaciones detectadas en tiempo de ejecución.

Los SAS han probado ser de utilidad en diversos ambientes, como el manejo de sistemas de transporte (Castagna y Dusparic, 2022; Khan *et al*., 2016), en sistemas robóticos de enjambre (Zhao *et al*., 2018; Zhu *et al*., 2024), en sistemas de comportamiento emergente (Cardozo, 2016) y, en general, en sistemas autonómicos (Cabrera *et al*., 2024; Kephart y Chess, 2003). La construcción de los SAS está compuesta por tres pasos principales: (1) definir el comportamiento base del sistema, (2) determinar las situaciones o condiciones para adaptar el comportamiento y (3) establecer el comportamiento especializado del sistema. Sin embargo, bajo estos parámetros, los SAS tienen una capacidad de adaptación y dinamicidad limitada a los problemas o situaciones que son identificados durante el diseño del sistema, las llamadas *conocidas situaciones desconocidas* (*known-unknowns*) (D'Angelo *et al*., 2019). Por lo tanto, la realización de SAS en situaciones desconocidas requiere afrontar los siguientes retos (Cardozo y Dusparic, 2021, 2022):

1. Las adaptaciones y las situaciones en las que aplican no deben ser prescritas a situaciones conocidas, pues se afecta la dinamicidad y adaptabilidad de los sistemas.
2. El manejo de interacciones y la composición entre adaptaciones no previstas puede causar errores en el comportamiento. Las soluciones actuales a dicho problema necesitan la definición inicial de las reglas de composición (Schmerl *et al*., 2017), lo que de nuevo afecta la dinamicidad y adaptabilidad de los sistemas.

Para abordar estos problemas, proponemos un mecanismo de aprendizaje que permite la generación dinámica de adaptaciones (Cardozo y Dusparic, 2023; Sanabria *et al*., 2024), denominado Auto-COP, y la mejor composición de dichas adaptaciones (Cardozo y Dusparic, 2020), denominado ComInA. De esta forma, el sistema responde a situaciones por completo desconocidas, en la que la estrategia de adaptación no está prescrita, sino más bien es aprendida, lo que ofrece una mayor flexibilidad para que el sistema responda correctamente a todo tipo de situaciones.

Conceptos preliminares

Para la construcción de SAS capaces de adaptar su comportamiento a situaciones desconocidas, utilizamos dos conceptos principales. Primero, la programación orientada al contexto (*context-oriented programming*, COP) se emplea como una herramienta para construir SAS de forma modular y con un alto grado de granularidad de las adaptaciones. Segundo, usamos aprendizaje por refuerzo (*reinforcement learning*, RL) como una estrategia de aprendizaje para la generación y composición de adaptaciones en tiempo de ejecución. El uso de RL permitirá aprender situaciones, condiciones de ejecución y comportamientos inicialmente desconocidos, en tiempo de ejecución. En concreto, utilizamos opciones de RL para aprender comportamientos y la técnica de aprendizaje multiobjetivo *aprendizaje W* para resolver la composición más adecuada con respecto a los objetivos del sistema.

Programación orientada al contexto

La COP (Hirschfeld *et al*., 2008; Salvaneschi *et al*., 2012) es un paradigma de programación para realizar adaptaciones de forma dinámica al comportamiento de los programas, a un alto nivel de granularidad (*e.g.,* métodos). La COP permite una clara independencia entre los módulos de adaptaciones y el

comportamiento base de los programas, así como de otras adaptaciones. Las adaptaciones se incorporan a un sistema durante la ejecución, mediante la recomposición dinámica del sistema; de esta manera, el modelo de composición dinámica utilizado en COP reifica la arquitectura de adaptación de *monitoring, analysis, plan, execute* (MAPE) (Rutten *et al.*, 2017).

Para implementar SAS capaces de adaptar su comportamiento a situaciones desconocidas, en los ejemplos a continuación utilizamos el lenguaje de ContextTraits (González *et al.*, 2013), una extensión de ECMAScript que permite adaptaciones dinámicas. Sin embargo, nótese que los conceptos que desarrollaremos no son específicos de este lenguaje y son aplicables a SAS en general. Hay tres conceptos principales detrás de las adaptaciones dinámicas en COP: *contextos*, *variaciones de comportamiento* y *activaciones de contexto*. Los contextos corresponden a objetos (p. ej., entidades de primera clase del sistema) que representan situaciones del entorno de ejecución capturadas por variables del sistema (p. ej., el estado) o eventos externos monitoreados por sensores. Siempre que se cumplan las condiciones específicas del entorno para los contextos, se dice que estos se encuentran *activos*; de lo contrario, estarán *inactivos*.

Por ejemplo, en el caso del asistente de navegación para un vehículo, el contexto `closeProximity` en el algoritmo 1 define la situación en la que un vehículo se acerca a otro vehículo delante de este.

```
closeProximity = new cop.Context ({
       description: "vehicle in close proximity"
})
```

Algoritmo 1. Definición estática de un contexto en ContextTraits

Cada contexto está asociado a un conjunto de variaciones de comportamiento (p. ej., métodos) que especifican adaptaciones al comportamiento base del sistema. Por ejemplo, en el algoritmo 2 se define el comportamiento especializado para gestionar la proximidad a un vehículo, al adaptar de forma efectiva el comportamiento base drive (p. ej., continuar en línea recta), definido para el vehículo, con nuevas acciones por ejecutar en su lugar —`steerLeft()` y `steerRight()`—, que permitan girar para evitar el vehículo que está delante. El contexto detectado `closeProximity` se asocia con su comportamiento por medio de la abstracción `adapt`, como se muestra en la línea 7 del algoritmo 2.

```
closeProximityBehavior = Trait ({
   drive: function () {
      steerLeft()
      steerRight()
   }
})
closeProximity.adapt(vehicle, closeProximityBehavior)
```

Algoritmo 2. Variación de comportamiento definida para evitar vehículos lentos delante

A medida que los contextos se activan, sus variaciones de comportamiento se asocian con el sistema y se presentan disponibles para la ejecución, es decir, las variaciones serán el comportamiento observable del sistema. Internamente, al activarse un contexto, sus variaciones de comportamiento asociadas se componen con el sistema en ejecución. La desactivación del contexto retira todas las variaciones de comportamiento asociadas con el contexto del sistema en tiempo de ejecución. En ambos casos, el comportamiento del sistema se adapta de forma dinámica. El contexto `closeProximity` se activa y desactiva, como se muestra en el algoritmo 3, basado en la información del sensor de proximidad.

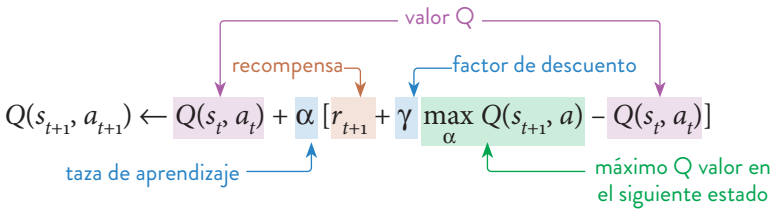
```
if(proximitySensor.receive() < 300)
       closeProximity.activate()
else
       closeProximity.deactivate()
```

Algoritmo 3. Condiciones de activaciones de contextos

Aprendizaje por refuerzo

En el RL, los agentes inteligentes aprenden a mapear situaciones del entorno (estados del entorno) sobre acciones, para maximizar una señal de recompensa numérica que reciben del entorno a largo plazo (Sutton y Barto, 2018). Los agentes de RL están definidos por: *S*, el espacio de estados, formado por todos los estados relevantes del entorno; *A*, el espacio de acciones, es decir, el conjunto de todas las acciones que un agente puede ejecutar y que afectan al entorno; y la recompensa *r*, la señal numérica que codifica el impacto positivo o negativo de la acción en cada paso de ejecución.

El aprendizaje Q (Watkins y Dayan, 1992) es una implementación libre de modelo muy utilizada en RL. La calidad a largo plazo de una acción realizada en un determinado estado se calcula de forma iterativa en una serie de pasos y está representada por un valor $Q(s, a)$. Formalmente, cada paso de ejecución t captura información del entorno y la asigna a un estado $s_t \in S$ de su espacio de estados. A continuación, selecciona una acción $a_t \in A$ de su espacio de acciones y la ejecuta. El agente recibe una recompensa r_t del entorno, cuando pasa al siguiente estado $s_{t+1} \in S$. La recompensa se utiliza para actualizar la optimalidad de realizar la acción a_t en el estado s_t. El objetivo del agente es aprender una política (p. ej., la acción más adecuada para cada estado) que maximice la recompensa del comportamiento a largo plazo. La tasa de aprendizaje α determina qué tanto las nuevas experiencias sobrescriben las experiencias aprendidas con anterioridad, y el factor de descuento γ determina cuánto se descuentan las recompensas futuras para que los agentes prioricen las acciones inmediatas y puedan planificar las mejores acciones a largo plazo. En cada paso de tiempo t, el valor Q de una acción a_t tomada en el estado s_t se actualiza mediante la ecuación de aprendizaje de Bellman, como se observa en la siguiente ecuación.



Opciones en aprendizaje por refuerzo

Las opciones (Sutton *et al.*, 1998) o macroacciones en RL son acciones extendidas temporalmente, que se utilizan para acelerar el aprendizaje, o minimizar los periodos de rendimiento subóptimo durante exploración del entorno, e incorporar acciones a diferentes niveles de granularidad. Las opciones son adecuadas para aprender e integrar secuencias de acciones en sistemas adaptativos basados en COP. Las adaptaciones en COP responden a cambios en el contexto, de forma similar a como se aprenden las acciones en condiciones observadas en el entorno.

Dentro del modelo de ejecución de RL, una opción codifica secuencias de acciones atómicas ejecutadas por el agente en acciones temporalmente extendidas. Las opciones se definen por tres componentes: una política π, que es la

correspondencia entre el espacio de estados *S* y el espacio de acciones *A*; una condición de inicio (p. ej., espacio de estados para comenzar la opción) $I \subseteq S$, y una condición de terminación, que determina la longitud de la opción. Existen numerosas formas de construir opciones a partir de acciones atómicas (Elfwing *et al.*, 2004; Girgin y Polat, 2005; McGovern y Sutton, 1998; Randløv, 1998; Stolle y Precup, 2002). En este trabajo, en particular utilizaremos técnicas estrechamente relacionadas con el cumplimiento de submetas (Stolle y Precup, 2002) y los comportamientos (Girgin y Polat, 2005).

Aprendizaje de múltiples objetivos

El aprendizaje Q utiliza una única fuente de recompensa, esto es, permite optimizar un único objetivo del sistema. Para aprender múltiples objetivos, es posible implementar varios procesos de aprendizaje Q; sin embargo, como un agente solo puede ejecutar una acción a la vez, es necesario añadir un método de arbitraje que resuelva cuál de los procesos de aprendizaje Q, es decir, cuál de los objetivos del agente, toma el control de la ejecución de la acción. El aprendizaje W (Humphrys, 1995, 1996) proporciona ese mecanismo. En cada paso de tiempo, cada proceso de aprendizaje Q propone una acción ejecutable, la más adecuada para el objetivo que representa. Al implementar el aprendizaje W, los agentes también aprenden, en términos de recompensas recibidas para cada uno de los estados, la importancia de la selección de su acción propuesta en contraposición a las acciones propuestas por otros agentes. Dicha importancia se expresa como un valor *W(s)*, aprendido por cada agente de aprendizaje Q para cada estado. El agente con el máximo valor *W* tiene prioridad en el siguiente paso temporal al ejecutar la acción propuesta.

Aprendizaje y generación de adaptaciones

Esta sección presenta el diseño de Auto-COP (Cardozo y Dusparic, 2018, 2022, 2023), nuestra solución para la generación de adaptaciones para SAS. Primero, presentamos la funcionalidad de alto nivel de Auto-COP y, luego, describimos sus dos componentes principales: (1) el aprendizaje de secuencias de acciones, mediante opciones de RL, y (2) la generación de adaptaciones basadas en las opciones aprendidas.

Generación de adaptaciones

En el desarrollo de los SAS, la información de las adaptaciones y los estados en las que deben suceder puede no estar disponible o no ser conocida de antemano por los desarrolladores. Esto limita la adaptabilidad del sistema al comportamiento conocido y especificado. Además, es difícil saber si las adaptaciones predefinidas corresponden realmente al comportamiento más apropiado para el estado actual, ya que es posible que el entorno evolucione durante la ejecución. Para abordar estos problemas, Auto-COP permite la generación dinámica de adaptaciones basadas en ejecuciones previas del sistema y la interacción con el entorno.

El proceso para la generación de adaptaciones en Auto-COP se muestra en la figura 2.1. Las acciones ejecutadas por el sistema generan una traza de ejecución de acciones. El origen de estas acciones puede ser un comportamiento predefinido, la intervención de usuarios (p. ej., acciones atómicas) o acciones generadas por otros procesos (p. ej., adaptaciones). La traza de las acciones ejecutadas se utiliza como entrada para el modelo de aprendizaje de opciones RL, que aprende las secuencias de acciones más adecuadas para las condiciones específicas del entorno. Estas secuencias se utilizan como entrada para el generador de adaptaciones, que las empaqueta en adaptaciones reutilizables, lo que define un contexto a partir del estado del sistema y las variaciones de comportamiento de dicho contexto con base en las opciones aprendidas.

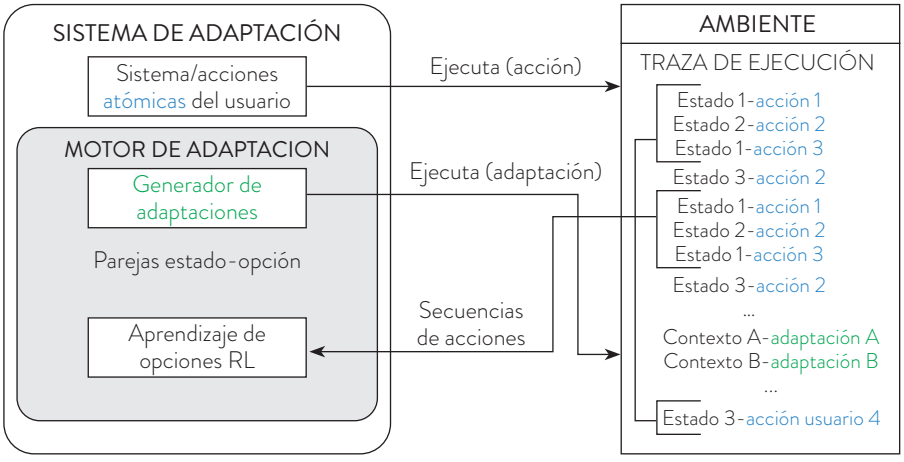


Figura 2.1. Modelo del proceso de aprendizaje de adaptaciones
Fuente: elaboración propia.

Aprendiendo opciones

El aprendizaje de opciones es un proceso que se lleva a cabo a lo largo de la ejecución del sistema. Para aprender opciones, observamos el conjunto de estados *S* del entorno (p. ej., variables monitoreadas), los cuales serán candidatos para convertirse en contextos, y el conjunto de acciones *A* (atómicas u opciones aprendidas), que pueden ejecutarse para cada estado. El módulo `Aprendizaje de opciones RL` aprende un conjunto de opciones *O*, que contiene un mapa de todos los estados $s_i \in S$, asociados a un conjunto de secuencias de acciones, definidas como opciones O_{si}, disponibles para su ejecución en ese estado. Todos los conjuntos de opciones O_{si} están inicialmente vacíos, ya que las opciones se aprenden durante la ejecución. Los conjuntos de opciones O_{si} se procesan en pequeños lotes de tamaño `batchSize`. Usamos el procesamiento por lotes, ya que el procesamiento después de cada paso de ejecución puede ser costoso y tener un efecto insignificante en las opciones generadas. Las líneas 4-15 del algoritmo 4 muestran el pseudocódigo con los detalles del proceso.

```
1 S := {s1, ..., sn} ; A := {a1, ..., am} ; O :
  = {{s1, Ø}, ..., {sn, Ø}} ; lastBatchEnd := 0
2 //EXTRACCION DE OPCIONES
3 while(true){
4  AtomicActionLog:=write(si, ai, r(si, ai))
5  if(timestep t mod batchSize == 0){
6  //Construir opciones de lotes de funciones en la
   traza de ejecucion (Log)
7     for(i=lastBatchEnd; i<logSize; i++){
8        loggedState := readLogLine(i, statePosition)
         //Obtener estado en el log
9        for(j=0; j<maxOptionLength&&!goalReached; j++)
10          actionSequencei += readLogLine(i+j, action
            Position) //Obtener acciones
11    }
12    Oi := {loggedState, actionSequencei} ; O.add(Oi)
13 }
14 lastBatchEnd := logSize
15 }
16 //GENERACION DE ADAPTACIONES
```

```
17 ReinforcementLearning.initialize(S, O)
18 currentState := senseEnvironment()
19 if(contextAdaptationAvailable(currentState)) {
   //seleccion de opciones
20    selectedOption := currentState.pickAnOption(E)
      //epsilon-greedy
21    currentState. activate ()
22    execute(selectedOption) //ejecutar la adaptacion
23    currentState. deactivate ()
24    ReinforcementLearning.updateOption(S, O,
      r(currentState, selectedOption))
25 } else { //no hay opciones disponibles, ejecute la
   accion atomica
26    atomicAction := currentState.pickAction()
27    execute(atomicAction)
28 }
29 newCOPAdaptation := generate(currentState,
   highestQOption)
30 O.add(newCOPAdaptation)
31 }
```

Algoritmo 4. Proceso continuo para la generación de adaptaciones

Durante las primeras fases de ejecución, solo se ejecutan acciones atómicas (predefinidas o ejecutadas por usuarios), ya que el sistema no dispone de adaptaciones generadas a partir de opciones. La ejecución de cada acción se registra en la traza de ejecución del programa (p. ej., el `log`) junto con el estado en las que la acción se ejecutó (a la derecha en la figura 2.1). Para cada acción también registramos la recompensa del efecto de la acción en el estado actual. Esta recompensa, $r(s_i, a_i)$, puede tener múltiples fuentes. Por ejemplo, si las acciones subyacentes se aprenden mediante un sistema basado en RL, la recompensa corresponde a la que se obtiene del entorno. Si las acciones las ejecuta un ser humano, es posible asociar a cada una de ellas una recompensa positiva fija, bajo el supuesto de que tales intervenciones son realizadas por expertos. Por último, puede darse una recompensa constante (p. ej., 1) cada vez que se encuentre un par estado-acción, suponiendo que las acciones que se ejecutan con más frecuencia son las más adecuadas.

El resultado de este proceso es *O*, el mapa de todos los estados identificados en la traza de ejecución, los cuales se consiguen por medio de la función `readLogLine`, que obtiene la información guardada dentro de la traza de ejecución. Cada estado está asociado a un conjunto de secuencias de opciones con tamaño entre 1 (p. ej., una acción atómica) y la longitud máxima de opción *n*; esta puede especificarse externamente, ser calculada de forma experimental, teniendo en cuenta la frecuencia y la utilidad de las secuencias registradas, o fijarse en el número máximo de acciones necesarias para alcanzar un estado específico en el sistema. El algoritmo 5 ilustra un ejemplo genérico de posibles secuencias de acciones generadas para los estados `stateVariablesSet1` y `stateVariablesSet2`.

```
1  state: [stateVariablesSet1]
2  reward:1 -> actions: [“action3”]
3  reward:4 -> actions: [“action3”, “action1”]
4  reward:10 -> actions: [“action3”,
   “action1”,”adaptation1”]
5  reward:10 -> actions: [“action3”, “action1”,
   “adaptation1”, “action4”]

7  state: [stateVariablesSet2]
8  reward:11 -> actions: [“action1”]
9  reward:17 -> actions: [“action1”, “adaptation1”]
```

Algoritmo 5. Secuencias de acciones extraídas como un conjunto ordenado (*O*)

Note que cada estado puede contener múltiples opciones extraídas, y es posible que las secuencias de acciones incluyan opciones generadas previamente (acciones `adaptation` en el fragmento de código). Sin embargo, la mayoría de las opciones extraídas serán inadecuadas para la adaptación, ya que podrían no llevar a un estado correcto. Auto-COP reduce las opciones, para seleccionar una única opción como el comportamiento más adecuado para cada estado (p. ej., contexto), el cual utilizamos para generar una adaptación.

Generación automatizada de adaptaciones

El módulo `Generador de adaptaciones` en Auto-COP toma como entrada las diferentes opciones generadas para cada estado explorado y utiliza un proceso para aprender la opción más adecuada, con el fin de ejecutarla como una

adaptación al comportamiento del sistema. Este proceso de generación se especifica en las líneas 17-30 del algoritmo 4. Para cada opción ejecutada en un estado (p. ej., `currentState`), registramos la recompensa de ejecutar la opción mediante actualizaciones estándar de *Q-learning* (estado, acción, recompensa), con el objetivo de maximizar el rendimiento del sistema a largo plazo. Luego del proceso de exploración, las opciones propuestas como adaptaciones son aquellas con un valor *Q* mayor (p. ej., las mayores recompensas esperadas a largo plazo) para cada estado. El modelo de recompensa de las opciones tiene en cuenta la recompensa del sistema por alcanzar el estado final de la opción. Utilizamos este modelo porque nos interesa que el sistema alcance su objetivo final; solo se consideran más apropiadas las opciones que cierran la brecha entre el estado actual y el estado objetivo. Las opciones con mayor recompensa son las que se utilizan en el módulo `Generador de adaptaciones`, para producir los objetos de contexto y las variaciones del comportamiento

```
1  ContextCurrentState = new cop.Context({name:
   “currentState” })
2  BehavioralVariation = Trait({
3     option:function () {
4        //Secuencia de acciones aprendida
5        action1();
6        ...
7        actionn();
8     }
9  });
10 ContextCurrentState.adapt(BaseSystem, Behavioral
   Variation);
```

Algoritmo 6. Template de la generación de adaptaciones

La definición de contextos generados se muestra en la línea 1 del algoritmo 6. Cada contexto recibe como nombre una cadena correspondiente al estado en el que se debe ejecutar. La generación de variaciones de comportamiento implica la redefinición del comportamiento base del sistema, mediante el uso de la secuencia de acciones de la opción seleccionada, tal y como se muestra en las líneas 2-9 del algoritmo 6. Por último, también generamos la asociación de las variaciones de comportamiento con su respectivo contexto; esto se hace en la línea 10 del algoritmo 6.

Una vez se generan las adaptaciones, mientras el sistema sigue ejecutándose, siempre que se detecte el estado asociado a una de estas (su contexto), el sistema activa el contexto correspondiente `ContextCurrentState.activate()`. Esto llama a la composición de la variación de comportamiento asociada al contexto con el sistema en ejecución. Una vez ejecutada la variación de comportamiento, el contexto se desactiva y el sistema vuelve a su comportamiento base, ejecutando acciones atómicas.

Este proceso se desarrolla de forma continua durante la ejecución del sistema: las trazas se registran, se procesan por lotes en opciones y se explora su idoneidad en interacción con el entorno, para utilizar las opciones más adecuadas como adaptaciones. A medida que el sistema se ejecuta, las opciones generadas también pueden considerarse para la generación de nuevas opciones, lo que compone opciones dentro de opciones. Además, hay que tener en cuenta que, si las condiciones del entorno cambian, la idoneidad de las adaptaciones generadas puede cambiar, y es posible que las nuevas opciones reciban una mejor recompensa que las opciones utilizadas para generar las adaptaciones actuales (p. ej., las adaptaciones obtienen una recompensa negativa, o se ejecutarán acciones atómicas diferentes por los usuarios del sistema). Este proceso permite la generación continua de adaptaciones. De este modo, Auto-COP elimina la necesidad de predefinir las adaptaciones en el momento del diseño y garantiza que el sistema se adapte continuamente a medida que cambian las condiciones, sin cambios manuales en el código fuente.

Aprendiendo estrategias de composición

Una vez generadas las adaptaciones de comportamiento del sistema, nos encontramos con el problema de cómo manejar su combinación. Cada una de las adaptaciones es generada como el comportamiento más apropiado para un estado específico del sistema; sin embargo, en los grandes sistemas, múltiples estados pueden estar presentes en un mismo momento de la ejecución, y el sistema debe responder con la mejor combinación de adaptaciones posibles.

Dado que predefinir todas las posibles combinaciones de adaptaciones para un sistema es imposible, junto con Auto-COP presentamos un novedoso enfoque de composición de adaptaciones, ComInA, que aprende de forma autónoma las interacciones entre adaptaciones, así como las composiciones de adaptaciones más apropiadas para cada combinación de contextos activos. Este proceso se basa en *w-learning* y está formado por: (1) agentes de contexto, los cuales tienen la tarea de realizar una adaptación específica para cada contexto y aprender la idoneidad de otras adaptaciones disponibles en el sistema;

(2) agentes de interacción, cuya tarea es aprender cómo los contextos y las adaptaciones afectan a un contexto determinado (p. ej., la relación entre contextos); y (3) compositor de contextos, que, con base en las aportaciones de los contextos, determina la adaptación o combinación que se debe ejecutar para cada combinación de contextos activos. Ahora, describiremos los algoritmos implementados por cada uno de los agentes.

Diseño de los agentes de contexto

ComInA define un conjunto de agentes de contexto A_{c1}, ..., A_{cn} para cada sistema, implementados mediante un proceso de aprendizaje Q para cada contexto c_i. Inicialmente, cada agente tiene un conjunto de estados Sq_{ci} que indican si su contexto está activo $c_i - 1$ o si esta inactivo $c_i - 0$, y un conjunto de acciones que contiene una única adaptación $A = \{a_{ci}\}$. a_{ci} es la adaptación necesaria para el contexto c_i. La definición de la adaptación, preespecificada o aprendida, es irrelevante, solo requerimos que la adaptación exista para la creación del agente. El proceso de aprendizaje de un agente de contexto se describe en el algoritmo 7. Cada vez que el contexto c_i está activo, el agente A_{ci} propone ejecutar la adaptación a_{ci}. Sin embargo, en función de otros contextos activos y de la decisión del compositor de contextos, podrían ejecutarse otras adaptaciones. En tal caso, el agente amplía su espacio de acción con la adaptación ejecutada (desconocida antes) y aprende (mediante Q-*learning*) su impacto en su sistema. El conjunto de acciones del agente de contexto se construye en tiempo de ejecución. Si los agentes conocen otras adaptaciones en el sistema, estas pueden afectar las preferencias a la hora de ejecutar las propias adaptaciones (p. ej., si se descubre una adaptación mejor).

```
1  Sq_i := {c_i - 0, c_i - 1}
2  A_i := {a_i}
3  QLearning.INITIALIZE(Sq_i, A_i)
4  while(true){ //Ejecucion continua del sistema
5     CurrentContexts[] := senseEnvConditions() //
       evaluar el estado de los contextos en el ambiente
6     if Context C is in CurrentContexts[] {
7        nominateAdaptationToExecute(A_i)
8        currentState:=c_i - 1
9     } else {
```

```
10        currentState:=c_i - 0 // obtiene la recompensa
          para el estado actual
11        reward := QLearning.getReward{currentState)
12     }
13 // si la adaptacion no se has visto anteriormente,
   expanda el conjunto de acciones
14     execAdapt := getExecutedAdaptation()
15     if execAdapt is not in A_i
16        A_i := A_i ∪ execAdaptation
17     QLearning.update(prevState, execAdapt, reward)
       //actualizacion de aprendizaje Q
18 }
```

Algoritmo 7. Definición del agente de contexto y proceso de aprendizaje

Diseño de los agentes de interacción

Los agentes de interacción Aw_{c1}, ... Aw_{cn}, definidos por contexto, aprenden cómo su contexto interactúa con otros agentes. Se implementan utilizando *w-learning*. Elegimos *w-learning* como base de nuestro enfoque por su capacidad de codificar relaciones entre adaptaciones, sin que sus prioridades relativas sean predefinidas o codificadas en el momento del diseño (Cardozo *et al*., 2017). La prioridad relativa inicial de las adaptaciones se expresa mediante recompensas a los agentes de contexto; no obstante, durante la ejecución del sistema, si la adaptación se "descuida" durante un tiempo, su valor *W* asociado acabará siendo superior a otros valores *W*, e incluso las adaptaciones menos prioritarias se harán con el control de la ejecución, permitiendo un cambio dinámico de prioridades. Además, el aprendizaje *W* permite implícitamente la detección de adaptaciones complementarias. Por ejemplo, varias adaptaciones pueden ser adecuadas para múltiples objetivos, es decir, es posible que una adaptación sea adecuada para otro contexto como "efecto secundario" de su ejecución, aunque haya sido designada para adaptarse a otro contexto. En tal caso, el valor *W* de ambos contextos será bajo, ya que ninguno de ellos tiene que competir para ejecutar su adaptación. Este tipo de interacción permite detectar relaciones entre múltiples adaptaciones al aprender la mejor adaptación o combinación de ellas.

El proceso de aprendizaje de un agente de interacción se describe en el algoritmo 8. Al principio, el espacio de estados de A_{wi} es idéntico al espacio de estados del agente de contexto, $S_{wc1} = S_{qc1}$. El espacio de estado de los agentes de interacción se amplía en tiempo de ejecución, a medida que se observan nuevos

contextos. Esta expansión permite a los agentes aprender (cuantitativamente) el impacto de aplicar la adaptación preferida en el rendimiento del sistema, para cada combinación de contextos concreta. Por ejemplo, un agente de interacción para el contexto `c1` podría ampliar su espacio de estado, en tiempo de ejecución, para representar todas las combinaciones activas/inactivas con el contexto `c2` (`statesW=["c1-0,c2-0","c1-0,c2-1","c1-1,c2-0","c1-1,c2-1"]`) y aprender los valores *W* para cada una de las combinaciones.

Se observa que, en sistemas muy grandes, los agentes de interacción no monitorean todos los contextos activos; solo se monitorean los contextos que afectan a los componentes parte de su propio contexto.

```
1  Sw_i := Sq_i
2  WLearning.initialize(Sw_i)
3  while(true){ // ejecucion continua del sistema
4     //todos los contextos activos se convierten a un
       estado
5     CurrentContexts[] := senseEnvConditions()
6     for All (Context c in cCurrentContexts)
7        currentWState += c
8  // si conoce el estado actual, obtiene su
   importancia (w-valor)
9     if (currentWState is in Sw_i) {
10       w := WLearning.getW(currentWState)
11       // nominar la accion de acuerdo al aprendizaje Q
12       WLearning.nominateAdaptation(w, a_i)
13    } else {
14       // Si no conoce el estado actual, expandir el
         conjunto de estados
15 }
16    Sw_i := Sw_i ∪ currentWState // actualizacion de
      aprendizaje W
17    WLearning.update(prevWState, execAdapt, reward)
18 }
```

Algoritmo 8. Definición del agente de interacción y el proceso de aprendizaje

Agentes compositores de contexto

ComInA contiene al menos un agente compositor de contexto[1]. Este módulo se encarga de componer adaptaciones seleccionadas procedentes de los otros agentes. En cada paso de tiempo, un compositor de contexto recibe todos los candidatos a adaptación de los agentes de contexto y su impacto asociado de los agentes de interacción. Así, se ejecuta la adaptación con el valor *W* más alto; sin embargo, como parte de la exploración, un compositor de contexto también ejecuta composiciones de adaptaciones, variando sus combinaciones y orden, para evaluar el impacto de las adaptaciones compuestas en el sistema. Estas adaptaciones compuestas se consideran igual que las adaptaciones individuales y, por tanto, se añaden al espacio de acciones de los compositores de contexto y se aprende su idoneidad para un contexto concreto (p. ej., `adapts=["A1", "A2","A1,A2","A2,A1"]`). Así, con el tiempo, los agentes de contexto individuales pueden sugerir las adaptaciones compuestas más adecuadas para su contexto.

```
1  currentWState
2  while(true){// obtener adaptaciones preferidas de
   los agentes de contexto
3     var adaptNominations[][], finalAdaptation
4     for All (Ac1 to Acn)
5        adaptNominations.push(Ac_i.adaptation, w_i)
6     maxAdapt := findMaxWvalue(adaptNominations)
7     if (exploring) // Intenta combinaciones de
      adaptaciones
8        finalAdaptation = Ac_i.adaptation ∪ Acj.
         adaptation
9     else
10       finalAdaptation := maxAdapt
11    finalAdaptation.execute()
12 }
```

Algoritmo 9. Selección y ejecución de la composición de contextos

1 En sistemas pequeños, un único compositor puede tener una visión de todos los agentes de contexto e interacción, mientras que, en grandes sistemas, múltiples agentes están a cargo de un número limitado de componentes que cooperan como un sistema multiagente, con el fin de garantizar un rendimiento global.

Las adaptaciones se seleccionan con base en las interacciones aprendidas mediante aprendizaje *W*, tanto en el contexto como en los agentes de interacción. Las combinaciones propuestas se ejecutan a continuación, utilizando la estrategia de composición basada en el orden de activación de los contextos de COP; no obstante, ComInA no fija la estrategia de composición, sino que permite definir continuamente estrategias de composición basadas en los valores *W* de los espacios de estado expandidos del agente.

Evaluación del aprendizaje de adaptaciones

Ahora nos disponemos a mostrar la aplicación del proceso automatizado para aprender adaptaciones y su composición dentro de SAS, sin la necesidad de predefinir dichos comportamientos o interacciones. Nótese que hasta el momento esta es la primera solución que permite dicho comportamiento para los SAS.

En el desarrollo de la evaluación, primero nos concentramos en la generación de las adaptaciones (p. ej., los contextos de ejecución y las variaciones de comportamiento). Luego mostramos cómo el sistema puede aprender la composición de las adaptaciones. Para la evaluación, utilizamos un sistema de gestión del transporte urbano.

Aprendiendo adaptaciones

Para evaluar la generación de adaptaciones usamos como aplicación un asistente de navegación para carros autónomos. En particular, evaluamos la utilidad y correctitud de las variaciones de comportamiento generadas en los diferentes estados (p. ej., contextos) en las que estas suceden.

Asistente de navegación

El asistente de navegación en una autopista de dos carriles utiliza cinco acciones de intervención atómicas para controlar los vehículos: `straight`, `steerRight`, `steerLeft`, `speedUp` y `slowDown` (las últimas acciones modifican la velocidad actual en ±10 km/h). El proceso de conducción consiste en una serie de acciones que se repiten con frecuencia en respuesta a las condiciones de la autopista. El comportamiento base del sistema es manejar por el carril derecho de la autopista. Para mantener un comportamiento adecuado en diferentes situaciones, el sistema debe adaptar su comportamiento ejecutando nuevas secuencias de acciones en situaciones particulares. Por ejemplo, la figura 2.2 muestra la situación cuando el vehículo *v1* se encuentra con *v2*, el cual maneja a una velocidad

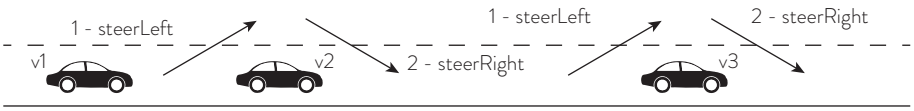


Figura 2.2. Escenario para adelantar en el asistente de navegación
Fuente: elaboración propia.

inferior a la suya; *v1* adelanta a *v2* mediante las acciones `steerLeft` y `steerRight`. Cuando se encuentra con el siguiente vehículo, *v3*, realiza las mismas acciones `steerLeft` y `steerRight`.

El comportamiento esperado del sistema es circular por el carril derecho y usar el carril izquierdo para adelantar, si el vehículo de delante circula demasiado despacio. El vehículo tiene tres objetivos: circular a la velocidad límite, evitar chocar con otros vehículos y no circular por el carril izquierdo. Utilizamos el comportamiento del vehículo con respecto a estos objetivos para medir el rendimiento del sistema. Las métricas de evaluación corresponden al número de veces que el vehículo choca, el número de veces que el vehículo se equivoca de carril y el número de veces que el vehículo supera el límite de velocidad. Las tres métricas deben reducirse al mínimo.

Para la aplicación, desarrollamos un entorno de simulación en JavaScript[2]. Este consiste en una autopista de dos carriles de 500 km. El límite de velocidad de la autopista es de 60 km/h. Otros vehículos aparecen con una probabilidad del 10 %, al menos tres pasos de tiempo después del último vehículo. Los vehículos de tráfico circulan a una velocidad de 30 km/h (para que puedan ser adelantados), siempre en el carril de circulación (el carril derecho del entorno).

Ejecución y resultados

Desarrollamos un algoritmo que aprende el comportamiento de conducción correcto para el vehículo a partir de las acciones atómicas. Observe que Auto-COP es agnóstico con respecto al origen de tales acciones en la traza de ejecución (p. ej., las acciones podrían generarse de las intervenciones humanas o automatizadas mediante RL). Al utilizar RL para generar acciones atómicas, de manera ocasional, el sistema ejecuta una acción errónea (cuando está explorando), lo que nos permite ilustrar cómo la generación de opciones corrige esas acciones, al producir adaptaciones a la ejecución del sistema en lugar de limitarse a repetir

2 Véase https://github.com/FLAGlab/DrivingAssistant

la traza de ejecución. En el proceso (atómico) de aprendizaje de acciones, el espacio de estados del vehículo es: la velocidad actual, que toma valores discretos múltiplos de 10 km/h en el intervalo [0, 70]; el carril actual, modelado como 0 para conducir por el carril derecho y 1 para el carril izquierdo; y la proximidad al vehículo de delante, dividida en el intervalo discreto [1, 4], que describe los pasos de tiempo para alcanzar al vehículo de delante. Una proximidad de 4 indica que no hay ningún vehículo delante, mientras que los valores 1, 2 y 3 denotan que hay un vehículo en proximidad inmediata y que se producirá un choque en 1, 2 o 3 pasos de tiempo en el futuro, si no se toma ninguna medida. El modelo de recompensa penaliza la colisión (−8), la conducción por el carril equivocado (−5), conducir por encima del límite de velocidad (−6), conducir demasiado despacio (−6) y proporciona una recompensa positiva cuando se conduce por el carril correcto sin un vehículo delante (8). La tasa de aprendizaje α se fija a 0,1; el factor de descuento γ a 0,6; y la selección de acciones es ε-*greedy*, con ε empezando en 0,2 en la etapa de exploración y reduciéndose a 0,001 en la etapa de exploración.

Dejamos que el sistema de entrenamiento se ejecute durante 8000 pasos, lo que genera una traza de ejecución de 8000 acciones atómicas. En la traza de ejecución, para cada acción registramos el estado actual del sistema (`speed, lane, vehicle_proximity`), la acción (atómica) ejecutada, el siguiente estado después de ejecutar la acción y la recompensa obtenida al ejecutar dicha acción.

Una vez registradas las trazas de ejecución, ejecutamos la etapa de extracción de opciones (líneas 4-15 del algoritmo 4), para construir las posibles opciones (secuencias de acciones). Con base en la traza de ejecución, en este paso generamos 26 opciones de ejecución en 13 estados diferentes. Solo se selecciona una opción en cada estado para generar la adaptación adecuada. Note que hay 72 estados en total en el entorno (de acuerdo con los valores posibles de las tres dimensiones de estado). Sin embargo, algunos de los estados no tienen ninguna opción asociada, dado que algunos estados no se experimentan durante la exploración (p. ej., no aparecen en la traza de ejecución) y solo nos interesa generar opciones para los estados en los que se requieren adaptaciones (únicamente cuando los objetivos del sistema no se cumplen). Si el vehículo circula al límite de velocidad correcto (60 km/h), en el carril correcto (0) y no hay ningún vehículo delante (4), no es necesaria ninguna adaptación.

Tabla 2.1. Espacio de estados para las opciones aprendidas y sus adaptaciones generadas

Estado	N.°	Secuencia de acciones	Frecuencia
50 0 1	**1**	**{steerLeft(), speedUp(), steerRight()}**	**29**
50 0 1	2	{steerLeft(), speedUp(), steerRight(), steerLeft(), steerRight()}	3
60 0 1	**1**	**{steerLeft(), steerRight()}**	**656**
60 0 1	2	{steerLeft(), steerLeft(), steerRight()}	1
60 0 1	3	{slowDown(), speedUp(), speedUp(), speedUp(), straight(), speedUp(), speedUp(), speedUp()}	1
60 0 1	4	{steerLeft(), straight(), steerRight(), steerLeft(), steerRight()}	1

Fuente: elaboración propia.

Para ilustrar este proceso, nos centramos en el ejemplo del comportamiento de adelantamiento (no predefinido como acción atómica), cuando un vehículo aparece delante del vehículo. La tabla 2.1 muestra todas las opciones extraídas para dos de esos estados. De forma intuitiva, adelantar a un vehículo que circula por delante se realiza mediante una secuencia de acciones atómicas; por lo tanto, la adaptación deseada en el estado [50,0,1] es la opción número 1 de la tabla, ya que adelanta al vehículo de delante a la vez que acelera hasta la velocidad objetivo de 60 km/h, para terminar en el estado objetivo [60,0,4]. Del mismo modo, en el estado [60,0,1] el comportamiento deseado es la opción 1, adelantar a un vehículo sin ninguna acción adicional. De hecho, la frecuencia de ejecución de estas acciones en la traza de ejecución muestra que estas son las opciones ejecutadas con más frecuencia para estos estados; sin embargo, la tabla también muestra opciones adicionales generadas para estos. Estas opciones se ejecutan con menor regularidad durante la exploración. Aunque no podemos asociar estas trazas a una situación exacta durante la ejecución, especulamos que la opción 2 en el estado [50,0,1] y la opción 4 en el estado [60,0,1] se deben a que el vehículo se encontró con un vehículo de tráfico en su carril inmediatamente después de adelantar a un primer vehículo, lo que resulta en tener que repetir la secuencia `{steerLeft(), steerRight()}` una vez más. Para la opción 3 en el estado [60,0,1], especulamos que la primera acción incorrecta en la secuencia `{slowDown()}` resultó en el choque del vehículo, lo que dio lugar

a una velocidad de 0, a partir de la cual el vehículo ejecutó `speedUp()` 6 veces para alcanzar el objetivo [60,0,4]. La generación de adaptaciones debería ser capaz de identificar estas secuencias como inadecuadas y aprender a extraer la opción 1 en una adaptación, para cada uno de los estados.

La lista de opciones extraídas, los estados y sus secuencias de acciones asociadas son la entrada para el proceso de aprendizaje (*adaptation generator*), en el que exploramos las opciones más a fondo a través de RL, para identificar qué opción es la más adecuada para el sistema (líneas 17-30 del algoritmo 4). Para generar adaptaciones, utilizamos el mismo espacio de estados, recompensas y parámetros de aprendizaje que para la generación de las acciones atómicas; no obstante, en este caso, permitimos el uso de todas las opciones disponibles para cada estado, junto con las opciones de aprendizaje y las cinco acciones atómicas. La opción con la mayor recompensa se selecciona para generar una adaptación para el estado dado mediante COP. Evaluamos y presentamos el rendimiento del sistema en la etapa de exploración, al comparar el rendimiento de solo el uso de las acciones atómicas y las combinaciones de acciones atómicas y adaptaciones generadas.

La figura 2.3 muestra la eficacia de las adaptaciones generadas, medida por el rendimiento respecto a los tres objetivos del sistema, al comparar la implementación de acciones atómicas y las adaptaciones generadas. Primero, se compara la cantidad de acciones ejecutadas entre los dos sistemas, durante los 8000 pasos de ejecución; cuantas más acciones atómicas ejecutadas, más adaptaciones utiliza el agente, lo que revela su idoneidad. La figura 2.3a muestra la cantidad de acciones ejecutadas por el vehículo al utilizar las adaptaciones generadas frente a las acciones atómicas. Se ejecutan un total de 1992 adaptaciones para los 8000 pasos de ejecución (p. ej., puntos de decisión), lo que resulta en un total de 15 516 acciones atómicas ejecutadas frente a las 8000 acciones atómicas en la implementación sin adaptaciones (ya que solo es posible ejecutar una acción atómica por paso de ejecución). Por lo tanto, confirmamos que el uso de opciones genera secuencias de acciones adecuadas y que el sistema aprende a utilizarlas como adaptaciones para mejorar su comportamiento, casi con el doble de eficacia de la ejecución del sistema o con la mitad de intervenciones necesarias.

En segundo lugar, comparamos la corrección de usar Auto-COP frente a acciones atómicas. La figura 2.3b muestra el comportamiento resultante de las adaptaciones generadas, en función del número de violaciones a los objetivos incurridas por el vehículo; cuantas menos violaciones, más correcto es el comportamiento. Así, observamos que hay muchas menos infracciones de carril (4) cuando se utilizan adaptaciones que cuando se emplean acciones atómicas (34).

Así mismo, el vehículo presenta un evento de colisión adicional (3 colisiones) en comparación con la ejecución de acciones atómicas (2 colisiones). Esto puede explicarse desde dos perspectivas. En primer lugar, al utilizar adaptaciones, ejecutamos efectivamente casi el doble de pasos que en el caso de las acciones atómicas, por lo que hay más del doble de vehículos en circulación (1497 en lugar de 692). Mientras que la cantidad total de choques aumenta, su porcentaje disminuye en 0,25 %. Esto constituye una mejora con respecto al comportamiento del sistema base. En segundo lugar, tras una inspección más detallada de la traza de ejecución, observamos que dos de los choques se producen durante la ejecución de acciones atómicas y no como consecuencia de la ejecución de las adaptaciones (como se señaló, el comportamiento final del sistema que implementa adaptaciones es una combinación de acciones atómicas y adaptaciones, ya que no todos los estados tienen adaptaciones asociadas). Por tanto, la cantidad de choques se reduce en un 50 % al utilizar las adaptaciones generadas. Por último, observamos que el número de infracciones de los límites de velocidad aumentó con respecto al caso en el que se utilizaron acciones atómicas, que superó el límite de velocidad 8 veces, frente a una infracción del límite de velocidad de las adaptaciones. Podemos concluir que el uso de adaptaciones generadas es beneficioso para el rendimiento del sistema, ya que las infracciones en las tres métricas disminuyen de forma significativa.

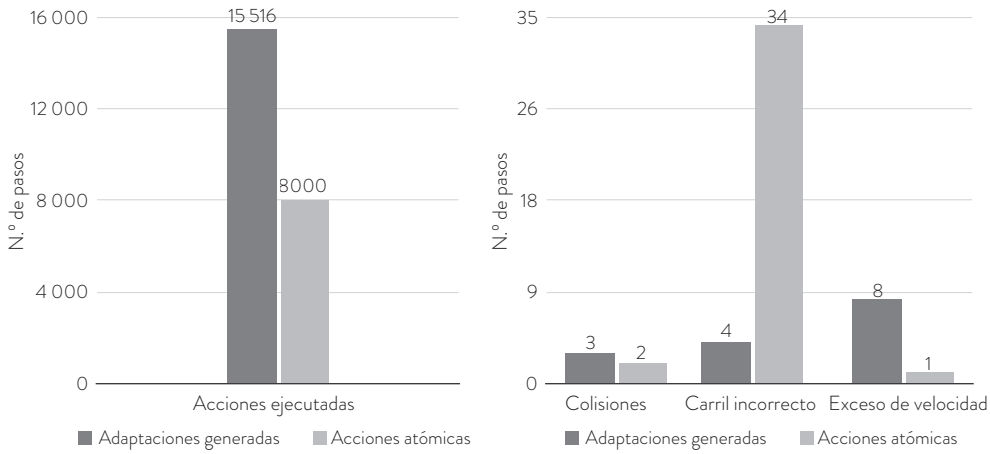


Figura 2.3. Correctitud y utilidad del comportamiento generado por las adaptaciones
Fuente: elaboración propia.

Como un ejemplo, las adaptaciones generadas para el estado [50,0,1] (contexto `BAContext5001`) se muestran en el algoritmo 10, las cuales coinciden correctamente con el comportamiento de adelantamiento deseado.

```
1  Context5001 = new cop.Contexto({ name:
   “Context5001”})
2  BAContext5001 = Trait ({
3     option: function (){
4        this.steerLeft();
5        this.speedUp();
6        this.steerRight();
7     }
8  })
9  Context5001.adapt (agent, BAContext5001)
```

Algoritmo 10. Adaptación generada para el estado `[50,0,1]`

Para integrar las adaptaciones generadas utilizamos COP, que adecua eficazmente el comportamiento de la aplicación: así, pasa de usar acciones atómicas a variaciones de comportamiento. El algoritmo 11 muestra la definición del vehículo base (líneas 1-7), junto con el comportamiento global del asistente de conducción. Para cada paso, de acuerdo con el estado actual, elegimos una acción (líneas 10-13). Si el estado actual no tiene una adaptación asociada, ejecutamos una acción primitiva. Si el estado tiene una adaptación asociada, entonces se ejecuta la variación de comportamiento correspondiente en la función `option()`. Para ejecutar la adaptación, seguimos el proceso de adaptaciones dinámicas utilizadas en COP. En primer lugar, activamos el contexto que representa el estado actual (línea 15). Esto compone la variación de comportamiento asociada al contexto con el sistema. En nuestro ejemplo, para el estado [50,0,1], la variación de comportamiento en el algoritmo 10 se compone con el sistema. El efecto de esta composición es que ahora se tiene una función `option()` definida. Luego, utilizamos la variación de comportamiento para llamar a la opción generada (línea 16). Por último, el sistema pasa a un nuevo estado, como consecuencia de la ejecución de la opción, y el contexto se desactiva (línea 17).

```
class Agent {
   speedUp() { ... }
   slowDown() { ... }
   steerRight() { ... }
   steerLeft() { ... }
   straight() { ... }
   }

while (true) {
   if(qtable[ this .currentState])
      action = qtable[ this .currentState]
   else
   action = this.randomAtomicAction()
   if (action >= Agent.actions.length) {
      eval('Context${state}'.activate ())
      agent.option()
      eval('Context${state}'.deactivate ())
   } else
   eval('agent.${actions[action]}'())
}
```

Algoritmo 11. Integración de las adaptaciones generadas mediante COP

Composición de adaptaciones en un sistema de transporte

La segunda aplicación de la validación TranCity es una aplicación para el control del servicio de transporte de una ciudad. El comportamiento base de TranCity permite supervisar los servicios de autobús de una ciudad, observando la ocupación del sistema (autobuses y estaciones), la frecuencia de autobuses, el número de autobuses que operan en una ruta y el estado de las carreteras (p. ej., si están bloqueadas). La figura 2.4 muestra una vista de TranCity y resalta la ruta utilizada en la evaluación (R4).

Como las condiciones de movilidad en la ciudad están en continuo cambio, el sistema necesita tomar medidas que modifiquen el comportamiento básico de TranCity en función del contexto de ejecución. Por ejemplo, si una estación supera su capacidad, se activa un contexto y se ejecutan las adaptaciones asociadas. La tabla 2.2 presenta los contextos y su variación de comportamiento asociada utilizados en la evaluación.

Durante la ejecución del sistema, varios contextos pueden activarse de forma simultánea. El sistema debe decidir qué adaptación o combinación de adaptaciones es la mejor para maximizar el rendimiento.

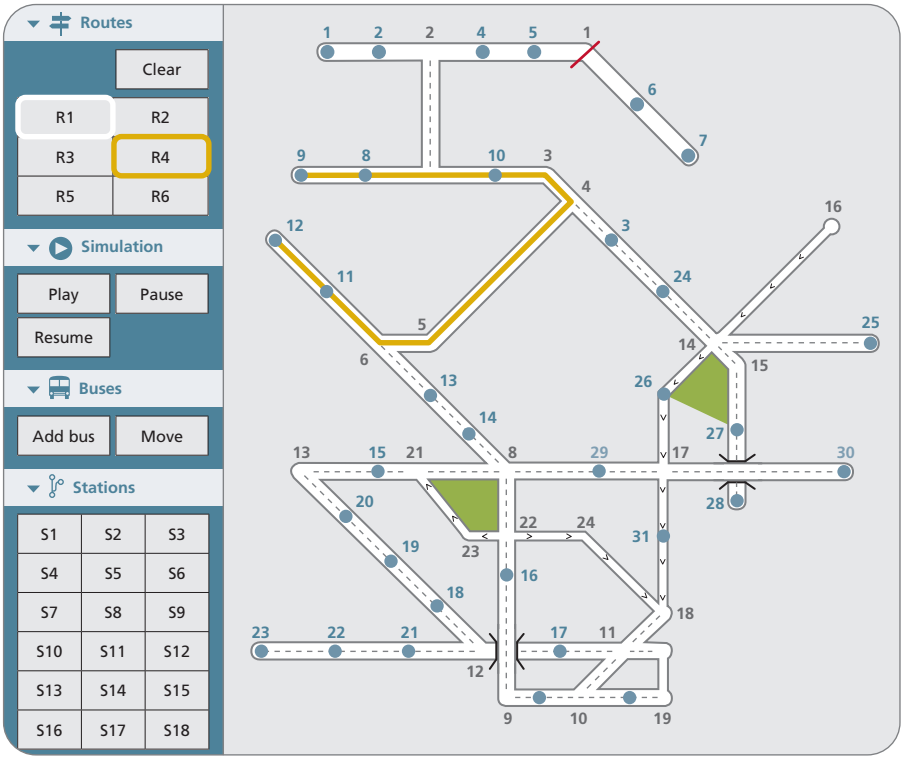


Figura 2.4. Rutas de bus en el sistema de TranCity
Fuente: elaboración propia.

Tabla 2.2. Contextos y sus variaciones de comportamiento asociadas

Contexto	Variación de comportamiento
`FullBus`	Saltar la próxima estación, enviar un bus a la estación llena.
`FullStation`	Cerrar la estación, enviar un bus, redirigir los pasajeros a otras estaciones.
`DepotEmpty`	Retornar un bus al parqueadero.

Fuente: elaboración propia.

Escenarios y parámetros

Evaluamos el comportamiento y el rendimiento de la combinación de adaptaciones en los siguientes escenarios específicos. Cada escenario se ejecutó durante 5000 episodios de aprendizaje, de los cuales dos tercios de los episodios se dedicaron a, por ejemplo, aprender los valores *Q* y *W*, y un tercio a explorar el comportamiento aprendido.

- Escenario 1: adaptaciones independientes/complementarias, en el que los contextos `FullBus` y `FullStation` pueden estar activos de forma simultánea.
- Escenario 2: adaptaciones conflictivas, en el que los contextos `Depot Empty` y `FullStation` pueden estar activos de forma simultánea.

En cada uno de los escenarios, comparamos el rendimiento de tres variaciones, dos alternativas de composición por medio de ComInA y la adaptación predefinida:

- ComInA individual. El sistema decide dinámicamente qué adaptación ejecutar para uno de los contextos activos.
- ComInA componible. El sistema puede ejecutar alguna de las adaptaciones individuales o una composición de las adaptaciones asociadas a ambos contextos.
- Adaptación predefinida. El comportamiento base proporcionado por enfoques existentes. La adaptación que se ejecuta en caso de múltiples contextos activos está predefinida. Nosotros tenemos dos líneas de base, una para cada uno de los dos contextos activos, siendo siempre un "ganador" predefinido.

El rendimiento del sistema se mide a través de dos métricas:

- El número de alertas de contexto, donde menos alertas significa un mejor rendimiento. Las alertas de contexto se producen como resultado de un funcionamiento anormal del sistema. Estas pueden provenir de un aumento/disminución de pasajeros en los autobuses/estaciones; sin embargo, las alertas no resueltas seguirán produciéndose en cada paso de tiempo hasta que se resuelvan. Las estrategias de adaptación efectivas tendrán un menor número de alertas.
- Retraso de los pasajeros, calculado en cada paso temporal como la diferencia entre el retraso acumulado total de todos los pasajeros en el paso temporal t y el retraso que experimentaron en el paso temporal anterior $t - 1$. Idealmente, si el sistema funciona sin problemas, el retraso adicional introducido en cada paso temporal es 0.

La combinación de las implementaciones de ComInA, comparado con las líneas base, permite evaluar si aprender la composición de adaptaciones efectivamente mejora el rendimiento del sistema, contra usar reglas de composición predefinidas. En particular, observamos si ComInA puede detectar relaciones entre adaptaciones activas de forma simultánea y aprender a componerlas o a ejecutar solo una de ellas.

Ejecución y resultados

Para el escenario 1, la figura 2.5 muestra el retraso de los pasajeros y la tabla 2.3 presenta el número de alertas lanzadas.

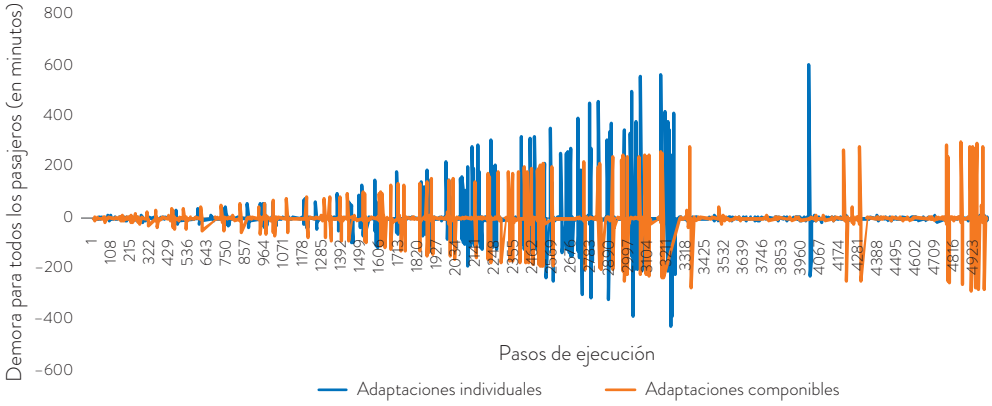


Figura 2.5. Demora por cada paso de tiempo para el escenario 1
Fuente: elaboración propia.

Tabla 2.3. Número de alertas de contexto en el escenario 1

	ComInA individual	ComInA componible	Predefinidas BusStation	
`FullBus`	2170	1766	1236	3315
`FullStation`	303	234	175	477
`Ambos`	173	109	48	320

Fuente: elaboración propia.

Intuitivamente, las adaptaciones para los dos contextos `FullStation` y `FullBus` son compatibles o complementarias, ya que ambas envían un bus adicional desde el parqueadero hasta una estación específica. Además, la adaptación `FullStation` cierra la estación para los nuevos pasajeros, hasta que se libere la capacidad. Observamos un mejor rendimiento general cuando se permite la composición de las dos adaptaciones (en términos de menor retraso máximo y número de alertas) al caso donde el sistema tiene que elegir una sola adaptación para ejecutar (el descenso del retraso alrededor del paso 3700 corresponde a que el sistema pasa a explorar las composiciones aprendidas). De igual manera, se observa que en 5000 pasos el número de veces que los contextos `FullStation` y `FullBus` se activan simultáneamente varía entre 48 y 320, lo que indica la frecuencia de estas situaciones y la necesidad de enfoques dinámicos para resolverlas. También observamos que en la implementación base, en la que solo se ejecuta la adaptación `FullBus`, se produce el menor retraso promedio de los pasajeros (que oscila en los 15 minutos, frente a 300 minutos en el caso compuesto) y el menor número de alertas. Si bien los resultados muestran un mejor rendimiento global, dar siempre prioridad a la adaptación `Bus` podría desequilibrar el sistema. Para detallar mejor el impacto a nivel de pasajero, bus o estación, se debe incluir una métrica de equidad, lo que evita que solo se cumpla una adaptación.

Para el escenario 2, la tabla 2.4 muestra el número de alertas emitidas. Intuitivamente, la relación entre las adaptaciones asociadas a los contextos `FullStation` y `DepotEmpty` son conflictivas, ya que una requiere del envío de un bus a una estación y la otra, retirar el bus del sistema para que sirva de reserva de emergencia en el parqueadero. Por lo tanto, para este escenario no proporcionamos una solución base al ejecutar una adaptación, dado que retirar buses continuamente del sistema eliminaría el servicio completo. En términos de retraso (durante la exploración), ambas implementaciones consiguen una métrica estable; no obstante, en términos de alertas lanzadas, permitir adaptaciones compuestas es más de un 20 % peor en todos los tipos de alertas. Esto indica que las métricas del sistema identifican correctamente que las adaptaciones son conflictivas y no deben componerse, a diferencia del escenario 1, en el que las adaptaciones compuestas mejoraron el rendimiento.

Tabla 2.4. Número de alertas de contexto para el escenario 2

	ComInA individual	ComInA componible
`FullStation`	2380	2761
`DepotEmpty`	713	1055
`Compuestas`	623	981

Fuente: elaboración propia.

Conclusión

Este estudio presenta el beneficio de usar sistemas de aprendizaje aplicados al caso de la ingeniería de *software*, en general, y a los SAS, en particular. Más aún, por medio de técnicas de RL, proponemos una solución a un problema hasta ahora abierto en los SAS: adaptarse a situaciones completamente desconocidas (*unknown-unknowns*).

Para lograr la adaptación del sistema a situaciones desconocidas formulamos una solución que genera adaptaciones y su mejor composición, a partir del conocimiento adquirido por medio de interacciones con el ambiente. El RL se utiliza para detectar los estados en los cuales se requiere una adaptación y las secuencias de acciones asociadas a dichos estados. Las secuencias de acciones se seleccionan utilizando el concepto de opciones o macroacciones, lo que promueve la mejor secuencia de acciones para un estado dado. Las secuencias de acciones y los estados en los que deben tener efecto se generan como adaptaciones de COP, definidas respectivamente como variaciones de comportamiento y contextos. De esta forma, se crean módulos que dan la flexibilidad para componer y descomponer dinámicamente, con el fin de adaptar el comportamiento del sistema sin que este sea prescrito. Una vez generadas las adaptaciones individuales, proponemos un proceso para aprender la mejor forma de componerlas. Este proceso utiliza una aproximación multiobjetivo basada en el aprendizaje W, donde, por medio de la interacción, el sistema aprende cuál es la mejor combinación de adaptaciones para un estado donde múltiples adaptaciones son aplicables.

Nuestros resultados demuestran, a través de dos aplicaciones diferentes, que la solución propuesta es efectiva en la generación y composición de adaptaciones que son útiles para alcanzar el objetivo del sistema. Por último, la validación demuestra que las técnicas de aprendizaje, como el RL, pueden ser aplicadas en el dominio de la ingeniería de *software* para facilitar y mejorar el desarrollo de los SAS.

Referencias

Cabrera, C., Paleyes, A. y Lawrence, N. D. (2024). Self-sustaining software systems (S4): Towards improved interpretability and adaptation. En *Proceedings of the 1st International Workshop on New Trends in Software Architecture* (pp. 5-9). Association for Computing Machinery.

Cardozo, N. (2016). Emergent software services. En *Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software* (pp. 15-28). Association for Computing Machinery.

Cardozo, N. y Dusparic, I. (2018). *Generating software adaptations using machine learning* [ponencia]. International Workshop on Machine Learning techniques for Programming Languages. Amsterdam, Países Bajos.

Cardozo, N. y Dusparic, I. (2020). Learning runtime compositions of interacting adaptations. En *Proceedings of the International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)* (pp. 108-114). Association for Computing Machinery.

Cardozo, N. y Dusparic, I. (2021). Adaptation to unknown situations as the holy grail of learning-based self-adaptive systems: Research directions. En *International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (pp. 252-253). Institute of Electrical and Electronics Engineers (IEEE).

Cardozo, N. y Dusparic, I. (2022). Next generation context-oriented programming: Embracing Dynamic generation of adaptations. *Journal of Object Technology*, *21*(2), 1-6. http://dx.doi.org/10.5381/jot.2022.21.2.a5

Cardozo, N. y Dusparic, I. (2023). Auto-COP: Adaptation generation in context-oriented programming using reinforcement learning options. *Information and Software Technology*, *164*, 107308.

Cardozo, N., Dusparic, I. y Castro, J. H. (2017). Peace COrP: Learning to solve conflicts between contexts. En *International Workshop on Context-Oriented Programming*. https://doi.org/10.1145/3117802.3117803

Castagna, A. y Dusparic, I. (2022). Multi-agent transfer learning in reinforcement learning-based ride-sharing systems. En *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)* (pp. 120-130). Science and Technology Publications.

D'Angelo, M., Gerasimou, S., Ghahremani, S., Grohmann, J., Nunes, I., Pournaras, E. y Tomforde, S. (2019). On learning in collective self-adaptive systems: State of practice and a 3D framework. En *International Symposium on Software Engineering for Adaptive and Self-Managing*

Systems (pp. 13-24). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/SEAMS.2019.00012

Elfwing, S., Uchibe, E., Doya, K. y Christensen, H. I. (2004). Multi-agent reinforcement learning: Using macro actions to learn a mating task. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, *4*, 3164-3169.

Girgin, S. y Polat, F. (2005). Option discovery in reinforcement learning using frequent common subsequences of actions. En *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)* (pp. 371-376). Institute of Electrical and Electronics Engineers (IEEE).

González, S., Mens, K., Colacioiu, M. y Cazzola, W. (2013). Context traits: Dynamic behaviour adaptation through run-time trait recomposition. En *Proceedings of International Conference on Aspect-Oriented Software Development* (pp. 209-220). Association for Computing Machinery. https://doi.org/10.1145/2451436.2451461

Hirschfeld, R., Costanza, P. y Nierstrasz, O. (2008). Context-oriented programming. *Journal of Object Technology*, *7*(3), 125-151. http://www.jot.fm/issues/issue_ 2008_03/article4/

Humphrys, M. (1995). *W-learning: Competition among selfish Q-learners* (inf. téc. UCAM- CL-TR-362). University of Cambridge, Computer Laboratory. https://doi.org/10.48456/tr-362

Humphrys, M. (1996). Action selection methods using reinforcement learning. En *Proceedings of the International Conference on Simulation of Adaptive Behavior* (pp. 135-144). Association for Computing Machinery.

Kephart, J. O. y Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41-50.

Khan, N. A., Brujic-Okretic, V. y Khaddaj, S. (2016). Self-adaptive service driven architecture for intelligent transport system. En *IEEE Intl Conference on Computational Science and Engineering and IEEE Intl Conference on Embedded and Ubiquitous Computing and Intl Symposium on Distributed Computing and Applications for Business Engineering* (pp. 669-672). https://doi.org/10.1109/CSE-EUC-DCABES.2016.258

McGovern, A. y Sutton, R. S. (1998). *Macro-actions in reinforcement learning: An empirical analysis*. University of Massachusetts.

Randløv, J. (1998). Learning macro-actions in reinforcement learning. En *Proceedings of the International Conference on Neural Information Processing Systems* (pp. 1045-1051). Association for Computing Machinery. http://dl.acm.org/citation.cfm?id=3009055.3009201

Sutton, R. y Barto, G. (2018). *Reinforcement learning, an introduction* (2.ª ed.). Massachusetts Institute of Technology Press.

Rutten, E., Marchand, N. y Simon, D. (2017). Feedback control as MAPE-K loop in autonomic computing. En R. de Lemos, D. Garlan, C. Ghezzi y H. Giese (Eds.), *Software Engineering for Self-Adaptive Systems III. Assurances: International Seminar, Dagstuhl Castle, Germany, December 15-19, 2013, Revised Selected and Invited Papers* (pp. 349-373). Springer International Publishing.

Salehie, M. y Tahvildari, L. (2009). Self-adaptive software: Landscape and research challenges. *ACM Transactions on Autonomous and Adaptive Systems*, *4*(2), 1-42. https://doi.org/10.1145/1516533.1516538

Salvaneschi, G., Ghezzi, C. y Pradella, M. (2012). Context-oriented programming: A software engineering perspective. *Journal of Systems and Software*, *85*(8), 1801-1817. https://doi.org/10.1016/j.jss.2012.03.024

Sanabria, M., Dusparic, I. y Cardozo, N. (2024). Learning recovery strategies for dynamic self-healing in reactive systems. En *SEAMS 24: Proceedings of the 19th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (pp. 133-142). Association for Computing Machinery. https://doi.org/10.1145/3643915.3644097

Schmerl, B., Andersson, J., Vogel, T., Cohen, M. B., Rubira, C. M. F., Brun, Y., Gorla, A., Zambonelli, F. y Baresi, L. (2017). Challenges in composing and decomposing assurances for self-adaptive systems. En R. de Lemos, D. Garlan, C. Ghezzi y H. Giese (Eds.), *Software engineering for self-adaptive systems III. Assurances* (pp. 64-89). Springer International Publishing.

Stolle, M. y Precup, D. (2002). Learning options in reinforcement learning. En *Proceedings of the International Symposium on Abstraction, Reformulation and Approximation* (pp. 212-223). Springer International Publishing.

Sutton, R. S., Precup, D. y Singh, S. P. (1998). Intra-option learning about temporally abstract actions. *International Conference on Machine Learning*, *98*, 556-564.

Watkins, C. J. C. H. y Dayan, P. (1992). Technical note: Q-learning. *Machine Learning*, *8*(3), 279-292.

Zhao, H., Liu, H., Leung, Y.-W. y Chu, X. (2018). Self-adaptive collective motion of swarm robots. *IEEE Transactions on Automation Science and Engineering*, *15*(4), 1533-1545.

Zhu, W., Oguz, S., Heinrich, M. K., Allwright, M., Wahby, M., Christensen, A. L., Garone, E. y Dorigo, M. (2024). Self-organizing nervous systems for robot swarms. *Science Robotics*, *9*(96).

LA INTELIGENCIA ARTIFICIAL Y EL ARCHIVO MULTIMODAL EN HISTORIA

Laura Manrique Gómez,
Jaime Huberto Borja Gómez

Nota a manera de preámbulo

Laura estuvo comprometida y muy entusiasmada con la elaboración conjunta de este artículo. No es fácil encontrar colegas con la calidez humana y la empatía para producir investigaciones conjuntas. En Laura encontré no solo la persona que desarrollaba mi dirección de tesis más disciplinada y compleja, en mi larga experiencia académica, sino también una colega con quien las ideas fluían a la par. Lamentablemente falleció unas semanas antes de la publicación de este capítulo. Su investigación doctoral era pionera en el campo de los estudios históricos en Colombia; de hecho, existen pocos ejemplos globales de su propuesta: integrar herramientas de la inteligencia artificial a la historia. En este caso, aplicaba *machine learning* y procesamiento de lenguaje natural para descifrar tendencias y patrones en las formas como se expresaban y se entendían las emociones y sentimientos en la prensa latinoamericana del siglo XIX. Historical Ink es el nombre que le dio a todo su proyecto académico, del cual quedan varios artículos publicados. Este aquí presente forma parte de las discusiones contextuales que sostuvimos para el desarrollo de su tesis. Sin duda, su prematura muerte es una gran pérdida en todo sentido. Este artículo es un homenaje a su memoria.

Jaime Humberto Borja Gómez

La memoria de Laura Manrique permanece en estas páginas como testimonio de su rigor académico, su sensibilidad humana y su capacidad visionaria para integrar el procesamiento de lenguaje natural en los estudios históricos. En cada proyecto que compartimos se hizo evidente su talento para tender puentes entre lo cualitativo y lo cuantitativo, su disciplina inquebrantable y su generosidad como colega. Este capítulo, elaborado junto con su asesor Jaime Borja, refleja la profundidad de su pensamiento y constituye, a la vez, un homenaje a su legado como investigadora y como ser humano excepcional.

Rubén Manrique

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.03

Introducción

Los desarrollos tecnológicos que se habían iniciado durante la Segunda Guerra Mundial se potenciaron en las décadas de 1950 y 1960. Entre ellos, la computación logró grandes avances con la invención de los circuitos integrados y, posteriormente, con el microchip (Ceruzzi, 2012). La aparición de la computadora personal en los años ochenta, así como avances en las conexiones de redes desde la década anterior, hizo inminente la revolución digital. La popularización de la internet y la web en los primeros años de los noventa fue posible por la globalización política y económica de aquella década, lo que hizo patente la profunda huella que estaban dejando las emergentes tecnologías digitales y electrónicas en cada estrato de la sociedad. Este fenómeno denominado *cultura digital* (Levy, 2007) se fortaleció, mientras dispositivos como la televisión y los computadores daban sus primeros pasos hacia la ubiquidad. En menos de diez años, a comienzos del nuevo milenio, el "salto digital" impactaba todos los espacios de la cultura global: de la vida cotidiana a las experiencias políticas globales; de las emociones individuales y grupales a las nuevas formas de comercio y consumo. Todo había sido permeado por lo digital: emergía la *era de la información*.

Este fenómeno en el que hemos vivido los últimos treinta años ha sido tan impactante que se le equipara a los profundos efectos de la invención de la imprenta, los cuales dieron origen a la modernidad. Por supuesto, las ciencias sociales no quedaron atrás. Todos estos elementos proporcionaron un sustrato sobre el que florecerían intersecciones inesperadas entre la computación y las humanidades, desencadenando tanto la concepción de la historia digital como

de las humanidades digitales[1], es decir, el cruce investigativo, temático y metodológico entre tecnologías y humanidades. Las dinámicas de profundización entre estos diálogos de exploración interdisciplinarios cobraron un singular impulso en la década del 2010, cuando se inició una etapa denominada tercera primavera de la inteligencia artificial (IA), marcada por el entusiasmo generado en torno a estas tecnologías, con los desarrollos de *deep learning* en la Universidad de Toronto e hitos como el sucedido en el 2011, cuando la inteligencia de Watson, alojada en un supercomputador de IBM, venció a Ken Jennings, campeón humano del popular concurso estadounidense *Jeopardy!* Dicho juego implica para los participantes dominar unas habilidades lingüísticas, analíticas, de procesamiento de información contextual y de cultura general que hasta ese momento solo podían ser imaginadas en un humano.

Este nuevo florecimiento de las tecnologías de IA marcó un antes y un después en la forma en la que conceptualizamos y nos aproximamos al presente y al futuro, pero también a nuestro pasado, a la forma como hacemos historia. Nuevas preguntas surgen en torno a cómo nuestras interacciones con la información y el conocimiento están siendo transformadas en un mundo cada vez más mediado por lo digital. Con la ciencia de datos y la IA surgiendo como catalizador, se articulan nuevas narrativas sobre la materialidad y la dimensión temporal de la historia. Entre el 2012 y el 2016 se comenzaron a observar los primeros impactos de la IA en algunos sectores de historiadores interesados en desarrollar nuevas metodologías vinculadas con lo digital. Esta transformación iba más allá de una simple acumulación de información: reformulaba los métodos a través de los que se entendía el pasado, introduciendo la naturaleza dinámica de los datos, los cuales siguen siendo concebidos como entidades singulares y estáticas por una buena parte de los historiadores. Las nuevas propuestas buscan alternativas para aproximarse a la masividad y multidimensionalidad de los datos, ya que la información puede estar contenida en textos, imágenes, videos, impulsos electromagnéticos, frecuencias sonoras, entre otras fuentes. En consecuencia, se ha abierto un panorama de comprensión crítica del mundo para las ciencias sociales y las humanidades, por medio de su abstracción en representación de datos[2]. En particular para la historia, la datificación del pasado

1 El debate y la investigación sobre la formación de las humanidades computacionales, las humanidades digitales y su impacto en las disciplinas sociales y humanas es abundante. Para una revisión de sus problemáticas centrales, véase Le Deuff (2018).

2 Una definición oficial es proporcionada por la Comisión Económica de las Naciones Unidas para Europa (Unece, 2000): "*Dato* es la representación física de información de manera adecuada para comunicación, interpretación o procesamiento por seres humanos o por

ha puesto un nuevo reto: la urgencia de desarrollar nuevas competencias analíticas y metodológicas.

En este contexto, el presente capítulo pretende examinar cómo se ha dado este diálogo constructivo entre historiadores y científicos de datos, en el cual se resalta el potencial colaborativo inherentemente interdisciplinario de la práctica histórica contemporánea. Este nuevo escenario no solo ha ampliado el bagaje de métodos y técnicas a disposición del historiador, sino que ha impulsado una reflexión profunda sobre la naturaleza epistemológica del conocimiento histórico y sobre cómo este puede ser articulado dentro de marcos digitales dinámicos. Esta visión propone una historiografía rejuvenecida, adaptada a las complejidades del presente y preparada para dialogar con los desafíos del futuro, subrayando la importancia de una alfabetización digital que permita interpretar con propiedad crítica y profundidad el curso de nuestra historia. Para presentar este panorama, este texto se desarrolla en tres partes. Primero, se aborda un estado del arte sobre el avance del procesamiento de archivos multimodales con IA. La segunda parte da cuenta de una propuesta desde el sur global, la contribución del proyecto Arte Colonial Americano (ARCA) de la Universidad de los Andes, en cuanto a sus métodos y resultados principales. Por último, se presenta una reflexión sobre el futuro de la historia al incorporar en la práctica de investigación disciplinar la multidimensionalidad de las fuentes históricas y la discusión epistemológica de posibilidades, como el análisis de discurso multimodal, y retos como las metodologías para indagar sesgos éticos.

Del uso de fuentes históricas multimodales

Las formas de hacer investigación histórica han tenido varias etapas desde que se consolidó como ciencia. En el siglo XIX, la disciplina depositó toda su confianza en el documento escrito como la principal fuente que podía contar los hechos del pasado. De este modo, hasta la década de los sesenta la principal fuente para hacer historia era el documento de archivo. Debido a los cambios epistemológicos y a la transformación interdisciplinaria de las ciencias sociales, durante esa década se comenzó a emplear cada vez con mayor frecuencia fuentes afines a otras disciplinas: pinturas, fotografías, literatura, prensa, música, entre otras, se convirtieron en las nuevas fuentes para reconstruir el pasado. Hoy en día, esas siguen siendo fuentes, pero las documentales son con las

---

medios automáticos" (p. 6 , traducción propia). También véase Organización para la Cooperación y el Desarrollo Económico (OECD, 2008). Sobre su impacto en las ciencias sociales, véase Gitelman (2013).

que trabajan la mayoría de las personas dedicadas a la historia. Con el auge de la cultura digital, a comienzos de los 2000 se empezaron a emplear elementos digitales básicos, buena parte de ellos derivados del ímpetu de la digitalización. En este horizonte emerge una propuesta vital: las metodologías de la ciencia de datos y las herramientas de IA pueden no solo apoyar, sino enriquecer la investigación histórica, al transformar los archivos tradicionales en repositorios digitales expansivos, susceptibles a análisis computacionales a gran escala. Esta situación pone presente que la historia ya no reposaba únicamente en los anaqueles de bibliotecas y colecciones de manuscritos antiguos.

En este contexto, surgió a inicios de los 2000 una nueva corriente o forma de hacer historia denominada *historia digital* (Weller, 2013). En un comienzo, se trataba de utilizar lo digital para producir herramientas para la investigación o generar nuevos modelos de comunicación masiva del conocimiento histórico. Con el transcurrir de la década, y en la medida en que se complejizaban las tecnologías digitales, en la década de 2010 surgieron nuevas posibilidades de archivos multimodales digitales al acercar la IA al quehacer de los historiadores. La implementación de técnicas de *lectura distante* (Moretti, 2000), junto con la exploración de algoritmos y análisis de datos, refuerzan la posibilidad de desentrañar patrones y estructuras hasta entonces ocultas dentro de vastas colecciones de información. Quienes han acogido estas técnicas para el discurso histórico han transitado hacia una interpretación del documento no como una reliquia fija, sino como un vector dinámico de conocimiento, indicativo de una concepción más amplia y epistemológica de lo que representa una evidencia histórica. Esta reconsideración del archivo en la labor historiográfica ha introducido preguntas sobre las metodologías y las interpretaciones del pasado. Pero también ha abierto posibilidades para la reutilización de conjuntos de datos de colecciones de información preexistentes, lo que permite la exploración de nuevas fuentes históricas multimodales, brindando espacios inéditos para la investigación, que contempla ahora la naturaleza iterativa y experimental del análisis de datos.

De esta forma, contrario a lo que podría intuirse, la IA acerca más al historiador al pasado, porque le permite explorar en extensión y profundidad fuentes de textos, manuscritos, imágenes y otros múltiples artefactos históricos. Este nuevo archivo histórico multimodal permite integrar las ideas *core* de apertura de las humanidades digitales: conocimiento abierto, colaboración, diversidad y experimentación (Spiro, 2012), lo cual entraña una dimensión de democratización del pasado. Sin embargo, esta propuesta no ha estado exenta de críticas que ven en la historia digital un conjunto de reflexiones y promesas de lo que vendría para la disciplina histórica, una especie de "futuro perpetuo" (Blevins, 2016). Pero en realidad esas promesas se han venido materializando en proyectos de

investigación que integran el uso de ciencia de datos y modelos de *machine learning* (ML).

Una exploración panorámica de algunas de las investigaciones y proyectos recientes que han utilizado la IA para el análisis de archivos multimodales permite apreciar los avances que se han realizado a partir de estos trabajos pioneros y dan una idea de las posibilidades y retos que se concretan. En particular, los historiadores vinculados con lo digital están trabajando en aplicar modelos de IA para complementar su trabajo analítico, y extraer información semántica y de comparación diacrónica de textos antiguos. Pero también se aplica en otros espacios donde la IA es útil, por ejemplo, en páginas web, mapas, planos, información geoespacial, audios, música y sonidos, imágenes, arte y cultura visual, videos, la actividad artística del *performance* y la gestualidad de los cuerpos, incluso los silencios y ausencias que desvela el propio archivo. Analizaremos a continuación los principales ejes mencionados, lo cual afecta incluso las formas de argumentación.

Un nuevo tratamiento para las fuentes textuales

No es sorpresivo que la mayor parte de avances que se hayan realizado en la integración de la IA con el oficio propio de la historia sean trabajos enfocados en la dimensión textual del archivo. El primer gran impacto que recibieron las ciencias sociales y humanas de la cultura digital fue los procesos masivos de digitalización de archivos documentales y medios impresos, como las bibliotecas y la prensa. Esto comenzó a suceder en la década de 1980, también en la medida en que se popularizaban nuevos medios electrónicos de almacenamiento, como el disco compacto. La popularización de internet y el abaratamiento de los costos de digitalización en los años noventa permitió que los primeros humanistas digitales se centraran en la digitalización de colecciones curadas durante las últimas tres décadas (Berry y Fagerjord, 2017; Dobson, 2021). Esta oleada de archivos digitalizados abrió nuevas posibilidades para proyectos clásicos de humanidades digitales, como la Text Encoding Initiative (TEI)[3], y nuevos conceptos, como el de *lectura distante* (Moretti, 2000), aplicados a través de técnicas de visualización (Graham *et al*., 2016; Schreibman *et al*., 2016). Este proceso permitió que las bibliotecas y colecciones rompieran las limitaciones geográficas y temporales, pero la gran cantidad de información disponible también hizo que la investigación fuera más compleja (Milligan, 2022). La ciencia

3 Véase https://tei-c.org/

de datos y la IA aportan nuevas posibilidades metodológicas que permiten a los investigadores explorar tanto archivos web como documentos analógicos antiguos, incluso manuscritos. Así, los investigadores pueden acceder a grandes conjuntos de datos, producto de un trabajo de curaduría realizado por los primeros humanistas e historiadores digitales.

Un ejemplo, ya clásico, es el proyecto Old Bailey[4], que durante veinte años y en un proceso que cubre varias etapas ha puesto en línea cerca de 198 000 juicios del tribunal penal londinense entre 1674 y 1913. En su última versión, el sitio cuenta con sistemas de búsqueda avanzado y una *application programming interface* (API), que permite ajustar la información a los intereses del investigador. Como este caso, buena parte de los archivos históricos y las bibliotecas tienen diferentes porcentajes de digitalización de sus colecciones, pero son menos las instituciones que han avanzado más allá de la simple digitalización para introducir herramientas con tecnologías recientes que ofrezcan un mayor aprovechamiento de la información. Por supuesto, este es un caso en el que la brecha norte-sur es muy notoria. El problema no reposa solo en las instituciones, pues también en los últimos años algunas instituciones privadas y universidades han creado programas de IA para la extracción de grandes volúmenes de datos almacenados en bibliotecas y archivos históricos. Este es el caso del proyecto Translantis de la Universidad de Utrech, que se apoya en *software* de código abierto (xTAS) de la Universidad de Amsterdam[5].

Por otro lado, las técnicas de ML permiten, por ejemplo, usar algoritmos de *clustering* no supervisados, como el *modelado de tópicos*, que consiste en técnicas computacionales que generan agrupamientos de palabras de acuerdo con su proximidad vectorial. Son modelos computacionales que permiten inferir, sin necesidad de que un humano le indique qué palabras son más próximas entre sí, información de similitudes en un conjunto de palabras u otras unidades textuales. Estas técnicas no supervisadas posibilitan una primera aproximación interesante a grandes corpus de textos, y aunque inicialmente no significaron por sí mismas un avance disruptivo en la interpretación histórica (Blevins, 2016), siguen teniendo un potencial importante en la medida en que ayudan al historiador a identificar piezas del "rompecabezas" o puntos por conectar para comprender mejor el pasado (Villamor *et al*., 2023), en especial si se combinan con modelos supervisados de *embeddings contextuales*, como en el caso de la identificación del cambio histórico semántico del uso de las palabras en

4 Véase https://www.oldbaileyonline.org/

5 Véase https://translantis.wp.hum.uu.nl/projects/biland/

español durante el siglo XIX en Latinoamérica (Montes *et al*., 2024) o para el posible reconocimiento de autorías con técnicas de estilometría (García Serrano y Menta Garuz, 2021, 2022; Hu *et al*., 2023). Lo importante es que la aproximación de esta metodología sea intervenida durante todo el proceso por un grupo interdisciplinario de historiadores y científicos de datos, de tal manera que se asegure una aproximación crítica de los algoritmos y una curaduría de los datos de entrenamiento, las decisiones de modelación y los resultados. Tal como Lorella Viola (2023) lo presenta, los objetos digitales sobrepasan el debate de si son auténticos o no, ya que no son objetos terminados y su naturaleza implica la intervención de múltiples agentes, y no únicamente los individuos con conocimiento técnico. Viola denomina su propuesta como un enfoque *posauténtico para el modelado de tópicos*[6].

Otras posibilidades son las técnicas supervisadas, que requieren un conjunto de datos etiquetados por expertos, para indicarle al algoritmo el "deber ser" y entrenarle para que pueda "aprender" a realizar tareas de clasificación específicas. Este tipo de técnicas ha generado resultados más significativos para los historiadores digitales, porque son personalizables dependiendo del proyecto y las características propias del corpus de documentos. Estas técnicas tienen amplias posibilidades de aplicación, que pueden ir desde la identificación de posibles sesgos de género en corpus de prensa histórica (Wevers, 2019), la aplicación de *article separation,* mediante técnicas de *layout,* y el entrenamiento de modelos *deep learning* de *embeddings* para el perfeccionamiento de la datificación de documentos antiguos mediante uso de herramientas de *optical character recognition* (OCR), para transcribir textos (Manrique-Gómez *et al*., 2024), reconocer autorías —como en el caso del manuscrito antes desconocido de Lope de Vega (Cuéllar, 2023)— o incluso identificar textos líricos insertos en publicaciones seriadas antiguas (Soh *et al*., 2023). Este tipo de técnicas se han aplicado en particular a fuentes históricas impresas, principalmente prensa y libros del siglo XIX[7].

Uno de los principales retos que tienen aquellos historiadores que trabajan con periodos premodernos, anteriores al siglo XIX, es el trabajo paleográfico, es decir, la transcripción y comprensión de formas de escritura antigua. La IA aplicada

6 "Post-authentic approach to topic modelling" (Viola, 2023, p. 94).

7 Algunos ejemplos de proyectos de historia digital que exploran el uso de herramientas IA son Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines, véase http://www.viraltexts.org; Project Impresso: Media Monitoring the Past, véase https://impresso-project.ch/; y News Eye, véase https://www.newseye.eu/

a la historia ha avanzado en cuanto a su uso con propósitos de interpretación semántica, de lo cual hay ejemplos exitosos que aportan soluciones paleográficas para la transcripción no solo de textos manuscritos antiguos, sino también de lenguas muertas, como es el caso de los proyectos Transkribus, Kitab, D-Scribes y Emlo Project[8]. Los modelos supervisados amplían el espectro de posibilidades para analizar archivos históricos efímeros o no catalogados, entre los que se encuentran las notas sueltas, los diarios personales (Fields *et al*., 2023), la correspondencia (Spina, 2022) y los panfletos que pueden dan indicios, por ejemplo, de los discursos virales en la época victoriana (Hardaker *et al*., 2023).

Estas aplicaciones requieren del uso creciente de los modelos avanzados de *deep learning* basados en *word embeddings*. Estos modelos de espacio vectorial son representaciones numéricas de textos que describen la distribución y las relaciones modeladas entre el vocabulario extraído de los textos, lo cual lleva cierto refinamiento matemático para el científico social, pero también implican una dimensión teórica nueva de hermenéutica vectorial que permita interpretar estos objetos digitales y sus procesos subyacentes (Dobson, 2022). Por ese motivo, su uso para propósitos de interpretación histórica o predicción de comportamientos humanos debe ser abordado por los científicos sociales con base en un sólido pensamiento crítico, cierto nivel de conocimiento técnico y trabajo colaborativo interdisciplinario. Tal como indica la historiadora digital Jo Guldi (2023), este tipo de propuestas pueden tener ciertos sesgos subjetivos, tanto de la interpretación del historiador como de los datos, los parámetros y los algoritmos con los que se entrenan los modelos de IA.

Inteligencia artificial y las fuentes históricas no documentales

Como se mencionó, una de las grandes transformaciones que tuvo la disciplina histórica fue la diversificación del tipo de fuente que se empleaba para reconstruir el pasado. En las últimas tres décadas, se ha ampliado el concepto de *fuente*, en el sentido de que cualquier artefacto cultural —escrito, visual, material— puede considerarse fuente para obtener datos y reconstruir un pasado. La *revolución del documento*, como la llamó el medievalista Jacques Le Goff (1991, pp. 230-234), implicó la apertura del concepto de *archivo*. Este ya no es solo el lugar físico que alberga algún tipo de documentación, sino también una colección de objetos que, con las implicaciones de la cultura digital, también puede

8 Para información sobre Transkribus, véase https://www.transkribus.org/; sobre Kitab Project, veáse https://kitab-project.org/about/; sobre Emlo Project, véase http://emlo.bodleian.ox.ac.uk/; sobre D-scribes Project, véase https://d-scribes.philhist.unibas.ch/en/

ser virtual. Desde el siglo XIX, la historia lidió con el procesamiento en grandes volúmenes de archivos, pero el uso de estas otras fuentes propuso otro tipo de retos frente a las posibles complejidades semánticas. Por ejemplo, cuando se empezó a utilizar imágenes como la pintura o la fotografía, se plantearon problemas epistemológicos sobre cómo hacerlas "hablar", ya que no daban la misma seguridad que el dato escrito. Las soluciones tradicionales implicaron el desarrollo de investigaciones históricas que tomaron décadas para reunir un acervo de archivo de suficiente amplitud que permitiera inferir tendencias de larga duración y rupturas dentro de tales estructuras. También se probaron metodologías cuantitativas como la cliometría, en conjunto con técnicas cualitativas como el análisis de discurso y el análisis iconográfico.

Sin embargo, la IA parece abrir posibilidades definitivas para el procesamiento del *big data* histórico y la extracción de información de fuentes no textuales. A partir de una revisión del trabajo multimodal de las humanidades digitales desde 1966 al 2004, de acuerdo con Sula y Hill (2019), los experimentos sobre la multimodalidad procesada computacionalmente pueden agruparse en siete categorías: texto, imagen y mapas, sonido, objetos, números, multimedia (video y videojuegos) y tecnología (AI, bases de datos, *hardware*). Esta propuesta insinúa la preeminencia del trabajo sobre textos, pero también encuentra que un rango significativo, entre el 30 % y el 40 % de la muestra de trabajos, se realizó sobre estas fuentes no escritas (Sula y Hill, 2019, p. 202). Esta situación provee evidencia del interés que los métodos computacionales generan en las posibilidades de expansión de archivos no tradicionales.

El trabajo con IA sobre imágenes, una de las fuentes más atractivas, requiere también una vectorización de su contenido, de manera que los avances más significativos en este campo se realizan con modelos avanzados de *deep learning* basados en redes neuronales con arquitecturas complejas. El trabajo de Amanda Wasielewski (2023), que recoge metodologías y ejemplos del uso de la IA en la historia del arte, es un buen ejemplo de los avances de la aplicación, en especial del ML aplicado a las imágenes. Por otro lado, la cultura visual histórica, la cual es cercana a la historia de las imágenes, pero guarda grandes diferencias en cuanto a las formas como una cultura visualiza la realidad, también está siendo explorada por la historia digital (Wevers y Smits, 2020). Existen diversas soluciones de análisis de imágenes, incluidos los sistemas de visualización, la búsqueda visual, la clasificación, el reconocimiento de objetos y la restauración de imágenes. La clasificación de imágenes, por ejemplo, proporciona una vía para que los investigadores naveguen por los materiales visuales de los archivos digitales sin depender exclusivamente de la búsqueda basada en texto; así, facilitan

la identificación de documentos similares en grandes colecciones digitales y ayudan en el análisis de tendencias visuales (Chen *et al.*, 2024).

Persiste una brecha significativa entre la interpretación humana de las imágenes y lo que se puede extraer automáticamente de los datos visuales, como en el desafío de comprender la difusión en la cultura visual de algunas imágenes icónicas, cuyo significado se extiende mucho más allá de su composición visual inmediata (Van Noord, 2022). Pero también hay avances prometedores, como el reconocimiento iconográfico en objetos como monedas (Pavlek *et al.*, 2022) o el estudio de videos, que permiten inaugurar posibilidades de estudio sistemático de los actos de *performance* y la gestualidad de los cuerpos. En este sentido, Hou y Kenderdine (2024) son un buen ejemplo de qué es cultura visual: proponen conceptualizar el cuerpo como una vasija que encarna discursos socioculturales y un complejo sistema de lo físico, lo perceptual, lo social y lo ideacional, para abordar el reto de datificar el complejo conocimiento corporal, histórico y cultural inherente a las prácticas de artes marciales. En su innovador estudio sobre artes marciales, Hou y Kenderdine aplican una metodología basada en ontologías y grafos de conocimiento computacional propios de una forma alternativa de IA: la web semántica. Con estas técnicas, más basadas en relaciones analíticas y menos en conjuntos de datos numerosos, plantean representar el complejo mundo de las artes marciales. Para construir esta ontología, los autores realizaron un análisis minucioso de las prácticas de artes marciales, integrando datos de diversas fuentes: la representación sistemática de movimientos, las técnicas, las filosofías y los contextos culturales. Este enfoque ofrece una manera prometedora de abordar los desafíos de la documentación de conocimientos y prácticas intangibles, abriendo nuevas avenidas para la preservación y exploración digital del patrimonio cultural humano.

Un caso interesante es el uso de mapas antiguos y planos, que son recursos históricos valiosos porque documentan muchos procesos culturales, pero su análisis sistemático había sido limitado por la complejidad de la tarea al realizarla de forma manual. Los avances digitales y de la IA permiten reconceptualizar el espacio con herramientas computacionales de geolocalización y realidad inmersiva, la posibilidad de "revivir" la historia y tener experiencias sensoriales del pasado[9]. Además, con modelos supervisados de IA, mapas catastrales de los siglos XVIII y XIX, por ejemplo, han sido procesados exitosamente como imágenes, en principio, para obtener de ellos información georreferenciada (Petitpierre y Guhennec, 2023). Esto ha abierto posibilidades de investigación al integrar la

9 Sobre historia sensorial con AI, véanse Soldovieri y Masini (2017), Kee y Compeau (2019) y Smith (2021).

ciudad como espacio geográfico extendido, lo que involucra territorios circundantes y ciudades vecinas, promoviendo un enfoque de contexto macroscópico o comparativo. Dos ejemplos adicionales se encuentran en la producción digital del historiador digital William J. Turkel, quien hace una historia de la ingeniería electrónica (Turkel y Jones-Imhotep, 2019) y de la ingeniería civil (Bartlett y Turkel, 2021), acudiendo a sus fuentes primarias particulares, es decir, los circuitos y los planos de construcciones de puentes. En ambos casos recurre a modelos ML para leer este tipo de imágenes y obtener información semántica, que le permite complementar sus fuentes tradicionales.

Desde la perspectiva de otro tipo de fuente que se emplea con frecuencia en historia, la memoria y la oralidad, que trabaja con entrevistas, las posibilidades de aplicación de técnicas de análisis de audio con IA han generado nuevos datos que tienden a sustituir las transcripciones textuales, las cuales a veces pasan por alto el potencial de la oralidad y la sonoridad en las colecciones de historia oral digital. El habla tiene características como el tono, el ritmo y el volumen que son inaccesibles para la transcripción, pero tienen significado social. En este contexto, la "oralidad" tiene una resonancia particular, aunque infrautilizada, ya que ofrece una ventana hacia las dinámicas de la creación de significado, que son objeto de experimentación computacional en la actualidad (Smyth *et al*., 2023). Otros artefactos de investigación sonora que se han explorado buscan reconstruir los sonidos del pasado en lugares históricos, a través de eventos musicales y sonoros cartografiados en distintos mapas[10], y podrían incorporarse fuentes adicionales procesadas con ML, como las características de la música en relación con la emocionalidad transmitida (Yang, 2021).

La teoría de la historia mostró ya hace décadas que los documentos no son objetivos y que casi nunca cuentan toda la verdad. Como afirma Michael de Certeau (1993), los documentos a veces son importantes por lo que callan, no por lo que dicen. Esta sospecha tiene hoy en día apoyo en la IA, a partir de la cual es posible reconocer los sesgos de una colección de documentos para obtener información histórica relevante (Ortolja-Baird y Nyhan, 2022). De esta forma, se puede corroborar que el archivo histórico multimodal se nutre de fuentes aún más insospechadas: los silencios del archivo en sí mismo. Aquellos faltantes y silencios del archivo permiten comprender qué tipo de fuentes contienen los archivos y representan a qué tipo de personajes; por ejemplo, determinar si contienen un sesgo político, social o de género. Así, es posible responder preguntas como: ¿qué actores están subrepresentados o sobrerrepresentados? ¿Qué

10 Véase *Paisajes sonoros históricos*, http://historicalsoundscapes.com/

periodos de tiempo están sin información? ¿Qué ideologías o intereses políticos se privilegian? ¿Qué casos constituyen valores atípicos o excepciones? ¿Qué memoria se está preservando y cuál se excluye? Estar consciente de los sesgos del archivo permite a los historiadores indagar con un lente crítico para connotar sus análisis y precisar sus conclusiones. Por el contrario, procesar el archivo con métodos computacionales sin estar consciente de tales silencios puede invalidar las interpretaciones, perpetuar estos silencios y profundizar las tendencias subjetivas de los datos, al conjugarlos con potenciales sesgos derivados de los modelos ML empleados (Lassen *et al*., 2024).

Una propuesta desde el sur global

Los nuevos horizontes abiertos en la última década en algunos sectores de la investigación histórica al aplicar algoritmos de la IA son muy interesantes, pero parece un trabajo exclusivo de académicos que trabajan en Estados Unidos y Europa. Es cierto que los proyectos de historia digital que aplican IA a grandes volúmenes de datos tienen un componente intensivo en financiación para obtener acceso a la infraestructura de cómputo y a las colecciones digitales requeridas (Crymble y Afanador-Llach, 2021); sin embargo, también hay que tener en cuenta que tanto en el norte como en el sur global existen resistencias académicas a los procesos de cambio, especialmente tecnológicos, lo cual impacta en la formación de habilidades y alfabetización digital, en este caso de los historiadores. Por esta razón, adicional a lo costoso que pueden resultar estos proyectos, las competencias digitales mínimas que debería tener un historiador se comportan como una barrera de entrada para establecer un diálogo equilibrado entre las disciplinas sociales y la tecnología. Esto también explica, en parte, la falta de representatividad de los estudios de historia digital con IA en el sur global.

No obstante, en los últimos años se ha abierto un espacio para herramientas y contribuciones teóricas y prácticas que posibilitan el trabajo de investigación digital desde el sur global. Existen plataformas de desarrollo gratuitas en la nube, que permiten el acceso a recursos computacionales temporales para la experimentación con ML. Cada vez gana más relevancia el principio de la *ciencia abierta*, que parte de que el conocimiento científico debe ser abierto al público y reproducible, de tal manera que es posible explorar gratuitamente corpus documentales digitales en nuevos trabajos de investigación (Führ y Bisset Álvarez, 2021). Pero esta reutilización de datos[11] debe realizarse con precaución. Para los

11 La reutilización de datos puede describirse como un proceso iterativo que incluye actividades de exploración, recolección y resignificación (Wang *et al*., 2021).

científicos sociales del sur global, es determinante que el entrenamiento de sus modelos de IA se realice con *datasets* etiquetados que consideren sus particularidades culturales, lingüísticas e históricas, y el código fuente sea público y abierto, de modo que se puedan "descolonizar" los sesgos datificados, se visibilicen las investigaciones y se posibilite la colaboración sur-sur (Ghosh, 2024).

La datificación implica un proceso de transformación de un fenómeno complejo, histórico en este caso, para representarlo de manera tabular. Esta representación permite preservar las características de relacionamiento definidas por quien diseña la tabla, para que los datos puedan ser sujetos de análisis. Finalmente, los datos deben ser transformados en representaciones binarias para garantizar su almacenamiento y procesamiento computacional. La datificación, en tanto proceso[12], parte del principio de que los datos son una evidencia que respalda la observación de un fenómeno, pero están sujetos a reinterpretaciones posteriores en el marco de la argumentación académica. En ciencia de datos, para ser fuente de conocimiento, los datos deben ser el resultado de un proceso de revisión cuidadosa para garantizar la calidad de la información.

Esto coincide con las propuestas de la historia como disciplina, donde los datos históricos también son curados para establecer la conexión entre la evidencia y las afirmaciones, es decir, tanto en ciencia de datos como en la práctica histórica el investigador desempeña un papel activo en el proceso de abstracción de la realidad en datos. En el contexto de la historia digital, la datificación se entiende entonces como el proceso de conversión material de un documento histórico para que pueda habitar en un espacio digital y ser sujeto de una interpretación humana y computacional. De ahí que se amplifique de forma exponencial la vulnerabilidad que tiene el proceso de catalogar y usar información de cualquier tipo, ahora con el manejo de grandes volúmenes. Por ese motivo, la datificación es un proceso que debe ser conducido de manera rigurosa por los científicos sociales, en especial desde el sur global, considerando que la mayor parte de archivos convertidos en *datasets* no son necesariamente representativos de los fenómenos culturales o sociales en un momento histórico dado.

Una propuesta teórica interesante para abordar propuestas desde el sur global es el *deep data*. Los conjuntos de datos curados y las colecciones históricas tienen la particularidad de ser densas en significados, aunque no necesariamente demasiado grandes en tamaño. A esta característica de los datos propios de las humanidades y ciencias sociales Štular y Belak (2022) le han denominado *deep data*. El concepto enfatiza en que es posible que los conjuntos

12 Véase Shilova (2024).

de datos no tengan el tamaño de millones de observaciones que se manejan en *big data*, pero son muy complejos semánticamente, porque son los resultados de proyectos de investigación de largo plazo, con una estricta curaduría digital. Así, pueden existir *datasets* "profundos" por la cantidad de características o etiquetas asociadas a cada observación, los cuales también pueden ser objeto interesante de investigación digital. Esta característica reduce las exigencias de capacidad computacional para procesar *big data*.

Un ejemplo de aplicación del *deep data* en la producción historiográfica digital latinoamericana es el proyecto Arte Colonial Americano (ARCA) de Borja Gómez (2023), una base de datos que integra 25 000 pinturas coloniales americanas producidas entre 1550 y 1830 en territorios españoles, anglosajones y lusitanos de la América colonial. Se trata de un portal para consulta de la base de datos con visualizaciones interactivas[13], un libro impreso (Borja Gómez, 2021) y un *e-book* que contiene formas diferentes de narrativas históricas digitales[14]. El proyecto, desarrollado con apoyo de la Universidad de los Andes, tiene dos versiones publicadas: la primera, del 2015, se compone de 19 000 piezas artísticas e incluye técnicas de visualización de datos con tecnología Tableau y diagramas de fuerzas; la segunda versión, del 2023, enriqueció la colección a 25 000 pinturas y parte de una modelación de base de datos relacional, para posibilitar una propuesta visual más compleja.

El proyecto ARCA, cuyo trabajo se ha extendido por más de una década, se inscribe dentro de esta tendencia de crear curaduría de archivos no tradicionales explorados digitalmente, ya que se ha constituido a partir de datos dispersos en miles de lugares y se ha enriquecido con la caracterización compleja de cada pieza de arte, testimonio de un pasado visual que se ha incorporado en la colección. Estas imágenes proceden de archivos analógicos, que al ser datificadas asumen una condición ambigua, pues esto no los preserva definitivamente. Al contrario, el documento hoy en día puede ser

> volátil, inestable, frágil y muchas veces efímero. Por eso mismo, tampoco es singular, único, sino que existen un sinfín de avatares del original y, no obstante, está sujeto a perderse, ya sea por su misma fugacidad o por su propia dinámica cambiante, su variabilidad, su continua actualización. (Pons, 2017, p. 289)

13 Véase Borja Gómez (2024b).

14 Véase Borja Gómez (2024a).

Estas paradójicas características se encuentran claramente en las imágenes barrocas, de modo que para la preservación y aprovechamiento de la información se necesitó pensar su estructuración en una base de datos que reflejara su carácter simbólico y dinámico.

El trabajo con pinturas coloniales en el proyecto ARCA reveló al menos cinco obstáculos derivados de los objetos culturales en sí mismos y su contexto histórico, que pueden inducir sesgos en la composición de los datos digitales. El primero es la construcción de la pintura como fuente. Hasta el siglo XIX, las sociedades que dieron lugar a estos artefactos no tuvieron preocupaciones en cuanto a fecha, autoría o clasificación. Las repúblicas decimonónicas se "inventaron" la idea del *arte colonial* e iniciaron el proceso de construirlas como fuentes, por eso se hace urgente el dato que las autentica como originales. Este breve contexto muestra cómo este tipo de pintura contiene dos importantes problemas: la vulnerabilidad del dato y su descontextualización. Lo primero porque cerca del 70 % de las pinturas son anónimas o atribuidas, es decir, no están firmadas ni fechadas; luego, al introducirse estas dentro de los mercados del arte y en los diferentes tipos de coleccionismo se pierde la precisión de su origen, se descontextualizan. Hoy en día, la mayoría de estos objetos se encuentran fuera del lugar para el cual fueron producidos, desperdigados a lo largo y ancho de América.

El segundo obstáculo está relacionado con el anterior: la pérdida de identidad como objeto visual. Desde el siglo XIX cada país "nacionalizó" su arte, vinculando las obras a *escuelas nacionales*, de manera que las separó de su área de producción cultural colonial y las denominó *arte colonial*, cuando en realidad su función original no era artística, sino devocional, en la mayor parte de los casos. Esto lleva al tercer problema: el desarrollo visual en las diferentes regiones coloniales se llevó a cabo con profundas diferencias, tanto en el volumen de la producción como en los ejes temáticos y en las características visuales de estas representaciones. Con el despertar de los nacionalismos del siglo XIX, los nuevos países implementaron mecanismos —como la exhibición, el museo o el coleccionismo— con los que pretendían rescatar y valorar la obra colonial. Y este es el cuarto problema: la constitución del canon del arte colonial, para lo cual la naciente investigación se preocupó por el dato de la obra, en especial su autoría y fecha. De allí provienen las invenciones de lo colonial que tanto afectan la posibilidad de restituir las pinturas a su contexto original.

Estos problemas tienen una dimensión diferente en el momento en el que este acervo de 25 000 imágenes se convierte en un contenido digital. Y aquí reside el último problema, pues al convertir una pintura en un registro dentro de una base de datos pierde la relación con el conjunto de obras al que pertenece; y,

posteriormente, además de perder su contexto de producción, pierde su contexto de reproducción. Por ejemplo, las pinturas en museos, conventos o iglesias muchas veces tienen sentido dentro de una curaduría o porque forman parte de una serie.

Sin embargo, estos obstáculos se convierten en el faro desde el cual se limpian los datos y se buscan caminos para generar una mayor precisión en la información. La constitución de los metadatos, a partir del trabajo de análisis en cuarenta campos que responden a diferentes preguntas, permite subsanar de algún modo las limitaciones que imponen los obstáculos mencionados. Por ejemplo, la identificación de temas visuales exactos o los gestos de las manos (quirología) pueden permitir una ubicación del lugar de producción del artefacto visual. La base de datos sirve a múltiples otros propósitos de investigación: para poner en contexto geográfico la secuencia de producción de la cultura visual, para mostrar cómo se dieron los circuitos de producción de las pinturas, los contenidos temáticos y cómo estos aspectos manifiestan una cultura del cuerpo y de los gestos.

Las pinturas coloniales fueron datificadas a través de un análisis de cada una de las imágenes en cuarenta características, lo que arroja un *dataset* multidimensional por el conjunto de metadatos que se acerca a 875 000 observaciones etiquetadas manualmente. A diferencia de las bases de datos de imágenes tradicionales de museos e instituciones, esta se construyó bajo el modelo relacional, que permite hacer colecciones particulares de casi cualquier elemento, y cuenta con un sistema de visualizaciones en tiempo real, lo cual da la posibilidad de construir patrones, ver modelos y responder nuevas preguntas. Además, es de acceso público y gratuito, se inscribe dentro de los principios de ciencia abierta, tiene un API pública y proporciona acceso al código fuente del desarrollo.

Esta particularidad de ARCA permite pensarla como insumo de series de datos históricos y como un *dataset* curado que contiene la datificación de la cultura visual americana colonial, para entrenar modelos ML, con el fin de predecir los comportamientos o explicaciones más probables de nuevos datos. Cada una de las cuarenta dimensiones de cada pintura son posibles casos de uso para la IA. Por ejemplo, la característica que se centra en los gestos identificados dentro de la composición artística detalló cerca de doscientos gestos quirológicos con sus respectivos significados, mediante técnicas mixtas de datos semiestructurados y modelos de aprendizaje de máquina. Los metadatos de estos gestos pueden permitir la "predicción" o identificación de otros gestos, que se convierten en insumo semántico para comprender la cultura gestual colonial americana y hacer estudios que pretendan rastrear la asimilación de posibles modelos iconográficos europeos o, por el contrario, la subversión de estos modelos barrocos en la producción artística local. En un futuro próximo, esta base de datos

se puede constituir en el campo de un modelo que ayude a identificar posibles autorías a partir de aquellas imágenes que están firmadas o fechadas, teniendo en cuenta que en este momento menos del 20 % de las pinturas tienen autor plenamente establecido.

Reflexiones finales. El futuro de la historia y la inteligencia artificial

Encontrar conjunciones disciplinares entre la IA y la disciplina histórica esboza un horizonte de exploración y análisis, que permite repensar los límites tradicionales del conocimiento histórico en función de la posibilidad de análisis sistemáticos de fuentes multimodales. Al integrar modelos ML en el estudio histórico, nos enfrentamos a un inmenso territorio de patrones y relaciones inéditas que aguardan ser descubiertas en vastos conjuntos de datos, lo que nos permitiría profundizar en el entendimiento de fenómenos históricos complejos. Las metodologías emergentes asociadas al procesamiento de archivos multimodales despliegan un abanico de posibilidades en cuanto al análisis de un espectro amplio de fuentes históricas, que incluyen información textual, iconográfica, gestual y de sonidos, proponiendo un reajuste epistemológico y metodológico con miras a un análisis de discurso multimodal.

Este horizonte trae consigo la emergencia de nuevas preguntas de investigación, que a su vez exigen herramientas analíticas capaces de abordar la complejidad y diversidad de datos disponibles para el historiador contemporáneo que sea seducido por esta línea metodológica de trabajo. En esta perspectiva, el análisis de Smits y Wevers (2023) destaca los modelos de *deep learning*, por su capacidad sustancial para trascender las barreras tradicionales entre texto e imagen, lo que facilita un enfoque multimodal en las humanidades digitales y la historiografía. Este enfoque instiga una reflexión sobre posibilidades y limitaciones investigativas, fomentando una mayor inclusión de diferentes voces y perspectivas. Sin embargo, implica desafíos éticos y sesgos significativos, entre los que se encuentran la crítica necesaria hacia la curaduría de datos, es decir, la aplicación de crítica de fuentes multimodales, además de los sesgos propios de los modelos de IA empleados.

Las implicaciones éticas de adoptar IA en la historiografía no deben soslayarse, exigiendo una permanente reflexión sobre la selección y análisis de datos, así como sobre las interpretaciones y narrativas generadas. La *alfabetización en datos* emerge entonces como una destreza esencial, la cual permite a los historiadores emplear herramientas de IA en sus investigaciones, al tiempo que confrontan los desafíos éticos inmanentes. Este diálogo interdisciplinario

suscita una práctica historiográfica más colaborativa e interdisciplinaria, para proyectar un futuro en el que la multidimensionalidad de las fuentes históricas y el uso ético de la tecnología redefinan la manera de entender y representar el pasado. Los dilemas sobre la objetividad y los sesgos asociados a la ciencia de datos, como contempla Wasielewski (2023), enfatizan en la necesidad de una traducción, interpretación y representación cuidadosa de los datos en humanidades. Este requerimiento reside en validar el carácter cultural de los datos y establecer parámetros para estandarizar, etiquetar y balancear conjuntos de datos, evitando modelos computacionales sesgados que puedan conducir a conclusiones erróneas. La captura y las interpretaciones de los datos, determinadas por el entorno cultural, subrayan la importancia de una datificación consciente, que haga explícitas y públicas las colecciones de datos empleadas, al igual que las consideraciones metodológicas y materiales detrás de las decisiones interpretativas.

La responsabilidad ética en el uso de la IA en la historiografía lleva a considerar no solo cómo seleccionamos y analizamos los datos, sino también cómo interpretamos y compartimos esos análisis. Esto significa una reflexión continua sobre quién tiene el poder de narrar la historia y cómo estas narrativas pueden influir en la comprensión pública del pasado. En últimas, todos los historiadores deberían ser "historiadores digitales", pensando esta como una habilidad clave que les permite tanto manejar herramientas analíticas de IA de manera competente como abordar con propiedad los retos éticos implicados. En este sentido, los historiadores están llamados no solo a utilizar la tecnología para acceder y analizar fuentes históricas, sino también a ser actores críticos en la configuración del desarrollo tecnológico, velando por que este sirva a una práctica historiográfica ética, inclusiva y reflexiva. De cara al futuro, la multidimensionalidad de las fuentes históricas y el análisis asistido por IA pueden transformar radicalmente la disciplina histórica, al promover una mayor inclusión de voces y perspectivas diversas y fomentar una práctica historiográfica más colaborativa e interdisciplinaria. Este futuro augura una expansión en las formas de entender y representar el pasado, donde los límites entre disciplinas se diluyen en función de un objetivo común: comprender la multiplicidad de realidades humanas a través del tiempo.

La colaboración entre historiadores y científicos de datos debe orientarse a aprovechar las capacidades analíticas de la IA, así como a promover una investigación consciente de sus limitaciones y sesgos, prestando especial atención a la selección de los datos, los modelos de interpretación y la representación de la historia. La transparencia en los modelos algorítmicos y la inclusión de perspectivas plurales en los conjuntos de datos serán cruciales para mitigar los riesgos

de sesgo y garantizar una representación más equitativa y exhaustiva del pasado. En última instancia, el futuro de la historia en la era de la IA depende de nuestra capacidad de equilibrar la innovación tecnológica con un compromiso ético y crítico. Al hacerlo, podemos continuar la indispensable tarea de explorar, entender y compartir las complejas tramas de nuestro pasado, asegurando que la riqueza de la historia humana sea accesible y significativa para las generaciones futuras, pero también explícita en los sesgos y limitaciones que hasta ahora han configurado nuestra comprensión del mundo.

Referencias

Bartlett, F. M., y Turkel, W. (2021). *Automatically harvesting high-quality images of historic bridges* [conferencia]. Canadian Society of Digital Humanities Conference 2021 Edmonton, Canadá. https://doi.org/10.17613/x03a-cr15

Berry, D. M. y Fagerjord, A. (2017). *Digital humanities: Knowledge and critique in a digital age*. Polity Press.

Blevins, C. (2016). Digital history's perpetual future tense. En M. K. Gold y L. F. Klein (Eds.), *Debates in digital humanities*. University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled/section/4555da10-0561-42c1-9e34-112f0695f523#ch26

Borja Gómez, J. H. (2021). *Los ingenios del pincel. Geografía de la pintura y la cultura visual en la América colonial*. Ediciones Uniandes. https://dx.doi.org/10.30778/2020.07

Borja Gómez, J. H. (2023). *ARCA (Arte Colonial Americano)* [data set]. Zenodo. https://doi.org/10.5281/zenodo.10864546

Borja Gómez, J. (2024a, 12 de julio). Los ingenios del pincel. https://losingeniosdelpincel.uniandes.edu.co/#!/intro/2

Borja Gómez, J. (2024b, 12 de julio). Proyecto ARCA. https://arca.uniandes.edu.co/

Ceruzzi, P. E. (2012). *Breve historia de la computación*. Fondo de Cultura Económica.

Crymble, A. y Afanador-Llach, M. J. (2021). Digital history: The globally unequal promise of digital tools for history: UK and Colombia's case study. En A. Nye y J. Clark (Eds.), *Teaching history for the contemporary world* (pp. 85-98). Springer. https://doi.org/10.1007/978-981-16-0247-4_7

Chen, J., Hou, J., Tsai, R., Liao, H., Chen, S. y Chang, M. (2024). Image classification for historical documents: A study on Chinese local

gazetteers. *Digital Scholarship in the Humanities*, *39*(1), 61-73. https://doi.org/10.1093/llc/fqad065

Comisión Económica de las Naciones Unidas para Europa (Unece). (2000). Terminology on Statistical Metadata. *Conference of European Statisticians Statistical Standards and Studies*, 53. https://digitallibrary.un.org/record/442455?ln=es&v=pdf

Cuéllar, A. (2023). La inteligencia artificial al rescate del Siglo de Oro. Transcripción y modernización automática de mil trescientos impresos y manuscritos teatrales. *Hipogrifo. Revista de Literatura y Cultura del Siglo de Oro*, *11*(1), 101-115. https://doi.org/10.13035/H.2023.11.01.08

de Certeau, M. (1993). *La escritura de la historia*. Universidad Iberoamericana.

Dobson, J. E. (2021). Interpretable outputs: Criteria for machine learning in the humanities. *DHQ: Digital Humanities Quarterly*, *15*(2). https://www.digitalhumanities.org/dhq/vol/15/2/000555/000555.html

Dobson, J. E. (2022). Vector hermeneutics: On the interpretation of vector space models of text. *Digital Scholarship in the Humanities*, *37*(1), 81-93. https://doi.org/10.1093/llc/fqab079

Fields, S., Lyans Cole, C., Oei, C. y Chen, A. (2023). Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman-Iraqi personal diaries. *Digital Scholarship in the Humanities*, *38*(1), 66-86. https://doi.org/10.1093/llc/fqac047

Führ, F. y Bisset Álvarez, E. (2021). Digital humanities and open science: Initial aspects. En E. Bisset Álvarez (Ed.), *Data and Information in Online Environments (DIONE) 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (pp. 154-173). Springer. https://doi.org/10.1007/978-3-030-77417-2_12

García Serrano, A. y Menta Garuz, A. (2021). *Orientaciones y evaluación de técnicas en humanidades digitales: De la estadística al deep-learning* [ponencia]. v Congreso de la Sociedad Internacional de Humanidades Digitales Hispánicas (HDH) 2021. Santiago de Compostela, España.

García Serrano, A. y Menta Garuz, A. (2022). La inteligencia artificial en las humanidades digitales: dos experiencias con corpus digitales. *Revista de Humanidades Digitales*, *7*, 19-39. https://doi.org/10.5944/rhd.vol.7.2022.30928

Gitelman, L. (2013). *Raw data is an oxymoron*. Massachusetts Institute of Technology Press.

Graham, S., Milligan, I. y Weingart, S. (2016). *Exploring big historical data. The historian's microscope.* Imperial College Press.

Ghosh, A. (2024). Recovering knowledge commons for the global south. *Journal of the Digital Humanities Association of Southern Africa*, *5*(1). https://doi.org/10.55492/dhasa.v5i1.5011

Guldi, J. (2023). *The dangerous art of text mining*. Cambridge University Press. https://doi.org/10.1017/9781009263016

Hardaker, C., Deignan, A., Semino, E., Coltman-Patel, T., Dance, W., Demjén, Z., Sanderson, C. y Gatherer, D. (2023). The Victorian anti-vaccination discourse corpus (VicVaDis): Construction and exploration. *Digital Scholarship in the Humanities*, *39*(1), 162-274. https://doi.org/10.1093/llc/fqad075

Hou, Y. y Kenderdine, S. (2024). Ontology-based knowledge representation for traditional martial arts. *Digital Scholarship in the Humanities. 39*(2), 575-592. https://doi.org/10.1093/llc/fqae005

Hu, X., Ou, W., Acharya, S., Ding, S., D'Gama, R. y Yu, H. (2023). Stylometric learning for authorship verification by Topic-Debiasing. *Expert Systems with Applications*, *233*. https://doi.org/10.1016/j.eswa.2023.120745

Kee, K. y Compeau, T. J. (2019). *Seeing the past with computers: Experiments with augmented reality and computer vision for history*. University of Michigan Press.

Lassen, I., Kristensen-McLachlan, R., Almasi, M., Enevoldsen, K. y Nielbo, K. (2024). Epistemic consequences of unfair tools. *Digital Scholarship in the Humanities*. *39*(1), 198-214. https://doi.org/10.1093/llc/fqad091

Le Deuff, O. (2018). *Digital humanities. History and development.* Wiley.

Le Goff, J. (1991). *El orden de la memoria. El tiempo como imaginario*. Paidós.

Levy, P. (2007). *Cibercultura. La cultura de la sociedad digital*. Anthropos.

Manrique-Gómez, L., Montes, T., Rodríguez Herrera A. y Manrique, R. (2024). Historical ink: 19th century Latin American Spanish newspaper corpus with LLM OCR correction. En *Proceedings of the 4th International Conference on natural Language for Digital Humanities NLP4DH*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4dh-1.13

Moretti, F. (2000). Conjectures on world literature. *New Left Review*. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Montes, T., Manrique-Gómez, L. y Manrique, R. (2024). Historical Ink: Semantic shift detection for 19th century Spanish. En *Proceedings of the 5th Workshop on Computational Approaches to Historical Language*

Change. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.lchange-1.4

Milligan, I. (2022). *The transformation of historical research in the digital age*. Cambridge University Press. https://doi.org/10.1017/9781009026055

Organización para la Cooperación y el Desarrollo Económico (OECD). (2008). *OECD glossary of statistical terms*. https://www.oecd.org/en/publications/oecd-glossary-of-statistical-terms_9789264055087-en.html

Ortolja-Baird, A. y Nyhan, J. (2022). Encoding the haunting of an object catalogue: On the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive. *Digital Scholarship in the Humanities*, *37*(3), 844-867. https://doi.org/10.1093/llc/fqab065

Pavlek, B., Winters, J. y Morin, O. (2022). Standards and quantification of coin iconography: Possibilities and challenges. *Digital Scholarship in the Humanities*, *37*(1), 202-217. https://doi.org/10.1093/llc/fqab030

Petitpierre, R. y Guhennec, P. (2023). Effective annotation for the automatic vectorization of cadastral maps. *Digital Scholarship in the Humanities*, *38*(3), 1227-1237. https://doi.org/10.1093/llc/fqad006

Pons. A. (2017). Archivos y documentos en la era digital. *Historia y Comunicación Social*, *22*(2), 283-292.

Schreibman, S., Siemens, R. y Unsworth, J. (2016). *A new companion to digital humanities*. Wiley Blackwell. https://doi.org/10.1002/9781118680605

Shilova, M. (2024, 12 de julio). The concept of datafication: Definition and examples. *Data Science Central*. https://www.datasciencecentral.com/the-concept-of-datafication-definition-amp-examples

Smits, T. y Wevers, M. (2023). A multimodal turn in digital humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities*, *38*(3), 1267-1280. https://doi.org/10.1093/llc/fqad008

Smith, M. (2021). *A sensory history manifesto*. Penn State University Press.

Smyth, H., Nyhan, J. y Flinn, A. (2023). Exploring the possibilities of Thomson's fourth paradigm transformation: The case for a multimodal approach to digital oral history? *Digital Scholarship in the Humanities*, *38*(2), 720-736. https://doi.org/10.1093/llc/fqac094

Soh, L., Lorang, L., Pack, C. y Yi Liu. (2023). Applying image analysis and machine learning to historical newspaper collections. *The American Historical Review*, *128*(3), 1382-1389. https://doi.org/10.1093/ahr/rhad369

Soldovieri, F. y Masini, N. (2017). *Sensing the past: From artifact to historical site*. Springer.

Spina, S. (2022). Historical network analysis and Htr tools for a digital methodological historical approach to the Biscari Archive of Catania. *Umanistica Digitale*, *6*(14), 163-181. https://doi.org/10.6092/issn.2532-8816/15159

Spiro, L. (2012). This is why we fight: Defining values of the digital humanities. En M. K. Gold (Ed.), *Debates in the Digital Humanities*. University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/9e014167-c688-43ab-8b12-0f6746095335

Štular, B. y Belak, M. (2022). Deep data example: Zbiva, early medieval data set for the Eastern Alps: Archaeology. *Research Data Journal for the Humanities and Social Sciences*, *7*(1). https://doi.org/10.1163/24523666-bja10024

Sula, C. y Hill, H. (2019). The early history of digital humanities: An analysis of computers and the humanities 1966-2004 and literary and linguistic computing 1986-2004. *Digital Scholarship in the Humanities*, *34*(1), 190-206. https://doi.org/10.1093/llc/fqz072

Turkel, W. y Jones-Imhotep, E. (2019). Sensors and sources: How a universal model of instrumentation affects our experiences of the past. En C. Stewart y S. Palmie (Eds.), *Varieties of historical experience* (pp. 232-253). Routledge.

van Noord, N. (2022). A survey of computational methods for iconic image analysis. *Digital Scholarship in the Humanities*, *37*(4), 1316-1338. https://doi.org/10.1093/llc/fqac003

Villamor, M., Kirsch, M. y Prieto-Nañez, F. (2023). The promise of machine-learning-driven text analysis techniques for historical research: Topic modeling and word embedding. *Management & Organizational History*, *18*(1), 81-96. https://doi.org/10.1080/17449359.2023.2181184

Viola, L. (2023). *The humanities in the digital: Beyond critical digital humanities*. Springer International Publishing. https://doi.org/10.1007/978-3-031-16950-2

Wang, X., Duan, Q. y Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, *72*(9), 1161-1182. https://doi.org/10.1002/asi.24483

Wasielewski, A. (2023). *Computational formalism. Art history and machine learning*. Massachusetts Institute of Technology Press.

Weller, T. (2013). *History in the digital age*. Routledge.

Wevers, M. (2019). Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. *arXiv.* https://doi.org/10.48550/arXiv.1907.08922

Wevers, M. y Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, *35*(1), 194-207. https://doi.org/10.1093/llc/fqy085

Yang, J. (2021). A novel music emotion recognition model using neural network technology. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.760060

TRANSPARENCIA, EXPLICABILIDAD Y CONFIANZA EN LOS SISTEMAS DE APRENDIZAJE AUTOMÁTICO

Andrés Páez

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.04

Introducción

La falta de transparencia de muchos algoritmos de inteligencia artificial (IA) se considera uno de los principales obstáculos para su desarrollo ético. Una reciente revisión de los lineamientos que han sido propuestos en diferentes lugares del mundo para el desarrollo de la IA arrojó que la transparencia era el principio ético más comúnmente citado: aparece en 73 de los 84 lineamientos examinados (Jobin *et al*., 2019). La falta de transparencia mencionada en estos lineamientos en realidad se refiere a dos problemas diferentes. Por una parte, existe una preocupación por la *implementación* transparente de aquellos sistemas de IA que afectan directamente la vida y los derechos de las personas (Coddou y Smart, 2021). La transparencia algorítmica, en este sentido, incluye elementos tan diversos como informar a las personas que una decisión que las afecta está basada en una herramienta algorítmica, implementar mecanismos de rendición de cuentas, exigir garantías de que los algoritmos no tienen efectos discriminatorios, entre otros (Llamas *et al*., 2022; Global Partnership on Artificial Intelligence [GPAI], 2024). Por otra parte, la transparencia se puede referir a la posibilidad de *comprender* el funcionamiento interno del algoritmo, la forma en que procesa los datos de entrada y llega a una predicción o una clasificación. El Reglamento General de Protección de Datos (RGPD, 2016) de la Unión Europea, por ejemplo, exige que se les proporcione a los usuarios "información significativa acerca de la lógica involucrada" en los sistemas de decisión automatizada (art. 13). En este capítulo solo me ocuparé de este segundo sentido de transparencia[1].

1 Además, vale la pena aclarar que solo trataré sistemas de IA *discriminatorios*, es decir, aquellos que llevan a cabo tareas de clasificación y predicción. La transparencia en los sistemas de IA *generativa* requiere de una aproximación diferente a la ofrecida aquí.

La transparencia algorítmica puede verse opacada por dos razones diferentes. Por una parte, muchos algoritmos complejos, como las redes neuronales profundas (*deep neural networks*, DNN), procesan en paralelo un volumen enorme de datos subsimbólicos, a través de múltiples capas ocultas de nodos interconectados. Esta arquitectura hace que el funcionamiento interno del algoritmo sea epistémicamente inaccesible para cualquier ser humano, incluidos sus diseñadores y desarrolladores. Sin embargo, la opacidad no siempre es el resultado de la complejidad técnica. En muchos casos, el funcionamiento del algoritmo es escondido intencionalmente y protegido como un secreto industrial, en especial en los sistemas de decisión automatizados de uso comercial; de este modo, se genera un tipo de opacidad totalmente diferente. En la primera parte del capítulo examinaremos estas dos formas distintas de entender la opacidad algorítmica. Cada una de ellas generará retos éticos diferentes que tendremos que analizar.

Para intentar disminuir el efecto de la opacidad producto de la complejidad técnica de los algoritmos, en años recientes se han desarrollado métodos de explicabilidad e interpretabilidad que intentan ofrecer atisbos acerca de su funcionamiento. El desarrollo de estos métodos, del que se ocupa una subdisciplina de las ciencias de la computación conocida como *IA explicable* (*explainable artificial intelligence*, XAI), se enfrenta a numerosos retos y no es evidente que vayan a presentarse grandes avances en la comprensibilidad de los algoritmos en el futuro cercano. Estas limitaciones de la XAI, que examinaremos en la segunda parte del capítulo, plantean un reto importante al ideal de transparencia que se persigue en los marcos éticos de la IA.

Afirmar que un algoritmo es transparente epistémicamente es equivalente a decir que es comprensible. Desde luego, la comprensibilidad es una cuestión de grado, determinada en gran medida por el conocimiento de fondo que tengan los usuarios del algoritmo. Un sistema de IA puede ser medianamente transparente para su desarrollador y totalmente opaco para el usuario final. Para caracterizar de manera adecuada la transparencia algorítmica es necesario hacer un análisis pragmático —contextual y situado— de qué significa *comprender* un sistema de IA. Hay varios sentidos en los que se puede decir que se comprende un algoritmo. Por una parte, podemos decir que, a través de un método de XAI como el *local interpretable model-agnostic explanations* (LIME) (Ribeiro *et al.*, 2016), entendemos cuáles elementos del *input* fueron mayormente responsables de un *output* dado. Por otra parte, la comprensión también se puede referir a entender el funcionamiento global del algoritmo por medio de un algoritmo sustituto más simple que lo represente. En la tercera parte del capítulo analizaremos estos sentidos del concepto de comprensión, para aclarar la meta de los métodos de explicabilidad.

La última tarea de este capítulo será examinar la relación entre transparencia y confianza, en especial en el contexto de las decisiones automatizadas basadas en sistemas de IA. En las relaciones humanas, la confianza es un concepto ético, en la medida en que atribuimos honestidad, justicia y competencia a los demás y depositamos en ellos la responsabilidad por nuestro bienestar y el cumplimiento de nuestros objetivos teóricos y prácticos. En el caso de los sistemas de IA, se pretende que la explicabilidad nos proporcione una forma de juzgar la confiabilidad e imparcialidad de las decisiones del sistema; sin embargo, la evidencia empírica muestra que la explicabilidad no siempre promueve la confianza, y en algunos casos incluso puede disminuirla. Aunque todavía falta estudiar más a fondo el problema, no debemos tomar como un dogma la idea de que la transparencia y la confianza en la IA van de la mano.

Dos tipos de opacidad algorítmica

Es posible entender la opacidad algorítmica de dos maneras muy diferentes. Por una parte, se refiere a modelos cuya estructura, *dataset*, características (*features*), pesos y sesgos son propiedad de una compañía privada o pública, y son tratados como secretos industriales protegidos por leyes mercantiles y de derechos de autor. Estos modelos no son necesariamente complejos, pero su opacidad se deriva del hecho de que las personas afectadas por sus decisiones están impedidas de forma legal para acceder a sus datos y funcionamiento. Por otra parte, la opacidad algorítmica se refiere a algoritmos que escapan la comprensión humana, debido a su complejidad extrema, lo cual los hace epistémicamente inaccesibles. Algunos autores, como Rudin (2019), han sugerido que cuando las decisiones de un algoritmo afectan de manera significativa la vida de las personas, solo se deberían utilizar algoritmos transparentes; sin embargo, es bien sabido que hay una relación inversa entre la transparencia y la precisión de un algoritmo. Los algoritmos más precisos, como las DNN, son también los más opacos. Sacrificar la precisión de las decisiones en aras de la transparencia podría tener efectos muy negativos sobre los usuarios. En esta sección discutiré estos dos tipos de opacidad algorítmica; para facilitar la discusión, llamaré al primer tipo *opacidad jurídica* y al segundo, *opacidad epistémica*.

La opacidad jurídica

A medida que se extiende el uso de herramientas basadas en IA tanto en la empresa privada como en las agencias del Estado, los modelos que utilizan se han convertido en bienes valiosos que deben ser protegidos. La IA está siendo

empleada para tomar decisiones importantes que afectan la vida de las personas en ámbitos como la salud, la vivienda, la educación, la asignación de subsidios, el crédito y el acceso al empleo, así como en procesos penales, en los que sirve para evaluar el riesgo de fuga o reincidencia de las personas imputadas, y en general en el sistema judicial en aplicaciones como la detección de la evasión fiscal y la vigilancia predictiva. En esta sección vamos a examinar las implicaciones éticas de algunos casos en los que se han utilizado sistemas de IA jurídicamente opacos.

Quizás el caso más conocido de opacidad jurídica es el Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). Se trata de un algoritmo utilizado en el sistema judicial estadounidense para evaluar el riesgo de reincidencia de las personas en diferentes momentos del proceso penal. El algoritmo fue creado y le pertenece a la compañía privada Northpointe, rebautizada Equivant. En una entrevista, su gerente general afirmó:

> La clave de nuestro producto son los algoritmos y son de nuestra propiedad. [...] Nosotros los creamos y no los hacemos públicos porque son una pieza esencial de nuestro negocio. No se trata de conocer los algoritmos. Se trata de conocer los resultados. (Citado en Smith, 2016)

El algoritmo genera escalas de riesgo para la reincidencia general y violenta, las cuales se usan para tomar decisiones sobre la libertad condicional de las personas encarceladas. Otro algoritmo es utilizado para decidir si las personas imputadas pueden esperar su juicio en libertad o si existe el riesgo de fuga, reincidencia o interferencia en el proceso. La naturaleza secreta del algoritmo se convirtió en el centro de una disputa jurídica que terminó en la decisión de *State v. Loomis*, tomada por la Corte Suprema del estado de Wisconsin, en los Estados Unidos, en el 2016. La demanda contra la opacidad jurídica del algoritmo fue impuesta por un imputado llamado Eric Loomis, quien alegó que utilizar la herramienta de evaluación de riesgo proporcionada por COMPAS en la decisión sobre la pena que se le debía imponer violaba su derecho al debido proceso, pues se infringía su derecho a ser condenado con base en información precisa (*State v. Loomis*, 2016, p. 757). La opacidad jurídica del algoritmo le impedía cuestionar su precisión y validez científica. También alegaba que COMPAS violaba su derecho al debido proceso, porque la Corte, de forma inconstitucional, tuvo en cuenta el género al tomar la decisión, ya que el algoritmo incluye este como una de sus variables.

La Corte Suprema de Wisconsin reafirmó la decisión que había tomado el juez del caso en contra de Loomis. Los argumentos de Loomis acerca de la violación del debido proceso fueron rechazados, porque COMPAS solo usa datos

disponibles públicamente y datos proporcionados por el acusado; en consecuencia, concluyó la Corte, Loomis hubiera podido corregir la información. Además, la Corte afirmó que el uso del género como un factor en la evaluación del riesgo sirve al propósito no discriminatorio de promover la precisión (*State v. Loomis*, 2016, pp. 766-767). La sentencia incluyó una nota de cautela para los jueces que utilicen herramientas de evaluación de riesgo y prescribió cómo se deben presentar las evaluaciones a las cortes y en qué medida las pueden utilizar (*State v. Loomis*, 2016, pp. 763-765).

Como lo han señalado varios autores (Freeman, 2016; " State v. Loomis: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing", 2017), la cuestión que *State v. Loomis* nunca resolvió tenía que ver, de forma paradójica, con el riesgo involucrado en utilizar algoritmos jurídicamente opacos al tomar decisiones judiciales. COMPAS y otros algoritmos opacos de evaluación de riesgo han sido criticados por reforzar desigualdades preexistentes, violar el derecho a la no discriminación con base en raza (Thomas y Pontón-Núñez, 2022) y disfrazar "discriminación evidente basada en demografía y estatus socioeconómico" (Starr, 2014, p. 806). COMPAS también ha sido acusado de ser menos preciso al evaluar acusados afrodescendientes (Angwin *et al*., 2016), aunque esta afirmación ha sido puesta en duda (Corbett-Davies *et al*., 2016). También se ha encontrado que no se ajusta a algunas métricas de justicia algorítmica (Gursoy y Kakadiaris, 2022). En términos generales, las herramientas actuariales y algorítmicas sufren de problemas en la calidad de los datos y de incertidumbre en sus resultados (Páez, 2016), con lo cual parece factualmente irresponsable basar cualquier decisión en ellas.

Otro ejemplo bastante conocido de opacidad jurídica es el algoritmo System Risk Indication (SyRI), el cual fue desarrollado por el Gobierno neerlandés en el 2014, como una herramienta para evaluar el riesgo de evasión fiscal y abuso de subsidios. El sistema generaba perfiles individuales basados en datos personales que habían sido recolectados de diversos organismos públicos. En el 2020, la Corte de Distrito de La Haya prohibió seguir usando SyRI por violar el artículo 8.° de la Convención Europea de Derechos Humanos (ECHR), el cual protege el derecho al respeto de la vida privada y familiar (Lazcoz y Castillo, 2020). Al igual que COMPAS, SyRI era jurídicamente opaco. En el 2017, el Ministerio de Asuntos Sociales decidió que los modelos de riesgo que utilizaba debían ser secretos. La justificación era que los potenciales infractores podrían adaptar su comportamiento si el Estado permitía el acceso al algoritmo de riesgo; sin embargo, la gran mayoría de los factores en el modelo eran estáticos, es decir, imposibles de cambiar. En consecuencia, la justificación de la opacidad judicial era muy poco convincente. Varios estudios han mostrado que

el uso de SyRI genera muchos de los mismos problemas éticos acerca de raza y clase que ya habían sido detectados en el caso de COMPAS, debido a su efecto desproporcionado sobre grupos sociales desfavorecidos (Bekker, 2021; Rachovitsa y Johann, 2022).

En Colombia, el Estado usa un sistema algorítmico conocido como Sisbén IV para estimar los ingresos de las personas y determinar su elegibilidad frente a subsidios y programas sociales. El algoritmo base también es "reservado", es decir, jurídicamente opaco, según lo establece el artículo 70 del Plan Nacional de Desarrollo 2022-2026 (Ley 2294 del 2023). La justificación de la reserva es la misma ofrecida por el Gobierno neerlandés: evitar que los ciudadanos manipulen el algoritmo al reportar información falsa. En este caso, la reserva está un poco más justificada, porque gran parte de la información que sirve de base para el algoritmo es autorreportada por los potenciales beneficiarios y se han detectado múltiples incidentes de fraude[2].

No existe un remedio efectivo contra la opacidad jurídica cuando esta es innecesaria o éticamente problemática. Es poco probable que haya cambios drásticos en las leyes mercantiles y de protección de derechos de autor, y muchos sectores están interesados en diseñar e implementar este tipo de algoritmos. Desde mi perspectiva, el principal riesgo de la utilización indiscriminada de este tipo de algoritmos es que cada vez más reemplazarán la capacidad de juzgar de los humanos, pues es fácil caer en la tentación de aceptar acríticamente los puntajes y las recomendaciones que generan. Varios estudios empíricos respaldan esta conclusión. Algunas investigaciones en economía comportamental y psicología social muestran que es psicológicamente difícil, y muy poco frecuente, actuar en contra de las recomendaciones de los algoritmos (Christin *et al*., 2015; Thaler, 1999). Estos estudios revelan que los resultados generados por los algoritmos actúan como anclas para las decisiones humanas, eliminando la libre discreción en el proceso de evaluación. No hay ninguna garantía de que el acceso a los datos sobre el funcionamiento del algoritmo contrarreste esta tendencia, pero al menos habría una base para la atribución de responsabilidades en la medida en que se conociera el peso dado a cada una de las variables. Así, las personas encargadas de tomar las decisiones asumirían su responsabilidad con base en las variables potencialmente sesgadas o discriminatorias del algoritmo.

Por último, algunos autores han señalado la necesidad de que existan auditorías independientes de los datos de entrenamiento y de la estructura de los algoritmos jurídicamente opacos. Quizás la propuesta más elaborada en este

2 Agradezco a Juan David Gutiérrez por señalarme la existencia de este caso local de opacidad jurídica.

sentido es la de Langenkamp *et al.* (2020), que solo puedo esbozar aquí de manera breve[3]. Los autores introducen la idea de *reportes de transparencia algorítmica*, los cuales cubren cuatro categorías: (1) *intención*: ¿cuál es el propósito del modelo?; (2) *dataset*: información sobre demografía, características y *dataset* de prueba; (3) *métricas*: medidas del desempeño del modelo, umbrales de prueba y definiciones de "justicia"; y (4) *aplicaciones*: cómo va a usarse el modelo en la toma de decisiones. Estos reportes, que tendrían carácter confidencial, serían la base fáctica para cualquier reclamo acerca de la ausencia de medidas precautelares por parte de las compañías para evitar la discriminación (Páez, 2021a).

La opacidad epistémica

El segundo tipo de opacidad algorítmica no es el resultado de las acciones o decisiones de ningún individuo; más bien, es la consecuencia de la forma en que funcionan los algoritmos más complejos de IA. No todos los modelos de IA son epistémicamente opacos; algunos usan arquitecturas simples, como árboles de decisión o funciones lineales sencillas, que no requieren de conocimiento técnico para ser entendidas. Pero el funcionamiento de los algoritmos más sofisticados, con las capacidades predictivas más poderosas, escapan a la comprensión humana.

Consideremos el caso de las DNN, que son el tipo más común de algoritmo epistémicamente opaco. Las DNN están diseñadas para identificar correlaciones y patrones en los datos, muchos de los cuales no son simbólicos, lo cual los hace incomprensibles para los humanos. La red usa esos patrones y correlaciones con el fin de hacer las predicciones y las clasificaciones para las cuales ha sido entrenada. Dentro de la red, los *inputs* pasan a través de múltiples capas ocultas de nodos o "neuronas" interconectados, cada uno de los cuales transforma los datos de diferentes maneras antes de pasarlos a la siguiente capa de nodos. Estas transformaciones a menudo incluyen operaciones no lineales, las cuales, al combinarse con las interacciones entre las neuronas a lo largo de las diferentes capas, le permiten a la red neuronal modelar límites de decisión complejos y multidimensionales. Incluso si pudiéramos ver todos los pesos y sesgos en la red[4], es decir, los parámetros que el modelo aprendió durante el

3 Para otras contribuciones importantes en esta misma dirección, véanse Gebru *et al.* (2018) y Mitchell *et al.* (2019).

4 Los pesos determinan la fuerza de la conexión entre neuronas. Los sesgos son constantes asociadas a cada neurona, que sirven como una forma de umbral, lo que permite que las

entrenamiento, no sería claro cómo interpretarlos a la luz de las características del *input* original. Así, la opacidad de las DNN no surge de las funciones no lineales, sino de la forma en que se usan en la red. Cada *input* de la red pasa por una serie de transformaciones complejas e interconectadas que hace imposible entender cómo se relaciona con el *output*. La naturaleza no lineal de las funciones de activación aumenta esa complejidad. En un sistema lineal, el efecto de cada *input* sobre el *output* puede considerarse independientemente de los demás, pero en un sistema no lineal el efecto de cambiar un *input* depende del valor de todos los demás *inputs*. Esto hace que sea epistémicamente imposible entender cómo cada *input* influencia el *output*.

El segundo problema es que no hay manera de verificar qué parámetros están siendo usados en las capas ocultas de la DNN y, por lo tanto, cuál es el modelo que resultó del entrenamiento. Los modelos profundos a menudo tienen un número muy grande de óptimos con un grado de precisión predictiva semejante. A este estado de cosas se le conoce como el *problema de la identificabilidad de modelos*:

> Un modelo es identificable si un conjunto de datos de entrenamiento lo suficientemente grande puede descartar todas las configuraciones posibles de los parámetros del modelo excepto una. Los modelos con variables latentes a menudo no son identificables porque podemos obtener modelos equivalentes intercambiando variables latentes entre sí. (Goodfellow *et al*., 2016, p. 284)

Por lo tanto, es imposible verificar cuál de los muchos modelos equivalentes es el que generó un *output* en un caso particular. Sin identificar el modelo utilizado, es imposible "explicar" las predicciones del modelo. Por supuesto, existe una descripción verdadera del modelo, pero es inaccesible al conocimiento humano.

La opacidad epistémica es una característica problemática de los algoritmos por varias razones. Desde el punto de vista ético, plantea muchas de las mismas preguntas que la opacidad jurídica con respecto a la discriminación oculta y la violación de los derechos humanos. Desde el punto de vista técnico, es un obstáculo para los desarrolladores que quieran mejorar el desempeño del modelo, y detectar y resolver sesgos y otros riesgos semejantes. Desde el punto de vista regulatorio, el GDPR y otros ordenamientos más recientes en otras jurisdicciones requieren que la lógica de las decisiones automatizadas que afecten la vida de las personas de manera significativa sea conocida por los usuarios. Sigue siendo

---

neuronas se activen incluso cuando la suma ponderada de sus *inputs* no es suficiente para hacerlo por sí sola.

una pregunta abierta si los desarrolladores van a poder cumplir este mandato legal. La opacidad epistémica también se considera a menudo un obstáculo para la generación de confianza en los modelos por parte de los usuarios, como veremos en detalle más adelante. Por estas y muchas otras razones, se han desarrollado diferentes métodos para eliminar, o al menos disminuir, la opacidad epistémica. En la siguiente sección estudiaremos la efectividad de estos métodos.

La XAI y los esfuerzos para restaurar la transparencia

La XAI es un programa de investigación en las ciencias de la computación que busca desarrollar métodos que provean algún grado de comprensión del funcionamiento de los modelos de aprendizaje automático. Las aproximaciones más comunes a la XAI son: (1) intentar explicar una predicción particular de un modelo, al encontrar los elementos del *input* responsables de ese *output*; o (2) proporcionar una explicación global de cómo funciona el modelo y cuáles son sus capacidades a través de un modelo más simple. La primera aproximación usa métodos locales de interpretación *post hoc*, es decir, posteriores al entrenamiento del modelo. Estos incluyen sondeos contrafácticos (Wachter *et al*., 2018; Mothilal *et al*., 2020) y diferentes tipos de métodos de perturbación del *input*, como LIME (Ribeiro *et al*., 2016), *gradient-weighted class activation mapping* (Grad-CAM) (Selvaraju *et al*., 2017; Ancona *et al*., 2019), *shapely additive explanations* (SHAP) (Lundberg y Lee, 2017), *testing with concept activation vectors* (TCAV) (Kim *et al*., 2018), entre otros[5]. La segunda aproximación está basada en el uso de modelos *proxy*, interpretativos o sustitutos. Las clases de modelos sustitutos más usados son las aproximaciones lineales o de gradiente, las reglas de decisión y los árboles de decisión (Frosst y Hinton, 2017; Wu *et al*., 2018). Ninguna de las dos aproximaciones logra eliminar por completo la opacidad epistémica.

Durante mucho tiempo, los métodos locales de interpretación *post hoc* se consideraron —al menos dentro de la comunidad de desarrolladores— el camino más prometedor para abrir la caja negra de la IA. Sin embargo, más recientemente, han sido objeto de muchas críticas debido a sus limitaciones y debilidades intrínsecas (Páez, 2024) y a la inescrutabilidad de las "explicaciones" que producen para no expertos y usuarios finales (Ehsan y Riedl, 2020). Quizás el problema más grave de los métodos locales es su bajo desempeño en diversas métricas de robustez. Lo ideal es que una alteración mínima del *input*

5 Para un estudio comprehensivo, véase Ivanovs *et al.* (2021).

no resulte en una explicación muy diferente del mismo *output*; no obstante, una transformación simple del *input*, o repetir el proceso de muestreo, puede generar explicaciones muy diferentes. Kindermans *et al*. (2019) revelan que añadir un cambio (*shift*) constante a los datos de entrada, lo cual es un paso simple y común de preprocesamiento que no afecta el desempeño del modelo, hace que muchos métodos locales de interpretación arrojen resultados equivocados. Slack *et al*. (2020) descubrieron la vulnerabilidad de LIME y SHAP a los ataques adversariales; y Ghorbani *et al*. (2019) muestran cómo generar perturbaciones adversariales que producen *inputs* perceptualmente indistinguibles, a los que se les asigna la misma etiqueta predictiva y, sin embargo, arrojan interpretaciones muy diferentes a través de métodos locales *post hoc*.

Otra limitación de los métodos locales de interpretación *post hoc,* como los mapas de calor o de prominencia (*saliency maps*), es que carecen de precisión. Por ejemplo, Rajpurkar *et al*. (2017) propusieron un mapa de calor para explicar las predicciones de una red neuronal convolucional de uso médico. El método resalta en rojo las áreas de una placa de rayos X que son más relevantes en el diagnóstico positivo de neumonía, y en azul las menos relevantes; no obstante, algunos autores han cuestionado su utilidad. Ghassemi *et al*. (2021), por ejemplo, arguyen que incluso las partes más calientes del mapa contienen información útil e inútil, desde la perspectiva de un agente humano, y que simplemente localizar la región más caliente de la placa no revela con exactitud qué elemento en esa región fue el que el modelo consideró importante:

> Un médico clínico no puede saber si el modelo estableció apropiadamente que la presencia de una opacidad en un conducto de aire fue importante en la decisión, si la forma del borde del corazón o de la arteria pulmonar izquierda fueron el factor decisivo, o si el modelo se basó una característica inhumana, tal como el valor o la textura de un píxel que pudo estar más relacionado con el proceso de toma de la imagen que con la enfermedad subyacente. (p. e746)

Más aún, la información proporcionada en el área caliente debe ser interpretada, lo que abre la puerta a las creencias previas del médico y al riesgo de que se cuele el sesgo de confirmación. La explicación también carece de cualquier tipo de justificación de por qué esa área en particular era más relevante que otras, porque no existe conocimiento causal que sustente la explicación. Por último, el sesgo de automatización (Lyell y Coiera, 2017) puede llevar a sobrestimar el desempeño del sistema y a un abandono de una actitud crítica frente a los resultados.

Los métodos contrafácticos no son inmunes al problema de la robustez. Al igual que los métodos de perturbación del *input*, los métodos contrafácticos pueden ser manipulados y converger hacia explicaciones drásticamente diferentes al introducir pequeñas perturbaciones (Slack *et al*., 2021). Los métodos contrafácticos también dependen críticamente de métricas de cercanía, pero no hay una forma fundamentada para decidir cuál métrica usar en un caso particular. Y tal como los métodos de prominencia, la falta de una base causal adecuada puede generar explicaciones subóptimas e incluso por completo equivocadas (Chou *et al*., 2021).

Esta es apenas una pequeña muestra de los problemas a los que se enfrentan los métodos locales de interpretación. Su fragilidad e imprecisión son lo suficientemente graves como para recomendar combinarlos con los métodos globales de explicación; sin embargo, estos tampoco son la panacea. Por ejemplo, para crear un modelo lineal que se asemeje funcionalmente a un modelo opaco, se necesita conocimiento experto para seleccionar las características que deben incluirse. Solo aquellas características que excedan un cierto umbral de correlación con las predicciones deseadas deben ser usadas, pero existe el riesgo de que muchas no muestren correlación alguna cuando son examinadas de forma individual o de que su contribución solo se pueda apreciar en combinación con otras características. La ventaja de los modelos lineales es que son usados con frecuencia en las ciencias sociales y naturales, incluida la medicina, lo cual los hace una herramienta conocida y aceptada por la mayoría de sus usuarios. No obstante, no siempre es posible encontrar modelos lineales sustitutos, en especial cuando el modelo está basado en datos subsimbólicos, como en los modelos de visión por computador.

Los árboles de decisión, por otra parte, se usan como sustitutos en aquellos casos en los que la relación entre las características y las predicciones son lineales o cuando las características interactúan entre sí. También pueden expresarse como reglas de decisión; sin embargo, su naturaleza escalonada no los hace muy eficientes. También son muy sensibles a cualquier cambio en los datos de entrenamiento o a cualquier cambio en las características escogidas: un cambio en una bifurcación al comienzo de un árbol lo afecta por completo.

Por último, muchos métodos globales de explicación que intentan preservar la precisión del modelo opaco original terminan generando "cajas grises". Por ejemplo, Xu *et al*. (2018) comprimieron una DNN en una red neuronal superficial, pero esta última sigue siendo completamente opaca para un usuario no experto. Y cuando los modelos sustitutos son fáciles de entender —por ejemplo, los árboles de decisión de Bastani y Bastani (2019) para valorar el riesgo de diabetes—, sufren de sobreajuste y pérdida de precisión en comparación con

el modelo original. Por supuesto, se podría argüir que el propósito principal del modelo sustituto no es alcanzar un nivel de precisión similar al del modelo original —pues en ese caso el modelo original sería innecesario—, sino ayudar a los usuarios finales a obtener algún grado de comprensión de su funcionamiento. Una explicación funcional sencilla permitiría a los usuarios entender las capacidades y limitaciones del modelo original, para que puedan ajustar de ese modo sus expectativas. Por ejemplo, una explicación sencilla de cómo funciona un modelo de IA generativa como ChatGPT ayuda a sus usuarios a entender por qué no es confiable usarla cuando se trata de información fáctica. Los árboles de decisión simples, las listas de decisión, los métodos basados en ejemplos e incluso las explicaciones dialógicas pueden ser más útiles para hacer comprensible y transparente, en algún grado, el funcionamiento de un sistema de IA. Pero ¿qué significa con exactitud que un método de XAI haga que un modelo sea *comprensible*? La siguiente sección estará dedicada a responder esta pregunta.

Explicabilidad y comprensión

Para explorar el concepto de *comprensión* debemos recurrir a la literatura filosófica al respecto. A primera vista, comprender o entender (usaré los dos términos indistintamente) una decisión específica de un sistema de IA, a través de un método local de interpretación *post hoc*, y un modelo como un todo por medio de un modelo sustituto son dos estados mentales diferentes que requieren de un análisis independiente. El primero parece corresponder a los que la literatura filosófica llama *entender por qué*, mientras que el segundo parece referirse a una *comprensión objetual*, es decir, a entender el objeto como un todo. Ambos tipos de comprensión han sido ampliamente discutidos en la epistemología y la filosofía de la ciencia.

La caracterización de la comprensión objetual en la literatura epistemológica coincide con el propósito de las explicaciones globales a través de modelos sustitutos. Zagzebski (2009), por ejemplo, afirma que la comprensión "implica lograr ver relaciones de unas partes con otras partes y quizás la relación de las partes con el todo" (p. 144). El tipo de relaciones que ella tiene en mente pueden ser espaciales, temporales o causales. Para Grimm (2011), por su parte, la comprensión global de un objeto complejo como el sistema de metro de Nueva York es un caso de "saber cómo":

> Si "saber cómo" implica una aprehensión de cómo funciona una cosa, entonces parece seguirse de ello que el objeto del "saber cómo" debe estar

> constituido por una estructura que pueda ser manejada, esto es, que pueda ser manipulada para determinar cómo los diversos elementos de la cosa se relacionan entre sí y dependen los unos de los otros. (p. 86)

Ambas descripciones asumen que la comprensión objetual requiere identificar las diversas partes de un objeto, describir sus interdependencias funcionales y usar esa información para hacer inferencias útiles. Esto es justo lo que ofrecen los modelos sustitutos, ya sea a través de ecuaciones lineales o reglas de asociación, o directamente en un árbol de decisión. Estos modelos pretenden proporcionar una versión simplificada del modelo original, al mostrar sus características (*features*) principales, las relaciones funcionales entre ellas y las decisiones del sistema. La meta es encontrar un *proxy* que restaure algún grado de transparencia, en el sentido descrito por Paul Humphreys (2004):

> En gran parte de los modelos estáticos, nuestra comprensión está basada en la habilidad de descomponer el proceso entre los *inputs* y los *outputs* del modelo en pasos modulares, cada uno de los cuales es metodológicamente aceptable tanto individualmente como en combinación con los demás. (p. 148)

El nivel de descomposición en el caso de los modelos *proxy* o sustitutos está determinado por consideraciones pragmáticas. Una vez el usuario ha entendido las relaciones funcionales que le interesan para sus metas prácticas o epistémicas, es posible afirmar que el modelo se ha vuelto transparente para él. La transparencia es un concepto asociado con el éxito práctico o epistémico, y depende de lograr ver la estructura funcional del modelo opaco a través del modelo sustituto.

Por otra parte, la descripción filosófica de *entender por qué* encaja muy bien con los métodos locales de interpretación *post hoc*. *Entender* por qué pasó *p* no es equivalente a *saber* por qué *p*. Saber que un sistema de reconocimiento de imágenes clasificó de forma correcta una imagen como un perro, porque le fue mostrada una foto de un perro, claramente no es suficiente para entender la decisión del sistema. La persona debe poder responder una amplia variedad de preguntas contrafácticas del tipo: ¿qué hubiera pasado si las cosas fueran de otro modo? (Woodward, 2003). ¿Qué hubiese pasado si las orejas del perro no fueran visibles? ¿O si la luz hubiera sido más tenue? ¿O si le hubiéramos mostrado una imagen espejo de la imagen original? Los métodos de interpretación local deben permitirles a los usuarios visualizar variaciones del *input* para lograr resolver estas preguntas. Solo su capacidad de responderlas puede demostrar que han entendido por qué el sistema hizo la clasificación correcta.

De hecho, muchos autores han argüido que no solo el entender por qué, sino la comprensión en general requiere de la habilidad de visualizar diferentes configuraciones de las partes de un objeto e inferir sus estados resultantes, es decir, la comprensión en general requiere pensar de forma contrafáctica (De Regt y Dieks, 2005; Wilkenfeld, 2013). Como afirma Knuuttila (2011), "la comprensión de lo posible es la manera de entender por qué emergió lo real y cómo funciona" (p. 269). A pesar de las apariencias, considero que los métodos locales de interpretabilidad *post hoc* no proveen las herramientas para pensar contrafácticamente. Unas pocas manipulaciones del *input* solo pueden dar a los usuarios una idea básica acerca de algunas correlaciones con las decisiones del modelo, pero no es posible generalizarlas fácilmente a casos similares. Lo que impide que estos métodos sean contrafácticos es la falta de información causal acerca de cómo funciona el modelo como un todo, esto es, la falta de comprensión objetual. Los modelos sustitutos proporcionan las reglas generales que han sido extraídas de los datos o directamente del modelo, lo que proporciona el soporte funcional necesario para pensar de forma contrafáctica acerca de cualquier predicción. Por ejemplo, en un árbol de decisión es posible seguir una rama u otra, y cada una de ellas será un caso contrafáctico cuyo resultado estará determinado por la estructura funcional estática representada en el árbol. Comprender por qué siempre requiere de algún grado de comprensión objetual (Páez, 2019).

Karimi *et al*. (2020) ofrecen un argumento similar, pero desde una perspectiva más práctica. Ellos se enfocan en el problema del recurso algorítmico: cuando una persona ha sido afectada desfavorablemente por una decisión automatizada (por ejemplo, le fue negado un crédito bancario), los métodos de XAI deberían poder sugerirle acciones posibles para mejorar o cambiar la decisión del sistema. Los autores se enfocan en uno de estos métodos: las explicaciones contrafácticas más cercanas desarrolladas por Wachter *et al*. (2018). Los autores muestran que tales contrafácticos "no resultan en un conjunto de acciones óptimas o factibles que podrían cambiar favorablemente la predicción de *h* si fueran implementadas. Este defecto se debe principalmente a no considerar las relaciones causales que gobiernan el mundo" (Karimi *et al*., 2020, p. 359). La información causal faltante forma parte del conocimiento teórico que compone la comprensión objetual del sistema. Si bien es cierto que un modelo sustituto no proporciona por sí solo la información causal faltante, al menos proporciona un conjunto robusto de correlaciones funcionales simbólicas, las cuales pueden ser investigadas y validadas de forma empírica. Estas correlaciones pueden considerarse un paso inicial hacia la obtención del conocimiento causal requerido para diseñar un sistema de decisión verdaderamente operable (Buijsman, 2023).

En suma, los métodos locales de interpretabilidad *post hoc* por sí solos no proporcionan una comprensión adecuada de un sistema de IA. Al revelar las causas de predicciones específicas, estos métodos contribuyen a establecer las interconexiones entre características y decisiones, pero solo la habilidad de un agente para razonar contrafácticamente sobre el modelo, solo su capacidad de usarlo y manipularlo puede considerarse evidencia de que lo ha comprendido y, por ende, de que se ha vuelto transparente en alguna medida para ese usuario. Los modelos sustitutos son quizás la mejor herramienta epistémica para que una amplia variedad de partes interesadas comprenda el funcionamiento de los sistemas epistémicamente opacos. La pregunta final que quiero discutir es si la comprensión de un modelo obtenido a través de un método de XAI tiene como resultado un mayor nivel de confianza en sus predicciones.

Transparencia, explicabilidad y confianza en la inteligencia artificial

Existe la creencia generalizada de que la transparencia y la explicabilidad son condiciones necesarias para que los usuarios confíen en los sistemas de IA. La *Recommendation on the ethics of artificial intelligence* (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [Unesco], 2021), por ejemplo, le dedica un capítulo entero a la transparencia y la explicabilidad; afirma que estas "tienen como objetivo proveer de información apropiada a sus destinatarios respectivos para permitirles entender y para promover la confianza" (III, §39). En esa misma línea, el marco ético AI4People, propuesto por Floridi *et al.* (2018), afirma que "es especialmente importante que la IA sea explicable, pues la explicabilidad es una herramienta crítica para generar confianza y comprensión hacia la tecnología" (p. 701). Los ejemplos de afirmaciones semejantes son numerosos[6]. Pero a pesar de la popularidad de la idea de que la transparencia promueve la confianza, tanto la evidencia disponible como la naturaleza de la confianza complican esta imagen tan sencilla.

Antes de examinar la conexión entre explicabilidad y confianza, es importante analizar la naturaleza de la segunda. En las relaciones humano-humano, la honestidad, la competencia y los valores compartidos son esenciales para establecer confianza *cognitiva* y *emocional* (Gambetta, 1991)[7]. La confianza

6 Véase Kästner *et al.* (2021, p. 169), para una revisión de las opiniones favorables acerca de la conexión entre explicabilidad y confianza en la IA.

7 En la literatura sobre la interacción entre humanos y robots, la confianza cognitiva y la emocional son llamadas *objetiva* y *subjetiva*, respectivamente (Witkowski y Pitt, 2000; Witkowski *et al.*, 2001; Tong *et al.*, 2013).

cognitiva está basada en buenas razones racionales (Lewis y Weigert, 1985), en qué tanto conocemos a la persona en quien depositamos nuestra confianza y en la evidencia sobre su confiabilidad. La confianza emocional, por su parte, se fundamenta en los sentimientos positivos generados por nuestras interacciones con los demás; es altamente contextual y depende de características sociales y culturales que no son fáciles de codificar. Los dos tipos de confianza son independientes entre sí. La gente a menudo confía en personas con las que no tiene ninguna conexión emocional, si tiene evidencia clara de su competencia y de sus habilidades. En otras ocasiones, confía en personas que le generan sensaciones positivas, quizás con base en claves sociales compartidas, sin conocer sus capacidades cognitivas.

La honestidad, la competencia y los valores compartidos solo pueden atribuirse a los demás, si además les asignamos intenciones y creencias de las cuales es posible inferir estos rasgos. En el caso de la IA, la honestidad y los valores compartidos son irrelevantes en gran medida, excepto quizás en el estudio de las relaciones humano-robot, donde es importante determinar si los usuarios atribuyen a los robots estados mentales de los cuales se pueden deducir características como la honestidad y otras intenciones (Páez, 2021b). En los demás contextos, la confianza en la IA se reduce a la competencia y la confiabilidad, esto es, a la confianza cognitiva. En la literatura al respecto, la confianza en la IA se define como "el grado en el que la persona que confía cree que el sistema automatizado se comportará como se espera" (Papenmeier *et al*., 2022, p. 3). Otra definición popular se enfoca en el desempeño del sistema:

> [La confianza es] la voluntad de una de las partes de hacerse vulnerable a las acciones de la contraparte con base en la expectativa de que la segunda llevará a cabo una acción de importancia para la primera, independientemente de su habilidad para monitorearlo o controlarlo. (Mayer *et al*., 1995, p. 712)

Hay otro contexto en el que la honestidad y los valores compartidos son importantes, pero no como rasgos atribuidos a los sistemas de IA, sino como contrapeso a las decisiones impersonales del sistema. Hay varios estudios que muestran que, en ciertos contextos, como el ámbito médico, las personas tienden a preferir las decisiones tomadas por humanos, incluso cuando son menos precisas y confiables que las tomadas por sistemas automatizados (Ferrario y Loi, 2022; Longoni *et al*., 2019). En el contexto de los vehículos autónomos, la gente tiende a desconfiar de ellos incluso cuando las estadísticas indican que generan un menor número de accidentes (Hutson, 2017; Brenan, 2018). Dietvorst *et al*. (2014) llaman a este fenómeno *aversión algorítmica*.

Una de las posibles explicaciones de la aversión algorítmica es que los seres humanos juzgamos de manera diferente a las máquinas y a nuestros congéneres. En un estudio reciente, Hidalgo *et al.* (2021) compararon las reacciones de las personas a una amplia variedad de acciones llevadas a cabo por humanos y por máquinas; concluyeron que, en general, "los humanos son juzgados por sus intenciones, mientras que las máquinas son juzgadas por sus resultados" (p. 139). Un metanálisis anterior de factores que afectan la confianza en las interacciones humano-robot (HRI) también reveló que "las características de los robots, en particular los factores relacionados con su desempeño, son la influencia más grande actualmente sobre la confianza percibida en HRI" (Hancock *et al.*, 2011, p. 523). Este resultado se alinea muy bien con las definiciones de confianza que se encuentran en la literatura sobre sistemas multiagente. En el campo de la IA médica, Hatherley (2020) arguye que es un error usar categorías consideradas relevantes para la confianza interpersonal en las interacciones entre humanos y la IA médica; es posible decir que uno depende de estos sistemas, pero no parecen ser el tipo de objetos en los que uno confía. En esta misma línea, Ferrario *et al.* (2021) afirman que la confianza que los médicos tienen en los sistemas de IA no requiere monitorearlos con respecto a propiedades que solo los humanos pueden tener. El metanálisis de Hancock *et al.* (2011) también encontró que los factores relacionados con actitudes humanas hacia los robots tenían un papel menor en la construcción de confianza.

Por lo tanto, parece que el desempeño del sistema es el principal factor en la construcción de confianza hacia las máquinas y que a menudo no es suficiente, como lo muestra la evidencia en el campo de la IA médica. La pregunta es si la explicabilidad puede agregarse como un factor que complemente al desempeño como una fuente de confianza. La respuesta es que la evidencia acerca de la utilidad de la explicabilidad para este propósito no es concluyente. Algunos estudios indican que tiene un efecto positivo. Shin (2021) hizo un estudio con 350 individuos que usaban regularmente servicios automatizados de noticias; sus resultados señalan que la transparencia y la explicabilidad impactan de forma positiva en la confianza de los usuarios. En el área de la IA de apoyo en decisiones clínicas, Liu *et al.* (2022) y Wysocki *et al.* (2023) reportaron que la transparencia y la explicabilidad fueron efectivas en la construcción de confianza entre el personal médico, aunque acentuaba el sesgo de confirmación y el exceso de confianza en el modelo.

A pesar de estos resultados positivos, la evidencia que muestra la inefectividad de la explicabilidad es mucho más extensa y convincente, incluso en los mismos sectores y en contextos similares. Papenmeier *et al.* (2019) encontraron que, de hecho, las explicaciones de alta fidelidad *disminuían* la confianza de los

usuarios en varios algoritmos de clasificación altamente precisos utilizados en redes sociales. Schmidt *et al.* (2020) también reportaron que un mayor grado de transparencia en un algoritmo de clasificación de texto puede tener un impacto negativo sobre la confianza; más aún, los autores afirman que

> este efecto ocurre predominantemente en aquellos casos en los que las predicciones del sistema de aprendizaje automático son correctas, mostrando de este modo que el uso descuidado de la transparencia en herramientas de asistencia basadas en ia puede de hecho desmejorar el desempeño humano. (p. 261)

Estos son algunos de los muchos ejemplos en los que la investigación empírica no ha encontrado ningún soporte para la hipótesis de la relación positiva entre explicabilidad y confianza[8]. Kästner *et al.* (2021) creen que hay tres razones por las que las explicaciones fallan en la promoción de la confianza:

> (1) Si la confianza que una persona le tiene al sistema ya está en su grado máximo, una explicación no puede aumentarla; (2) si la explicación revela un problema en el sistema, la explicación puede disminuir en lugar de aumentar la confianza; (3) si una persona no puede comprender la explicación o no puede usarla para evaluar el sistema, es posible que la explicación no cambie su confianza en el sistema. (p. 2)

Esta revisión de literatura indica que existe una aguda controversia en torno a la efectividad de la explicabilidad en la construcción de confianza. Hay una premisa implícita en esta discusión: la creencia de que la confianza en la IA es un fin deseable. Es evidente que no confiar en un sistema de decisión automatizado que sea útil, explicable y de gran desempeño parece irracional, incluso antiético, si la falta de confianza impide que la gente reciba sus beneficios sin un riesgo significativo; sin embargo, hay voces que invitan a la cautela. Peters y Visser (2023) advierten acerca del exceso de confianza en el modelo y recomiendan una dosis saludable de desconfianza. Ghassemi *et al.* (2021) también advierten acerca del peligro de usar explicaciones superficiales o poco confiables en las aplicaciones en el sector salud, que pueden generar falsas esperanzas en el poder de la IA y llevar al sesgo de automatización; los autores llegan al punto de afirmar que la explicabilidad no debería ser un requisito de los modelos utilizados en el ámbito clínico. Lakkaraju y Bastani (2020) también advierten que no se deben usar métodos de XAI que solo optimicen la fidelidad, esto es, que solo

8 Otros ejemplos incluyen: Chen *et al.* (2019), Cheng *et al.* (2019), Kizilcec (2016) y Langer *et al.* (2018, 2021), así como las referencias incluidas en cada uno.

se preocupen por encontrar explicaciones que dupliquen de forma correcta las predicciones del modelo de caja negra, porque es posible obtener la fidelidad aun si las explicaciones usan características completamente diferentes a las utilizadas por el modelo original. Las explicaciones de alta fidelidad "pueden incluso engañar al tomador de decisiones y hacerlo confiar en una caja negra problemática" (p. 79). Incluso cuando los métodos de XAI mejoran los reportes de confianza y comprensión de un sistema de IA, hay evidencia de que esos reportes no se traducen en una mejora en el desempeño en tareas que se apoyan en el sistema (Kandul *et al*., 2023; Papenmeier *et al*., 2022).

Finalmente, existe el riesgo ético de promover métodos de explicabilidad que fomenten la confianza emocional, a través de interfaces amigables que provean racionalizaciones *post hoc* falsas, pero persuasivas, de decisiones complejas. Esto tiende a ocurrir en el campo de la robótica social, donde la transparencia algorítmica puede interferir con la integración del robot en su entorno social. Danaher (2020) ha llamado a estos sistemas autónomos "robots engañosos". ¿Deberíamos descalificarlos como herramientas útiles debido a su naturaleza engañosa? Con frecuencia, los seres humanos inventamos racionalizaciones *post hoc* falsas para justificar las cosas que hacemos (Nisbett y Wilson, 1977). ¿Por qué las aceptamos en el caso de los humanos, pero no en el de las máquinas? Zerilli *et al*. (2019) han expresado su preocupación: "las decisiones automatizadas están siendo sometidas a estándares altos poco realistas, probablemente debido a un estimado muy alto y poco realista del grado de transparencia que es posible alcanzar en el caso de los decisores humanos" (p. 661). Isaac y Bridewell (2017) defienden la idea de que el engaño de los robots es aceptable cuando lo hacen en aras de un bien superior, incluida la integración social fluida. ¿Deberíamos entonces tolerar el mismo grado de falta de sinceridad que encontramos en las relaciones entre humanos? Con frecuencia, los robots tienen que tomar decisiones con consecuencias significativas, las cuales requieren de explicaciones complejas que no pueden ser empaquetadas en formatos estandarizados. El uso de *outputs* gráficos en tiempo real para representar los estados internos del proceso de toma de decisiones que ocurre dentro del robot es una forma prometedora para alcanzar la transparencia robótica (Wortham *et al*., 2017; Edmonds *et al*., 2019). Esta aproximación no requiere apelar a la confianza emocional de las personas; por el contrario, al hacer explícita la arquitectura jerárquica del *software* del robot, es más fácil pensar en él como un ser sin estados mentales humanos. En suma, en situaciones de alto riesgo para las personas es conveniente que la confianza cognitiva prime sobre la emocional.

Comentarios finales

A menudo se da por sentado que diseñar sistemas transparentes o explicables debe ser una meta del desarrollo responsable de la IA. Una de las principales razones que se aducen para ese desiderátum es que es deseable que la gente confíe en esos sistemas. El análisis presentado en este capítulo muestra no solo que la conexión entre explicabilidad y confianza no es obvia, sino también que el grado de transparencia al que podemos aspirar en este momento, dado el estado actual de los métodos de explicabilidad, es muy limitado. Esto ha llevado a que algunos investigadores adopten una actitud escéptica acerca de la posibilidad de comprender los sistemas de IA.

Por ejemplo, Humphreys (2004) argumenta que "debemos abandonar la insistencia en la transparencia epistémica para las ciencias de la computación"; en lugar de transparencia, podemos alcanzar las virtudes de los modelos computacionales "a través de procedimientos de prueba y error, tratando las conexiones entre la plantilla computacional y sus soluciones como una caja negra" (p. 150). Otros se preguntan si el uso cada vez más extendido de la IA en las ciencias "ejemplifica un cambio de paradigma que nos aleja del propósito explicativo tradicional de la ciencia, y nos acerca hacia el reconocimiento de patrones y la predicción" (Boge y Poznic, 2021, p. 171).

Yo quisiera resistir estos llamados a abandonar el proyecto de hacer que la IA sea comprensible. Hay razones epistémicas, éticas y jurídicas —es decir, razones normativas— para continuar desarrollando la IA explicable. En consecuencia, la discusión acerca de las posibilidades de la XAI no debe estar restringida a sus limitaciones técnicas. Así mismo, es una discusión filosófica acerca de la transparencia de una tecnología utilizada para tomar decisiones que afectan en gran medida la vida de las personas y, en ese sentido, es una discusión ética de comienzo a fin.

Referencias

Ancona, M., Ceolini, E., Öztireli, C. y Gross, M. (2019). Gradient-based attribution methods. En W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen y K. R. Müller (Eds.), *Explainable AI: Interpreting, explaining, and visualizing deep learning* (pp. 169-191). Springer Nature.

Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016, 23 de mayo). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bastani, O., Kim, C. y Bastani, H. (2017). Interpreting blackbox models via model extraction. *arXiv*. https://arxiv.org/abs/1705.08504

Bekker, S. (2021). Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. En O. Spijkers, W. G. Werner y R. A. Wessel (Eds.), *Netherlands yearbook of international law 2019* (pp. 289-307). Springer.

Boge, F. J. y Poznic, M. (2021). Machine learning and the future of scientific explanation. *Journal for General Philosophy of Science*, *52*(1), 171-176.

Brenan, M. (2018, 15 de mayo). Driverless cars are a hard sell to Americans. *Gallup*. https://news.gallup.com/poll/234416/driverless-cars-tough-sell-americans.aspx.

Buijsman, S. (2023). Causal scientific explanations from machine learning. *Synthese*, *202*(6), 202. https://link.springer.com/article/10.1007/s11229-023-04429-3

Chen, L., Yan, D. y Wang, F. (2019). User evaluations on sentiment-based recommendation explanations. *ACM Transactions on Interactive Intelligent Systems*, *9*(4), 1-38.

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M. y Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. En *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). Association for Computing Machinery (ACM).

Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C. y Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59-83.

Christin, A., Rosenblat, A. y Boyd, D. (2015). Courts and predictive algorithms. En *Data and Civil Rights: A New Era of Policing and Justice*. http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf

Coddou, A. y Smart, S. (2021). La transparencia y la no discriminación en el Estado de bienestar digital. *Revista Chilena de Derecho y Tecnología*, *10*(2), 301-332.

Corbett-Davies, S., Pierson, E., Feller, A. y Goel, S. (2016, 17 de octubre). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas

Danaher, R. (2020). Robot betrayal. A guide to the ethics of robot deception. *Ethics and Information Technology*, *22*, 117-128.

de Regt, H. W. y Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137-170.

Dietvorst, B. J., Simmons, J. P. y Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126.

Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y. N., Lu, H. y Zhu, S. C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, *4*(37), eaay4663.

Ehsan, U. y Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. En *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings 22* (pp. 449-466). Springer.

Ferrario, A., Loi, M. y Viganò, E. (2021). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics*, *47*(6), 437-438.

Ferrario, A. y Loi, M. (2022). How explainability contributes to trust in AI. En *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1457-1466). Association for Computing Machinery (ACM).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., *et al*. (2018). AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 689-707.

Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75-106.

Frosst, N. y Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv*. https://arxiv.org/abs/1711.09784.

Gambetta, D. (Ed.). (1988). *Trust: Making and breaking cooperative relations*. Basil Blackwell.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. y Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Ghassemi, M., Oakden-Rayner, L. y Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750.

Ghorbani, A., Abid, A. y Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 3681-3688.

Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep learning*. Massachusetts Institute of Technology Press.

Global Partnership on Artificial Intelligence (GPAI). (2024). *Algorithmic transparency in the public sector: A state-of-the-art report of algorithmic transparency instruments*. https://gpai.ai/projects/responsible-ai/algorithmic-transparency-in-the-public-sector/algorithmic-transparency-in-the-public-sector.pdf.

Grimm, S. R. (2011). Understanding. En S. Bernecker y D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 84-94). Routledge.

Gursoy, F. y Kakadiaris, I. A. (2022). Equal confusion fairness: Measuring group-based disparities in automated decision systems. En *2022 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 137-146). Institute of Electrical and Electronics Engineers (IEEE).

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J. y Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517-527.

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7), 478-481.

Hidalgo, C. A., Orghiain, D., Canals, J. A., De Almeida, F. y Martín, N. (2021). *How humans judge machines*. Massachusetts Institute of Technology Press.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Hutson, M. (2017). A matter of trust. *Science*, *358*, 1375-1377.

Isaac, A. M. C. y Bridewell, W. (2017). White lies on silver tongues: Why robots need to deceive (and how). En P. Lin, R. Jenkins y K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 157-172). Oxford University Press.

Ivanovs, M., Kadikis, R. y Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228-234.

Jobin, A., Ienca, M. y Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Kandul, S., Micheli, V., Beck, J., Kneer, M., Burri, T., Fleuret, F. y Christen, M. (2023). Explainable AI: A review of the empirical literature. *SSRN*. http://dx.doi.org/10.2139/ssrn.4325219

Karimi, A. H., Schölkopf, B. y Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. En *Proceedings of the 2021*

ACM Conference on Fairness, Accountability, and Transparency (pp. 353-362). Association for Computing Machinery (ACM).

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T. y Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. En *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169-175). Institute of Electrical and Electronics Engineers (IEEE).

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J. y Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). En *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668-2677). https://proceedings.mlr.press/v80/kim18d/kim18d.pdf

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D. y Kim, B. (2019). The (un)reliability of saliency methods. En W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen y K. R. Müller (Eds.), *Explainable AI: Interpreting, explaining, and visualizing deep learning* (pp. 267-280). Springer.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. En *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390-2395). Association for Computing Machinery (ACM).

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science Part A*, *42*(2), 262-271.

Lakkaraju, H. y Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. En *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79-85). Association for Computing Machinery (ACM).

Langenkamp, M., Costa, A. y Cheung, C. (2020). Hiring fairly in the age of algorithms. *arXiv*. https://arxiv.org/abs/2004.07132

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D. y Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, *29*(2), 154-169.

Langer, M., König, C. J. y Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19-30.

Lazcoz, G. y Castillo, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. *Revista Chilena de Derecho y Tecnología*, *9*(1), 207-225.

Lewis, J. D. y Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967-985.

Ley 2294 del 2023 (19 de mayo), por el cual se expide el Plan Nacional de Desarrollo 2022- 2026 "Colombia Potencia Mundial de la Vida". *Diario Oficial* 52 400.

Liu, C. F., Chen, Z. C., Kuo, S. C. y Lin, T. C. (2022). Does AI explainability affect physicians' intention to use AI? *International Journal of Medical Informatics*, *168*, 104884.

Llamas, Z. J., Mendoza, O. A. y Graff, M. (2022). Enfoques regulatorios para la inteligencia artificial (IA). *Revista Chilena de Derecho*, *49*(3), 31-62.

Longoni, C., Bonezzi, A. y Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629-650.

Lundberg, S. M. y Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874

Lyell, D. y Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423-431.

Mayer, R. C., Davis, J. H. y Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. *et al.* (2019). Model cards for model reporting. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). Association for Computing Machinery (ACM).

Mothilal, R. K., Sharma, A. y Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. En *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607-617). Association for Computing Machinery (ACM).

Nisbett, R. E. y Wilson T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco). (2021). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Páez, A. (2016). The prediction of future behavior: The empty promises of expert clinical and actuarial testimony. *Teoría Jurídica Contemporánea*, *1*, 75-101.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441-459.

Páez, A. (2021a). Negligent algorithmic discrimination. *Law and Contemporary Problems*, *84*(3), 19-33.

Páez, A. (2021b). Robot mindreading and the problem of trust. *AISB Convention 2021: Communication and Conversation* (pp. 140-143). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Páez, A. (2024). Understanding with toy surrogate models in machine learning. *Minds and Machines*, *34*(4), 45.

Papenmeier, A., Englebienne, G. y Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv*. https://arxiv.org/abs/1907.12652

Papenmeier, A., Kern, D., Englebienne, G. y Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *29*(4), 1-33.

Peters, T. M. y Visser, R. W. (2023). The importance of distrust in AI. En L. Longo (Ed.), *Explainable artificial intelligence: First world conference, XAI 2023* (pp. 301-317). Springer.

Rachovitsa, A. y Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review*, *22*(2), 1-15.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T. *et al.* (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv*. https://arxiv.org/abs/1711.05225

Reglamento General de Protección de Datos (RGPD). (2016). Reglamento (UE) 2016/679. https://www.boe.es/doue/2016/119/L00001-00088.pdf

Ribeiro, M. T., Singh, S. y Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). Association for Computing Machinery (ACM).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206-215.

Schmidt, P., Biessmann, F. y Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260-278.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. y Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. En *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626). Institute of Electrical and Electronics Engineers (IEEE).

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551.

Slack, D., Hilgard, S., Jia, E., Singh, S. y Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. En *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186). Association for Computing Machinery (ACM).

Slack, D., Hilgard, A., Lakkaraju, H. y Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, *34*, 62-75.

Smith, M. (2016, 22 de junio). In Wisconsin, a backlash against using data to foretell defendants' futures. *The New York Times*. https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803-872.

State v. Loomis (2016). Wisconsin Supreme Court Decisions, 881 N.W.2d 749, cert. denied, 137 S. Ct. 2290 (2017).

State v. Loomis: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. (2017). *Harvard Law Review*, *130*, 1530-1537.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183-206.

Thomas, C. y Pontón-Núñez, A. (2022). Automating judicial discretion: How algorithmic risk assessments in pretrial adjudications violate equal protection rights on the basis of race. *Minnesota Journal of Law & Inequality*, *40*(2), 371-407.

Tong, X., Zhang, W., Long, Y. y Huang, H. (2013). Subjectivity and objectivity of trust. En *International Workshop on Agents and Data Mining Interaction. ADMI 2012* (pp. 105-114). Springer.

Wachter, S., Mittelstadt, B. y Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841-887.

Wilkenfeld, D. (2013). Understanding as representation manipulability. *Synthese*, *190*, 997-1016.

Witkowski, M., Artikis, A. y Pitt, J. (2001). Experiments in building experiential trust in a society of objective-trust based agents. En R. Falcone, M. Singh e Y. H. Tan (Eds.), *Trust in cyber-societies* (pp. 111-132). Springer.

Witkowski, M. y Pitt, J. (2000). Objective trust-based agents: Trust and trustworthiness in a multi-agent trading society. En *Proceedings of the Fourth International Conference on MultiAgent Systems* (pp. 463-464). Institute of Electrical and Electronics Engineers (IEEE).

Woodward, J. (2003). *Making things happen. A theory of causal explanation*. Oxford University Press.

Wortham, R. H., Theodorou, A. y Bryson, J. J. (2017). Robot transparency: Improving understanding of intelligent behavior for designers and users. En Y. Gao, S. Fallah, Y. Jin y C. Lakakou (Eds.), *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Proceedings* (pp. 274-289). Springer.

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V. y Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. *arXiv*. https://arxiv.org/abs/1711.06178v1

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R. y Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839.

Xu, K., Park, D. H., Yi, C. y Sutton, C. (2018). Interpreting deep classifier by visual distillation of dark knowledge. *arXiv*. https://arxiv.org/abs/1803.04042

Zagzebski, L. (2009). *On epistemology*. Wadsworth.

Zerilli, J., Knott, A., Maclaurin, J. y Gavagahn, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*, 661-683.

PARTE II

APLICACIONES Y EFECTOS DE LA INTELIGENCIA ARTIFICIAL

POTENCIAL DE LA INTELIGENCIA ARTIFICIAL EN TELEDETECCIÓN PARA EL DESARROLLO SOSTENIBLE Y LA GESTIÓN AMBIENTAL

Haydemar Núñez, Andrés Calderón, Nicolás Díaz, Rocío Sierra, David Vásquez

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.05

La teledetección ha emergido como una valiosa herramienta para la observación y comprensión de nuestro entorno terrestre, a través de la captura de datos con sensores remotos. Debido a la complejidad de los procesos relacionados con esta tecnología, en los últimos años se han venido utilizando técnicas de inteligencia artificial (IA) para el análisis e interpretación de imágenes satelitales. La combinación de teledetección e IA mejora nuestro entendimiento de los cambios ambientales a escala global, regional y local, y apoya las acciones favorables al desarrollo sostenible y la gestión efectiva de recursos naturales. Utilizando imágenes satelitales y técnicas avanzadas de aprendizaje automático (*machine learning*), es posible facilitar la planificación urbana sostenible, el seguimiento de la deforestación y la desertificación, la detección de desastres naturales, la evaluación de la calidad del aire y del agua, entre otros usos, al proporcionar información crucial para la implementación efectiva de políticas respetuosas con el medio ambiente.

Este trabajo presenta dos aplicaciones de *software* que apuntan en esta dirección: utilizan técnicas de aprendizaje automático sobre imágenes satelitales, para apoyar la toma de decisiones que promuevan un desarrollo sostenible. La primera es un atlas que permite estimar el potencial energético a partir de la biomasa residual poscosecha[1]. En la búsqueda de soluciones energéticas verdes,

1 Esta aplicación forma parte de un proyecto de cooperación triangular entre la Universidad de los Andes, la Universidad de Chile y la Gesellschaft für Internationale Zusammenarbeit (GIZ) de Alemania, para generar recomendaciones orientadas a impulsar la reactivación

la valorización de este tipo de residuos emerge como una estrategia clave[2]. En muchos sectores agropecuarios, una cantidad significativa de biomasa residual representa un recurso hasta ahora mayormente desperdiciado. Convertir estos residuos en fuentes de energía renovable no solo proporciona una alternativa a los combustibles fósiles, sino que también fortalece la capacidad de las comunidades para enfrentar desafíos ambientales y económicos. En este contexto, el atlas de biomasa ha sido diseñado como una herramienta interactiva que permite a los usuarios acceder a un mapa y obtener estimaciones del potencial energético de diversas zonas, según el tipo de cultivo presente y el tipo de residuo.

La segunda aplicación está dirigida a apoyar la monitorización del crecimiento periférico de ciudades[3]. La expansión urbana descontrolada puede llevar a problemas como la falta de vivienda adecuada, la contaminación ambiental y la segregación social. Así, se busca apoyar un desarrollo urbano planificado y sostenible, que incluya la construcción ordenada de infraestructuras y servicios básicos, la disponibilidad de viviendas seguras y asequibles, la protección del patrimonio cultural y natural, y la mejora de la calidad de vida en las urbes y sus alrededores. Por ello, es fundamental contar con herramientas de seguimiento que permitan a los planificadores y autoridades urbanas tomar decisiones informadas, basadas en datos actualizados, para identificar áreas de mejora y diseñar políticas que promuevan un progreso citadino más equitativo, resiliente y sostenible a largo plazo.

La siguiente sección está dedicada a explicar varios conceptos básicos sobre la teledetección y las imágenes satelitales. Luego, se presentan aplicaciones actuales de la IA en diferentes sectores relacionados. A continuación, se describe el atlas para la estimación de potencial energético a partir de la biomasa residual, para después hacer lo propio con la aplicación para la monitorización del crecimiento periférico de ciudades. Para finalizar, se dan algunas reflexiones sobre los resultados y elementos abordados en este trabajo.

---

económica de territorios vulnerables, a través de la elaboración y divulgación de mapas interactivos de potencial energético solar y de biomasa.

2 El objetivo de desarrollo sostenible 7 (ODS 7, "Energía asequible y no contaminante") está dirigido a garantizar el acceso a una energía asequible, fiable, sostenible y moderna para todos. Dentro de las metas específicas del ODS 7 se incluye el incremento en el uso de energías renovables, siendo la biomasa una fuente que puede ser utilizada para este propósito.

3 El objetivo de desarrollo sostenible 11 (ODS 11: "Ciudades y comunidades sostenibles") tiene como meta hacer que las ciudades y los asentamientos humanos sean inclusivos, seguros, resilientes y sostenibles.

Algunos conceptos sobre teledetección e imágenes satelitales

La teledetección es la práctica de obtener información sobre las superficies terrestres y acuáticas de la Tierra, mediante imágenes adquiridas desde una perspectiva aérea, utilizando radiación en una o más regiones del espectro electromagnético, reflejada o emitida por la superficie terrestre (Campbell, 2023). Es una herramienta muy útil para la gestión del territorio, ya que posibilita el análisis espacial de la cobertura terrestre para, por ejemplo, determinar cambios en la dinámica de la población, identificar patrones de actividades económicas, evaluar ecológicamente la naturaleza, gestionar recursos hídricos y hacer seguimiento a la desforestación (Chuvieco, 2016; Lillesand *et al*., 2015).

Los satélites, a través de sensores, recopilan datos en forma de imágenes que contienen información sobre la reflectancia de la luz (Campbell, 2023; Chuvieco, 2016; Lillesand *et al*., 2015). Estas imágenes, por lo general, tienen múltiples bandas que representan distintas longitudes de onda, desde las porciones ultravioleta hasta las visibles e infrarrojas del espectro electromagnético.

Las imágenes satelitales poseen diversas características clave. La *resolución espacial* determina la capacidad de distinguir objetos pequeños en una imagen (expresada en metros por píxel), mientras que la *resolución espectral* indica la habilidad para captar diferentes longitudes de onda del espectro electromagnético. La *resolución temporal* refleja la frecuencia con la que un satélite puede adquirir imágenes de una zona específica. Además, los satélites de observación terrestre ofrecen una cobertura global, ya que siguen una trayectoria que les permite capturar imágenes de cualquier lugar en la superficie terrestre bajo esta. Algunos satélites también utilizan información multisensorial, empleando múltiples sensores para capturar datos en diferentes espectros y resoluciones.

Existen diferentes misiones que han lanzado satélites al espacio para cumplir con objetivos científicos, comerciales o de seguridad. Entre ellas se pueden citar Landsat[4], programa conjunto entre la Administración Nacional de Aeronáutica y el Espacio (NASA) y el Servicio Geológico de los Estados Unidos (USGS); y Copernicus[5], programa de observación de la Tierra liderado por la Unión Europea y desarrollado en colaboración con la Agencia Espacial Europea. Este último, fuente de las imágenes para las aplicaciones desarrolladas, consta de una constelación de satélites, llamados *satélites de monitoreo ambiental* (Sentinel), y una serie de servicios operativos que utilizan los datos recopilados

4 Véase https://landsat.gsfc.nasa.gov/

5 Véase https://www.copernicus.eu/es

por estos satélites[6]. Los Sentinel están diseñados específicamente para recopilar información sobre la Tierra en una amplia gama de aspectos, como la atmósfera, los océanos y los ecosistemas terrestres.

Los Sentinel, que son lanzados con diferentes capacidades y funciones, procesan y distribuyen datos de forma gratuita, fomentando así el acceso abierto a la información para fines científicos y comerciales. Entre estos está Sentinel-1, que utiliza radar de apertura sintética para obtener imágenes de la Tierra cada seis días, independiente de las condiciones meteorológicas, con una resolución espacial de cinco metros; su dominio de medición abarca la topografía del paisaje, la humedad del suelo y la vegetación. Otro es el Sentinel-2, el cual captura imágenes multiespectrales con una resolución espacial de diez metros para las bandas espectrales visibles e infrarrojas cercanas y una resolución de veinte metros para las bandas infrarrojas de onda corta, con una capacidad de revisita de diez días. Esta multiespectralidad brinda una visión integral y detallada de la superficie terrestre, lo que permite analizar la salud de la vegetación, monitorear cambios en el uso del suelo, evaluar la calidad del agua y realizar diversas aplicaciones relacionadas con la gestión del medio ambiente y la agricultura.

Inteligencia artificial sobre imágenes satelitales

El avance vertiginoso de la IA ha potenciado radicalmente la capacidad de procesamiento y extracción de información a partir de diversos datos complejos, ofreciendo posibilidades sin precedentes en una amplia gama de aplicaciones. En particular, la integración de la IA con la teledetección está abriendo nuevas vías para respaldar el desarrollo sostenible en múltiples sectores, al permitir analizar, de manera eficiente y precisa, grandes cantidades de datos recopilados por satélites (Janga *et al*., 2023; Jeon, 2023; Li *et al*., 2018; Miller *et al*., 2024; Thapa *et al*., 2023). Esta capacidad de procesamiento sobre imágenes satelitales puede utilizarse para prever cambios ambientales y evaluar el progreso hacia metas específicas de los objetivos de desarrollo sostenible (ODS), al proporcionar a Gobiernos y organizaciones herramientas para la toma de decisiones informadas (Burke *et al*., 2021; Holloway y Mengersen, 2018; Ferreira *et al*., 2020; Persello *et al*., 2022). Potenciar esta capacidad ayuda a comprender y mitigar los impactos en el entorno natural, al promover prácticas sostenibles que contribuyan a un avance equitativo y resiliente a largo plazo.

6 Véase https://dataspace.copernicus.eu/ecosystem

Por ejemplo, la conjunción de estas dos tecnologías ha posibilitado diversas aplicaciones para el sector agrícola: el seguimiento dinámico de los cultivos, la predicción de la productividad agrícola, la segmentación y clasificación de cultivos, la detección de enfermedades y estrés en las plantaciones, entre otras (Benos *et al*., 2021; Teixeira *et al*., 2023). Además, ha facilitado una gestión óptima de la distribución de los cultivos y el aprovechamiento del suelo. En especial en la detección y monitoreo de áreas de cultivo, una tarea crítica para la gestión agrícola sostenible y la seguridad alimentaria en el mundo, las técnicas de IA han revelado un gran potencial al posibilitar la automatización de la identificación de plantaciones mediante el análisis de imágenes satelitales (Jung *et al*., 2021).

En el sector energético, la combinación de IA y teledetección puede apoyar en la planificación e implementación de proyectos relacionados con el uso de energías renovables (Lindahl *et al*., 2023; Paletta *et al*., 2023). En particular, ofrece posibilidades para desarrollar modelos para la estimación de biomasa residual, como residuos agrícolas o forestales, al igual que para identificar cambios en la distribución y cantidad de biomasa a lo largo del tiempo. Estos modelos ayudan a identificar áreas con alto potencial para la producción de biomasa y planificar estrategias de recolección y uso de manera eficiente, con el fin de generar energías alternativas (Carrijo *et al*., 2020; Liu *et al*., 2024; Senocak y Guner, 2022).

En el ámbito urbano, la integración de estas dos tecnologías promete revolucionar la forma en la que se diseñan, gestionan y optimizan las ciudades (Li *et al*., 2023). Al aprovechar técnicas de IA, junto con métodos de teledetección, se tiene entonces la capacidad de detectar y rastrear cambios en el uso del suelo, como la expansión urbana, la conversión de áreas verdes o el desarrollo de infraestructuras (Kim *et al*., 2024; Veneri *et al*., 2022). Este conocimiento puede utilizarse para la construcción de herramientas para la gestión ambiental y la evaluación del impacto de actividades humanas, promoviendo así un desarrollo urbano más sostenible y resiliente (Wu *et al*., 2024). Además, hace posible la identificación y clasificación de elementos urbanos como edificios, carreteras, parques y cuerpos de agua, lo cual contribuiría a la evaluación de infraestructuras existentes y a la planificación de nuevas construcciones.

Atlas para la estimación del potencial energético, a partir de la biomasa residual

En Colombia, los programas de desarrollo con enfoque territorial (PDET) buscan estabilizar y transformar las zonas más afectadas por la violencia y la pobreza, beneficiando a 170 municipios de 19 departamentos, que abarcan el 36 % del territorio y el 24 % de la población rural. Dada la alta proporción de población

agrícola en estos territorios, surge la oportunidad de impulsar su desarrollo mediante proyectos energéticos sustentables, aprovechando la biomasa residual poscosecha. Esto no solo beneficiaría a las comunidades locales, sino que también contribuiría a la transición energética del país, al promover el uso de residuos orgánicos[7]. En este contexto, la propuesta se centró en la construcción de un atlas de biomasa, el cual ofrezca información correspondiente a la generación de residuos poscosecha con enfoque en los territorios con PDET, incluidas las estimaciones del potencial energético generado a partir de su aprovechamiento[8]. Como se muestra en la figura 5.1, el atlas se basa en una metodología que combina técnicas de aprendizaje automático con imágenes satelitales y datos históricos de producción agrícola, para la construcción de modelos de identificación de plantaciones y la predicción de rendimiento de cultivos con base en el área sembrada. Estos modelos se integraron en una aplicación web que permite al usuario final interactuar con un mapa y obtener estimaciones del potencial energético, para zonas dentro de municipios con PDET seleccionados.

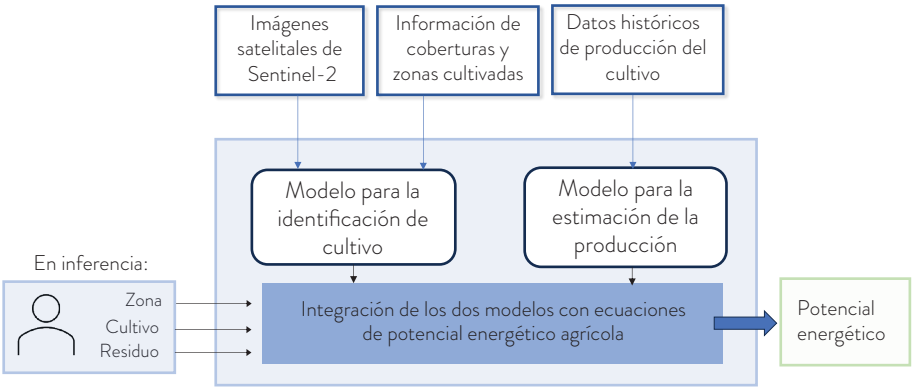


Figura 5.1. Metodología para la construcción del atlas de biomasa
Fuente: elaboración propia.

7 En el contexto del potencial del país para el aprovechamiento de residuos, en un estudio realizado por el Centro de Innovación y Desarrollo Tecnológico del Sector Eléctrico (CIDET) para identificar los distintos tipos de biomasa residual disponibles para la generación de biogás, se encontró que el país posee un potencial teórico energético de 149 436 TJ/año de biomasa residual agrícola, pecuaria, agroindustrial y urbana. Este potencial técnico corresponde al 26 % de la demanda nacional de gas natural, con respecto a datos del balance energético colombiano (Rincón Martínez *et al.*, 2019).

8 La Unidad de Planeación Minero Energética (UPME) construyó un atlas del potencial energético de la biomasa residual proveniente de ocho cultivos promisorios, tres actividades pecuarias y dieciocho municipios como fuente de residuos sólidos urbanos en plazas de

Para este estudio se seleccionaron municipios con PDET, aplicando un análisis multicriterio (MCA) con lógica difusa, basado en once indicadores socioeconómicos y físicos, como cobertura de servicios, rendimiento agrícola, factores ambientales, entre otros[9]. Este método permitió crear un índice de potencialidad, con valores destacados en Putumayo (0,38), Cesar (0,373) y Bolívar (0,363). Aunque el MCA facilita la toma de decisiones, la incorporación de múltiples capas de datos de carácter heterogéneo aumenta la complejidad y el riesgo de sobreajuste y ruido. Por ello, la selección final incluyó veintiún municipios con PDET en Bolívar y Cesar, considerando también la extensión de áreas cultivadas y protegidas.

Para la prueba de concepto se trabajó con palma de aceite, pero la metodología es aplicable a cualquier tipo de cultivo. A continuación, de manera general, se explican los pasos seguidos para la construcción de los modelos de identificación de cultivos y estimación de la producción, con base en el ciclo de aprendizaje automático (Geron, 2022)[10].

Construcción del modelo de identificación de cultivos

A través del portal de datos abiertos de Copernicus, se realizó la descarga de las imágenes de Sentinel-2 que cubren los municipios seleccionados de los departamentos de Cesar y Bolívar; para ello, se empleó información de coordenadas y rango de fechas (años 2018 y 2020). Para la identificación de áreas de cultivos de palma de aceite, se utilizó el mapa de Cobertura de la Tierra, desarrollado por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM)[11].

---

mercado, además de los generados por las podas en áreas urbanas. La información sobre áreas sembradas, ubicación de los cultivos, rendimientos y cantidades de residuos generados se obtuvo del Anuario Estadístico del Sector Agropecuario del 2006 (Unidad de Planificación Rural Agropecuaria, s. f.) y de los gremios del sector, como Cenipalma, Cenicaña, Cima, Cenicafé, Fedearroz, entre otros. La cantidad de biomasa residual se calculó a partir de los factores de generación de residuo, utilizando los valores promedio de la caracterización fisicoquímica de cada tipo de biomasa residual (UPME, 2021).

9 Los índices socioeconómicos fueron tomados de fuentes oficiales como Territata (Departamento Nacional de Planeación, s. f.), UPME (s. f.) y la Plataforma Nacional de Datos Abiertos de Colombia (Ministerio de Tecnologías de la Información y las Comunicaciones, s. f.). Los físicos fueron recolectados a través de portales como Copernicus Global Land Service (s. f.)

10 Para una descripción detallada, véase Díaz (2023).

11 El mapa contiene la clasificación de más de 130 clases diferentes de coberturas, entre las que se describen y caracterizan territorios como centros urbanos, zonas industriales, zonas

También se utilizaron polígonos suministrados por el Instituto Geográfico Agustín Codazzi (IGAC) sobre la zona de estudio. Además, fueron construidos polígonos asociados a otras coberturas para las siguientes cuatro clases: suelo desnudo, naturaleza, ciudad (zona urbana) y cuerpos de agua. Para tal efecto, se utilizó la herramienta de creación de polígonos provista por el programa QGIS[12].

Para aplicar los algoritmos de aprendizaje automático seleccionados en este trabajo, es necesario obtener una representación, a partir de las imágenes satelitales, en la forma de una matriz bidimensional, donde cada fila corresponde a un píxel de la imagen y cada columna representa un atributo asociado a dicho píxel, como los valores de las diferentes bandas espectrales. Además, fue necesario asignar a cada píxel una clase, que define su tipo de cobertura, lo cual se incluyó como una columna adicional en la tabla. Así, para transformar los valores de las bandas de los polígonos de palma de aceite y otras coberturas a este formato tabular, se desarrolló un *script* en Python utilizando la librería GDAL[13], el cual extrae y organiza la información de cada píxel en las distintas bandas espectrales, asignando también la clase correspondiente (coberturas y cultivos de palma en los diferentes polígonos). A partir de esta tabla, se derivó un conjunto de entrenamiento con el 80 % de los datos (68 732 registros) y un conjunto de pruebas (test) con el 20 % restante (17 183 registros).

Sobre el conjunto de entrenamiento se realizaron los pasos de limpieza e ingeniería de características. El primero consistió en la eliminación de registros duplicados, nulos y con valores en cero para alguna de las bandas. En el segundo se añadieron dos capas adicionales (variables) relacionadas con un modelo digital de elevación (DEM) de la zona de estudio y la pendiente (SLOPE), que representa la tasa de cambio de elevación para cada celda del modelo digital de elevación. Para esta tarea, se descargó el ALOS PALSAR DEM, recurso cartográfico disponible dentro de los productos del satélite ALOS[14].

---

mineras, territorios agrícolas, áreas con pastos para ganadería, áreas de bosques, llanuras, pantanos y cuerpos de agua. Dentro de las clases de cobertura disponibles, se tienen algunas específicas para diversos cultivos presentes en el país; en particular, existe una clase dedicada a la palma de aceite (IDEAM, s. f.).

12 Véase https://www.qgis.org/es/site/

13 Véase https://gdal.org/en/latest/

14 El Satélite Avanzado de Observación Terrestre (ALOS), también conocido como DAICHI, fue una misión satelital japonesa que operó del 2006 al 2011. Llevaba tres instrumentos, incluido el radar de apertura sintética de banda L tipo *phased array* (PALSAR), que se utilizó para obtener observaciones detalladas de la superficie de la Tierra, de día y de noche,

El modelado se hizo con XGBoost[15], el cual es un algoritmo de aprendizaje basado en una combinación de clasificadores. Se realizó una búsqueda de hiperparámetros con técnicas de selección de modelos. El modelo final (con el mejor rendimiento, con base en la técnica de validación) fue aplicado al conjunto de test y se calcularon las métricas de rendimiento, exactitud, sensibilidad o *recall*, precisión y el F1-*score*, para cada una de las clases y para el desempeño general del modelo[16] (Díaz, 2023). Por ejemplo, para la exactitud, se obtuvo un promedio por encima de 90 %. La figura 5.2 muestra las clasificaciones hechas por el modelo sobre una zona no utilizada durante el aprendizaje.

Figura 5.2. Clasificaciones realizadas por el modelo XGBoost
La segunda imagen muestra delimitadas las plantaciones de palma de aceite para la primera imagen. En la tercera se pueden observar las zonas, en verde oscuro, que el modelo identifica como plantaciones de palma de aceite.
Fuente: elaboración propia sobre imágenes generadas son el *software* QGIS[17].

---

en cualquier condición meteorológica. Los datos PALSAR podían adquirirse de múltiples modos, con diferentes polarizaciones, resoluciones, anchos de franja y ángulos fuera del nadir, lo que lo constituía como una herramienta adecuada para la generación de modelos digitales de elevación. PALSAR se utilizó para una variedad de aplicaciones, incluida la cartografía, la observación precisa de la cobertura del suelo a escala regional, el monitoreo de desastres y el estudio de recursos. Fue una herramienta valiosa para científicos e investigadores, y sus datos siguen utilizándose para diversos fines en la actualidad. Véase https://asf.alaska.edu/data-sets/sar-data-sets/alos-palsar/alos-palsar-about/

15 Véase https://xgboost.readthedocs.io/en/stable/

16 *Recall*: representa la proporción de datos de la clase que se clasificó de forma correcta. *Precisión*: indica cuántos datos clasificados en una clase realmente pertenecen a esta. *F1-score*: proporciona la media geométrica de estas dos medidas.

17 Se ajustó la imagen con la aplicación LetsEnhance, utilizando la opción Balancedx4, para obtener una resolución de 300 dpi.

Construcción del modelo de predicción de rendimiento agrícola

Para la construcción del modelo, se utilizaron datos de las evaluaciones agropecuarias (EVA) de la Unidad de Planificación Rural Agropecuaria (UPRA) correspondientes al periodo 2006-2021, que incluyen 271 801 registros y 17 columnas con información sobre área sembrada, cosechada, producción y rendimiento, por municipio, cultivo y periodo. Tras filtrar los datos para enfocarse en el cultivo de palma de aceite, se obtuvo un conjunto de 1796 registros correspondientes a 21 departamentos, con especial énfasis en los tres primeros con mayor número de registros, incluidos 23 municipios de Cesar y 18 de Bolívar. Se entrenó un modelo de regresión lineal utilizando el área sembrada como variable descriptora y la producción como variable objetivo, sin segmentar los datos por variables temporales o espaciales. El conjunto de entrenamiento comprendió el 80 % de los registros (1436 ejemplos), mientras que el 20 % restante (360 registros) se usó para las pruebas. El modelo fue evaluado con métricas como el coeficiente de determinación (R2) y la raíz del error cuadrático medio (RMSE), obteniendo valores de 0,87 para el primero y 7293,25 para el segundo

El atlas de biomasa

La vista general de la herramienta desarrollada muestra un mapa interactivo, en el que se pueden observar los puntos clasificados mediante el modelo XGBoost (Díaz, 2023). Dentro de la grilla, el usuario puede dibujar un polígono y obtener información general, como la latitud, la longitud y la ubicación, así como información histórica de producción para el municipio al que pertenece la zona seleccionada. Además, se utiliza la resolución de píxel de la imagen para estimar el área del cultivo en la zona, obteniendo el número de píxeles de palma dentro del polígono y transformando el valor resultante a hectáreas. Con este valor se realiza el estimado de la producción con el modelo de regresión. La aplicación permite al usuario seleccionar los residuos poscosecha del cultivo, que serán utilizados como entrada para la función de estimación del potencial, junto con la producción estimada. El resultado final se muestra como un valor único de energía resultante con base en ecuaciones de potencial energético para la palma de aceite, que permiten el cálculo de energía de acuerdo con el tipo de residuo y el área sembrada. Para validar la aplicación, se hicieron pruebas de usuario con las instituciones que participaron en el proyecto[18], las cuales certificaron su utilidad y usabilidad en este contexto.

18 En este proyecto de cooperación triangular participaron, como expertos y usuarios finales, la UPME, del Ministerio de Minas y Energía, el IGAC, entre otros entes.

Identificación del crecimiento periférico de ciudades

El catastro es un registro público que documenta las propiedades y características físicas de los bienes inmuebles dentro de un área geográfica específica. Es crucial mantenerlo actualizado, porque proporciona información precisa y detallada sobre la propiedad y el uso del suelo, lo que facilita la gestión adecuada de recursos, como agua y energía, la integración de áreas verdes y espacios públicos, y la provisión de servicios básicos de manera equitativa. Además, ayuda a prevenir conflictos relacionados con la tenencia de la tierra y apoya la gestión de riesgos asociados con desastres naturales, promoviendo así ciudades más resilientes y sostenibles en el largo plazo.

Los procesos de actualización catastral se componen en gran parte por trabajos manuales que se tornan costosos, debido al tiempo de revisión que toma buscar los cambios temporales en la cobertura del suelo para áreas muy extensas, como las principales ciudades. Una vía para apoyar este proceso es por medio del desarrollo de herramientas que faciliten la identificación de cambios en la cobertura de las ciudades, lo que permite a quienes toman decisiones implementar acciones de manera oportuna, con base en analítica espacial del crecimiento de zonas urbanas.

El objetivo principal de este trabajo fue desarrollar una aplicación para facilitar la detección de cambios en la cobertura urbana y el uso del suelo[19]. Este proceso se llevó a cabo mediante el análisis de imágenes de radar multitemporales proporcionadas por Sentinel-1, con técnicas de aprendizaje profundo (*deep learning*). Como se muestra en la figura 5.3, la herramienta se basa en una red neuronal convolucional (CNN), que permite segmentar imágenes de radar asociadas a tres periodos consecutivos (T1, T2 y T3), para detectar cambios en la cobertura a través del tiempo, utilizando para su implementación tecnologías de nube. A continuación, de manera general, se describen los pasos que se siguieron para la construcción del modelo de identificación de cambios de cobertura, con base en el ciclo de aprendizaje automático (Geron, 2022)[20].

19 El usuario final de esta aplicación es una empresa colombiana que se especializa en georreferenciación enfocada en la actualización cartográfica, que consiste en detectar y procesar cambios en el territorio, para crear productos de valor agregado para diversos sectores económicos en Latinoamérica de manera sostenible.

20 Para una descripción detallada, véase Vásquez (2023).

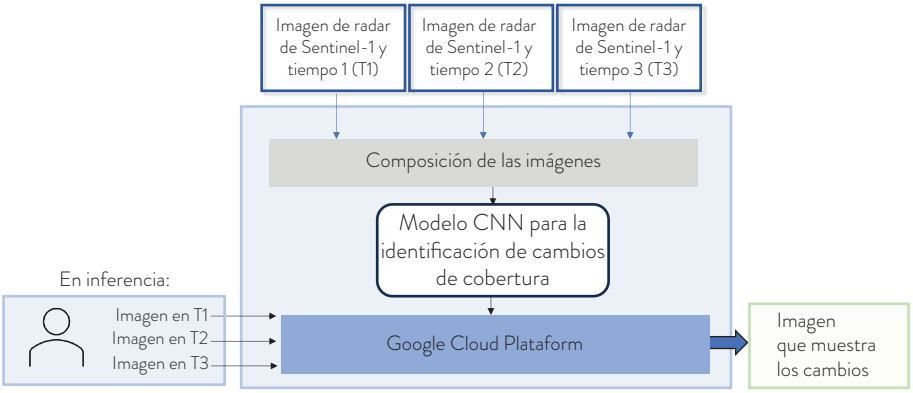


Figura 5.3. Diagrama general de la aplicación para la detección de cambios en zonas urbanas

Fuente: elaboración propia.

Construcción del modelo de identificación de cambios de cobertura

La metodología utilizada en este trabajo se basa en la propuesta realizada por Gruenhagen y Juergens (2022). Este estudio empleó imágenes de radar de Sentinel-1 para analizar cambios en la cobertura del suelo, como la demolición y la construcción de edificios. Para visualizar la dinámica espacial en la cobertura terrestre, se asignan las imágenes ordenadas cronológicamente (T1, T2 y T3) a los colores rojo, verde y azul, y se superponen en una imagen mediante la mezcla aditiva de colores. Los cambios en la ocupación del suelo entre los tres puntos temporales se revelan en la respectiva mezcla de los tres colores implicados (figura 5.4). Si se construyen o derriban edificios, estos cambios de la cubierta terrestre se muestran en el color correspondiente al periodo de tiempo, lo que permite un registro muy preciso del cambio de la cubierta terrestre. Las zonas en las que no se producen cambios en la ocupación del suelo se muestran en tonos blancos o grises.

Para construir el conjunto de aprendizaje, se descargaron imágenes de Sentinel-1 de los periodos 2019, 2020 y 2021, sobre las ciudades seleccionadas para el estudio (Barranquilla, Cali y Santiago de Chile), durante los tres primeros meses para cada T. Esto se realiza con el objetivo de tener un promedio que represente la textura real de uso del suelo y no registros momentáneos, como lo pueden ser contenedores en zonas de desembarque marítimo, parqueaderos de vehículos, entre otros. Después, se construyeron polígonos de las zonas que comprenden la categoría de cambios del uso del suelo en la imagen multitemporal,

utilizando la herramienta de edición Google Earth Engine (GEE)[21]. Por último, las imágenes se dividieron en *patches* de 512 × 512 píxeles, con sus respectivas máscaras (anotaciones) realizadas manualmente con GEE, y se obtuvo un total de 161, de los cuales el 12 % se destinó como conjunto de validación para la búsqueda de hiperparámetros. El test se diseñó en dos fases, la primera sobre un conjunto de trece *patches*, correspondientes a Santa Marta y Cartagena, los cuales tienen las anotaciones manuales correspondientes. Para la segunda fase, se realizaron veinte inferencias sobre todo Bogotá sin anotaciones manuales y solo visuales, comparadas con la construcción de la imagen de radar multitemporal de los años 2019, 2020 y 2021.

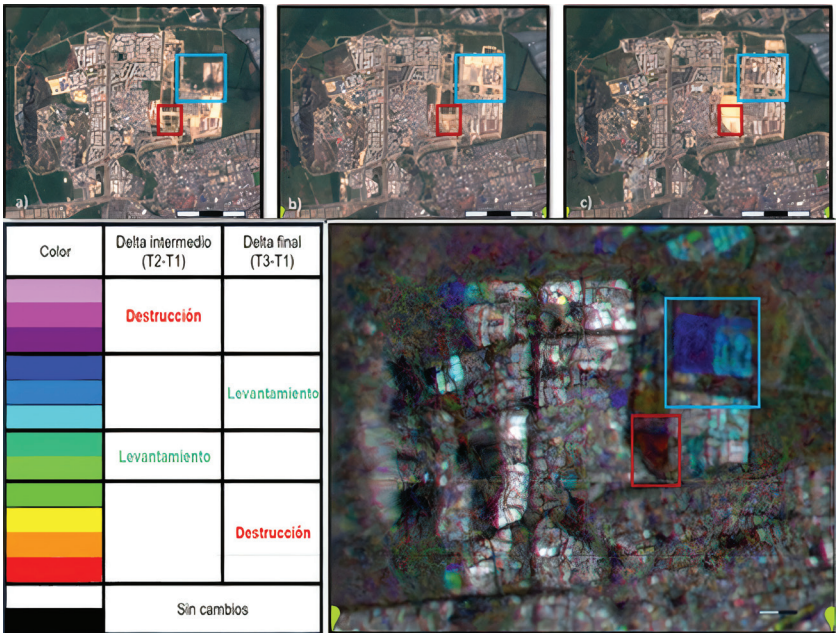


Figura 5.4. Imágenes ópticas correspondientes a Soacha en (a) T1, (b) T2 y (c) T3, así como la composición final de tres canales, los cuales representan los cambios en una escala de color RGB
Se observa que la composición es fiel a los cambios de los nuevos conjuntos residenciales construidos en la zona noroccidental (T1 corresponde a los primeros tres meses del 2019, T2 y T3 a 2020 y 2021, respectivamente). Al lado izquierdo se presenta la leyenda de la creación de la imagen multitemporal, con su interpretación.
Fuente: elaboración propia sobre imágenes generadas son el *software* QGIS[22].

21 Véase https://earthengine.google.com/

22 Se ajustó la imagen con la aplicación LetsEnhance, utilizando la opción Balancedx4, para obtener una resolución de 300 dpi.

Para el modelado se utilizó una arquitectura tipo U-net, la cual es una red neuronal convolucional diseñada para tareas de segmentación semántica de imágenes[23]. Se caracteriza por una estructura codificador-decodificador en forma de "U", de ahí su nombre. Se destaca por su capacidad de realizar una segmentación precisa de imágenes, que captura tanto información detallada como contextual a diferentes escalas. Para realzar sus capacidades, se sustituyó el codificador por la implementación de VGG16[24] de la librería Keras[25]. Luego, se llevó a cabo el entrenamiento con muestras de validación y búsqueda de hiperparámetros, los cuales permitieron ajustar la red al conjunto de datos. La plataforma utilizada para la construcción y validación del modelo fue Vertex AI[26], de Google Cloud. Con base en el conjunto de validación se obtuvo un promedio de exactitud de 92 %.

La figura 5.5 muestra nuevas inferencias del modelo sobre Bogotá, la ciudad de mayor población y crecimiento urbano en Colombia. La inferencia sobre Bogotá presenta una buena diferenciación con respecto a las zonas urbanas que mantienen colores blancos y grises, correspondientes a coberturas permanentes a lo largo del periodo de estudio. Por el contrario, la máscara resultante, que identifica cambios en la imagen de radar, deja al descubierto las zonas donde hay presencia de destrucción y construcción de nueva infraestructura, ya sea vial, residencial o industrial.

23 La segmentación semántica de imágenes es una técnica en el campo de la visión por computadora, que consiste en asignar a cada píxel de una imagen una etiqueta que identifique el objeto al que pertenece.

24 VGG16 es una red neuronal convolucional propuesta por K. Simonyan y A. Zisserman, de la Universidad de Oxford, la cual adquirió notoriedad al ganar el Desafío de Reconocimiento Visual a Gran Escala de ImageNet (ILSVRC) en el 2014. El modelo alcanzó una precisión del 92,7 %, una de las puntuaciones más altas logradas. Supone una mejora respecto a los modelos previos, al proponer núcleos de convolución más pequeños (3 × 3) en las capas de convolución de lo que se había hecho antes.

25 Véase https://keras.io/

26 Vertex AI es una plataforma que permite a los usuarios entrenar, implementar y administrar modelos de *machine learning*. Ofrece una gama de herramientas y servicios para ayudar a los usuarios a crear, entrenar y desplegar estos modelos de forma rápida y eficiente.

Figura 5.5. Inferencia de prueba sobre una imagen multitemporal del 2019-2020-2021 en Bogotá.
Datos que no fueron incluidos en el entrenamiento: (a) corresponde a la predicción del modelo; (b) corresponde a la imagen de radar multitemporal, y (c) corresponde a la superposición de las imágenes.
Fuente: elaboración propia sobre imágenes generadas con el *software* QGIS[27].

Implementación de la aplicación

El prototipo de la aplicación se realizó sobre GEE, en el módulo de desarrollo de aplicaciones, el cual está diseñado y orientado al despliegue rápido y sencillo de soluciones de prototipado. Este permite al usuario final la selección de la zona de estudio, la visualización de todos los datos que componen la imagen multitemporal de cambios y el cálculo del resultado de la segmentación de la zona donde se desea revelar el uso de suelo (Vásquez, 2023). Es importante resaltar que, con base en la opinión experta, la aplicación desarrollada es capaz de mejorar la asignación de recursos para la actualización cartográfica, al aumentar la frecuencia de monitoreo de los cambios en las zonas de interés.

Comentarios finales

Una estrategia para apoyar el desarrollo de zonas vulnerables es la implementación de proyectos sustentables con base en el aprovechamiento energético de la biomasa residual generada por la cosecha. El atlas de biomasa residual agrícola abre caminos en este sentido, al ser una herramienta que permite hacer una estimación del potencial energético para una zona y tipo de cultivo. El aprovechamiento de este recurso promueve la producción de energía limpia y renovable, lo cual no solo contribuye a la seguridad energética local, sino que reduce las emisiones de gases de efecto invernadero y fomenta prácticas agrícolas más eficientes y amigables con el ambiente. Así, se establece un modelo para un desarrollo rural sustentable, que equilibra las necesidades económicas y la conservación del entorno natural.

27 Se ajustó la imagen con la aplicación LetsEnhance, utilizando la opción Balancedx4, para obtener una resolución de 300 dpi.

Por su parte, la herramienta para la detección de cambios de cobertura en zonas urbanas representa un aporte significativo para lograr una expansión sostenible en este ámbito. Al identificar de manera precisa y oportuna los cambios en el entorno urbano, se pueden tomar decisiones informadas que promuevan la planificación responsable, la conservación de recursos naturales, la mejora de la calidad de vida de los habitantes y la reducción de impactos ambientales negativos, lo que fortalece el compromiso con un desarrollo sustentable y resiliente a largo plazo.

Un desafío importante que conlleva el uso de técnicas de IA es la disponibilidad de datos anotados, lo cual es crucial para entrenar modelos de *machine learning* y *deep learning* de manera efectiva. Una forma de solventar este problema es mediante el uso de modelos preentrenados, conocidos como *modelos fundacionales*, los cuales puedan transferir conocimiento aprendido a tareas específicas. Además, la aplicación de modelos generativos para la creación de datos sintéticos podría ser una estrategia prometedora para mitigar la escasez de datos anotados, lo que permite la expansión y diversificación del conjunto de datos disponible para el aprendizaje. Estas líneas de investigación podrían mejorar la robustez y eficiencia de los modelos, así como abrir nuevas posibilidades para la aplicación de técnicas avanzadas de IA en el análisis de imágenes satelitales.

La integración de la IA y la teledetección encierra un inmenso potencial para abordar los retos que implica un desarrollo sostenible y una gestión ambiental efectivos. Para avanzar, es fundamental establecer marcos regulatorios claros y éticamente sólidos, que guíen el desarrollo y despliegue de sistemas de IA en contextos de teledetección. Se debe fomentar la investigación en métodos que aseguren la interpretación y explicación de los resultados obtenidos por los modelos de IA, facilitando así la confianza y aceptación de estas tecnologías por parte de las comunidades y entidades comprometidas con el desarrollo de este tipo de proyectos. Además, la colaboración con diversas partes interesadas y la consulta pública son importantes para asegurar que estas tecnologías se implementen en beneficio de la sociedad en su conjunto.

Finalmente, es necesario recordar que si bien el binomio IA-teledetección alimenta grandes expectativas para favorecer la generación de energía verde y una expansión urbana con mayores garantías de calidad de vida para las personas, así como un acotado impacto ambiental, también es imperativo abordar, de manera activa y responsable, los riesgos para el ambiente natural asociados con su implementación y uso. Esto garantizará que los beneficios de la IA puedan maximizarse sin comprometer los objetivos de conservación ambiental y sostenibilidad a largo plazo.

Referencias

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D. y Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11). https://doi.org/10.3390/s21113758

Burke, M., Driscoll, A., Lobell, D. y Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, *371*(6535), eabe8628. https://doi.org/10.1126/science.abe8628.

Campbell, J. (2023). *Introduction to remote sensing* (6.ª ed.). The Guilford Press.

Carrijo, J., Miguel, E., Teixeira Do Vale, A., Matricardi, E., Monteiro, T., Rezende, A. y Inkotte, J. (2020). Artificial intelligence associated with satellite data in predicting energy potential in the Brazilian savanna woodland area. *iForest*, *13*, 48-55. https://doi.org/10.3832/ifor3209-012.

Chuvieco, E. (2016). *Fundamentals of satellite remote sensing. An environmental approach*. (2.ª ed.). CRC Press.

Copernicus Global Land Service. (s. f.). *Copernicus Global Land Service*. https://www1.upme.gov.co/

Departamento Nacional de Planeación (DNP). (s. f.). *TerriData: Sistema de Estadísticas Territoriales*. https://terridata.dnp.gov.co/

Díaz, N. (2023). *Herramienta para la estimación del potencial bioenergético en municipios vulnerables de Colombia mediante imágenes satelitales y machine learning* [tesis de maestría]. Universidad de los Andes.

Ferreira, B., Iten, M. y Silva, R. (2020). Monitoring sustainable development by means of earth observation data and machine learning: A review. *Environmental Sciences Europe*, *32*(120). https://doi.org/10.1186/s12302-020-00397-4.

Geron, A. (2022). *Hands-on machine learning with Scikit-Learn and TensorFlow. Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gruenhagen, L. y Juergens, C. (2022). Multitemporal change detection analysis in an urbanized environment based upon Sentinel-1 data. *Remote Sensing*, *14*(1043). https://doi.org/10.3390/rs14041043.

Holloway, J. y Mengersen, K. (2028). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, *10*(9), 1365. https://doi.org/10.3390/rs10091365.

Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). (s. f.). *Cobertura de la Tierra metodología CORINE Land Cover adaptada para Colombia*. https://visualizador.ideam.gov.co/geonetwork/srv/spa/catalog.search#/metadata/285c4d0a-6924-42c6-b4d4-6aef2c1aceb5

Janga, B., Asamani, G.P., Sun, Z. y Cristea, N. (2023). A review of practical AI for remote sensing in Earth sciences. *Remote Sensing*, *15*(16), 4112. https://doi.org/10.3390/rs15164112.

Jeon, G. (2023). Advanced machine learning and deep learning approaches for remote sensing. *Remote Sensing*, *15*(11), 2876. https://doi.org/10.3390/rs15112876.

Jung, J., Maeda, M., Chang, A., Bhandari, M., Ashapure, A. y Landivar-Bowles, J. (2021). The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology*, *70*, 15-22. https://doi.org/10.1016/j.copbio.2020.09.003.

Kim, J., Kim, D., Jun, H-J. y Heo, J-P. (2024). The detection of residential developments in urban areas: Exploring the potentials of deep-learning algorithms. *Computers, Environment and Urban Systems*, *107*(102053). https://doi.org/10.1016/j.compenvurbsys.2023.102053.

Li, F., Yigitcanlar, T., Nepal, M., Nguyen, K. y Dur, K. (2023). Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework. *Sustainable Cities and Society*, *96*(104653). https://doi.org/10.1016/j.scs.2023.104653.

Li, Y., Zhang, H., Xue, X., Jiang, Y. y Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(6). https://doi.org/10.1002/widm.1264

Lillesand, T., Kiefer, R. y Chipman, J. (2015). *Remote sensing and image interpretation*. Wiley.

Lindahl, J., Johansson, R. y Lingfors, D. (2023). Mapping of decentralized photovoltaic and solar thermal systems by remote sensing aerial imagery and deep machine learning for statistic generation. *Energy and AI*, *14*, 100300. https://doi.org/10.1016/j.egyai.2023.100300

Liu, H., Mou, C., Yuan, J., Chen, Z., Zhong, L. y Cui, X. (2024). Estimating urban forests biomass with LiDAR by using deep learning foundation models. *Remote Sensing*, *16*(9), 1643. https://doi.org/10.3390/rs16091643

Miller, L., Pelletier, G., Webb, G. (2024). Deep learning for satellite image time-series analysis: A review. *IEEE Geoscience and Remote Sensing Magazine*, *99*, 2-45. https://doi.org/10.1109/MGRS.2024.3393010

Ministerio de Tecnologías de la Información y las Comunicaciones. (s. f.). *Plataforma Nacional de Datos Abiertos de Colombia*. https://www.datos.gov.co/

Paletta, Q., Terrén-Serrano, G., Nie, Y., Li, B., Bieker, J., Zhang,W., Dubus, L., Dev, S. y Feng, C. (2023). Advances in solar forecasting: Computer vision with deep learning. *Advances in Applied Energy*, *11*, 100150. https://doi.org/10.1016/j.adapen.2023.100150.

Persello, C., Wegner, J., Hänsch, R., Tuia, D., Ghamisi, P., Koeva, M. y Camps-Valls, G. (2022). Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, *10*(2), 172-200. https://doi.org/10.1109/MGRS.2021.3136100.

Rincón Martínez, J. M., Durán Hernández, D. M., Quintero Montoya, O., Duarte González, C. S., Guevara Patiño, P. O. y Velásquez Lozano, M. E. (2019). Disponibilidad de biomasa residual y su potencial para la producción de biogas en Colombia. *Revista CIDET*, (19), 16-25. http://revista.cidet.org.co/revistas/revista-19/.

Senocak, A. y Guner, H. (2022). Forecasting the biomass-based energy potential using artificial intelligence and geographic information systems: A case study. *Engineering Science and Technology an International Journal*, *26*(5). https://doi.org/10.1016/j.jestch.2021.04.011

Teixeira, I., Morais, R., Sousa, J. y Cunha, A. (2023). Deep learning models for the classification of crops in aerial imagery: A review. *Agriculture*, *13*(5), 965; https://doi.org/10.3390/agriculture13050965.

Thapa, A., Horanont, T., Neupane, B. y Aryal, J. (2023). Deep learning for remote sensing image scene classification: A review and meta-analysis. *Remote Sensing*, *15*(19), 4804. https://doi.org/10.3390/rs15194804

Unidad de Planeación Minero Energética (UPME). (2021). *Atlas del potencial energético de la biomasa residual en Colombia*. https://www1.upme.gov.co/siame/Paginas/atlas-del-potencial-energetico-de-la-biomasa.aspx

Unidad de Planeación Minero Energética (UPME). (s. f.). *UPME: Unidad de Planeación Minero-Energética*. https://www1.upme.gov.co/

Unidad de Planificación Rural Agropecuaria (UPRA). (s. f.). *Reporte: Evaluaciones agropecuarias-EVA y Anuario Estadístico del Sector Agropecuario*. https://www.agronet.gov.co/estadistica/paginas/home.aspx?cod=59

Vásquez, D. (2023). *Detección de crecimiento urbano a través de imágenes de radar multitemporal e inteligencia artificial* [tesis de maestría]. Universidad de los Andes.

Veneri, P., Banquet, A., Delbouve, P. y Daams, M. (2022). *Monitoring land use in cities using satellite imagery and deep learning*. OECD *Regional*

Development Papers, *28*. Organización para la Cooperación y el Desarrollo Económico (OECD). https://doi.org/10.1787/dc8e85d5-en.

Wu, P., Zhang, Z. y Peng, X. (2024). Deep learning solutions for smart city challenges in urban development. *Scientific Reports*, *14*(5176). https://doi.org/10.1038/s41598-024-55928-3

INNOVACIONES DE LA WEB SEMÁNTICA PARA LA EDUCACIÓN SUPERIOR

Olga Mariño, Gilbert Paquette,
Rubén Manrique

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.06

Introducción

Desde sus orígenes en la década de los cincuenta, la inteligencia artificial avanzó en dos enfoques: el conexionista, que busca emular el cerebro humano con sus neuronas y conexiones; y el simbólico, que pretende emular el razonamiento humano, centrándonse en la lógica, las reglas causales y las conexiones entre conceptos y relaciones. El enfoque conexionista evolucionó hacia el *machine learning*, mientras que el simbólico lo hizo hacia la web semántica. Según su creador, Tim Berners Lee, la web semántica se basa en representaciones formales de conocimiento compartidas que pueden evolucionar y en agentes de *software* que pueden manipular estas representaciones (Berners-Lee *et al*., 2001). Por lo tanto, el objetivo principal de la web semántica es introducir descripciones explícitas sobre el significado de los recursos, para permitir que las máquinas tengan un nivel de comprensión del contenido de la web (Castells, 2003).

El potencial de la web semántica depende de la representación del conocimiento y de la manipulación automática de estas representaciones por aplicaciones de *software*. La representación del conocimiento en la web semántica ha evolucionado desde listas de conceptos, taxonomías o vocabularios controlados, hasta mapas de temas, modelos conceptuales y, más recientemente, ontologías y grafos de conocimiento (KG).

Una ontología se define como "una especificación formal explícita de una conceptualización compartida" (Gruber, 1993, p. 199). Las ontologías describen un dominio específico de conocimiento usando un vocabulario de clases y relaciones para caracterizar las entidades relacionadas. El componente principal de una ontología, llamado *tripleta*, inspirado en la estructura de una oración:

sujeto-verbo-predicado, está formado por dos conceptos y una relación entre ellos. Al combinar tripletas, la ontología se convierte en un grafo dirigido de conceptos y relaciones, que puede enriquecerse aún más con restricciones y reglas. Los principales lenguajes de la web semántica para la representación de ontologías son Resource Description Frameworks (RDF), RDF Schema (RDFS) y Web Ontology Language (OWL), mientras que el principal lenguaje de consulta para ontologías es Protocol and RDF Query Language (SPARQL), el cual resuelve una consulta mediante la unificación de sus variables con partes del grafo de tripletas.

Cada vez más ontologías están disponibles en la web, como la ontología de vida silvestre de la BBC[1], la ontología médica SNOMED[2] o GoodRelations[3], vocabulario de comercio electrónico.

Aunque las ontologías de dominio ofrecen una gran oportunidad, todo el poder de la web semántica se alcanza cuando estas son abiertas, libremente accesibles y están conectadas entre sí, creando una nube de datos abiertos enlazados (*linked open data*, LOD) (Ontotext, 2017). Estos datos pueden explotarse para construir sistemas más inteligentes o mejorar la efectividad de los algoritmos existentes (Musto *et al*., 2017). De este enorme espacio de datos, los KG —como OpenCyc, Freebase, Wikidata, YAGO y DBpedia— son en particular importantes, ya que concentran el conocimiento de múltiples dominios y especifican una gran cantidad de interrelaciones entre conceptos (Paulheim, 2017). Solo DBpedia cuenta con más de 4 800 000 instancias y 170 000 000 de tripletas.

La figura 6.1 muestra el concepto "red neuronal artificial" y algunos de sus enlaces existentes en el KG de DBpedia. El concepto pertenece a dos categorías que forman parte de una de las estructuras jerárquicas de DBpedia, la inducida por la relación "más general" (skos:broader). Los conceptos en un KG también están vinculados mediante propiedades no jerárquicas, las cuales expresan otras relaciones semánticas, como es el caso del concepto "aprendizaje automático (*machine learning*) en el ejemplo.

Las ontologías y el conocimiento de la web semántica pueden ser explotados por los humanos para varias tareas, como entender y aprender de un dominio u organizar recursos relacionados con el dominio al diseñar un curso. Pero la visión de Berners Lee de la web semántica funcional va más allá del uso

1 Véase https://www.bbc.co.uk/ontologies/wo

2 Véase http://www.snomed.org

3 Véase http://wiki.goodrelations-vocabulary.org/Quickstart

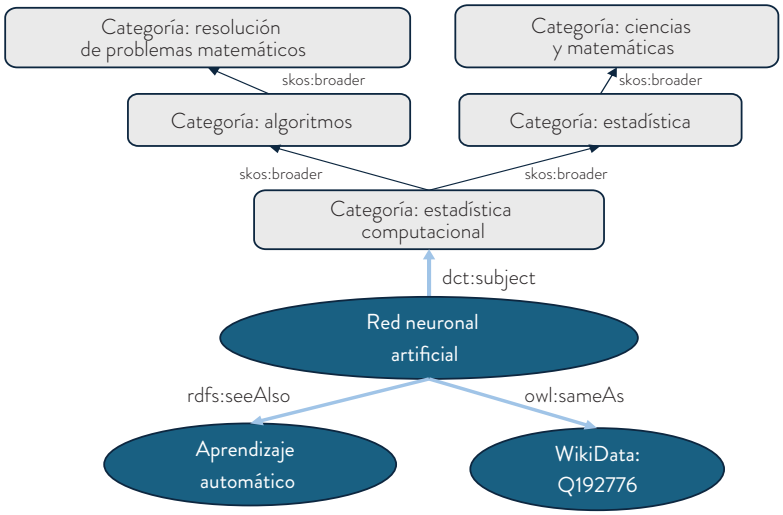


Figura 6.1. Extracto del conocimiento presentado en DBpedia para el concepto *red neuronal artificial*

Fuente: elaboración propia.

humano de las ontologías, incluye el razonamiento automatizado logrado por "agentes de software que recolectan contenido web de diversas fuentes, procesan la información" y "realizan tareas sofisticadas para los usuarios" o por servicios web que "intercambian los resultados con otros programas" (Berners-Lee *et al.*, 2001, pp. 39-40).

En la próxima sección analizamos los beneficios del uso de la web semántica para mejorar la educación superior. Luego, presentamos dos aplicaciones principales que desarrollamos en el Laboratorio de Informática Cognitiva y Ambientes de Formación (LICEF) de la Télé-université du Quebec (TELUQ), en Canadá. La siguiente sección muestra el conjunto de herramientas que desarrollamos en la Universidad de los Andes, en Colombia. Por último, concluimos con una reflexión global.

Oportunidades de la web semántica para enriquecer la educación superior

Aprendizaje enriquecido con computador y la web semántica

El uso de las tecnologías de la información y las comunicaciones —y en particular de la inteligencia artificial y la web semántica— para enriquecer entornos de aprendizaje ha evolucionado a la par con las teorías de aprendizaje.

En los orígenes del aprendizaje apoyado con computador (*computer based learning*, CBL), los entornos de aprendizaje tomaban por lo general la forma de instrucción programada basada en el conductismo (Skinner, 1954). El alumno adquiría pequeños fragmentos de contenido estático proporcionados en secuencias, seguidas de preguntas de opción múltiple. El sistema daba la misma retroalimentación a todos los alumnos, según las respuestas típicas. Estos sistemas de ejercicios y prácticas luego se mejoraron para incluir un modelo de alumno básico, con el fin de personalizar los puntos de entrada y la complejidad de los ejercicios sin dejar de estar alineados con el conductismo. A partir de los años ochenta y noventa, los entornos de aprendizaje avanzaron hacia sistemas más heurísticos, como juegos educativos, simuladores y entornos de aprendizaje por descubrimiento, que promueven el cognitivismo (Newell y Simon, 1972) y el constructivismo (Le Moigne, 2001; Piaget, 1936). Posteriormente, los entornos de aprendizaje colaborativo proporcionaron aplicaciones concretas del socioconstructivismo (Vygotsky *et al*., 1978).

Por su parte, la integración de inteligencia artificial a los sistemas de aprendizaje (*artificial intelligence in education*, AIED) comenzó con el uso de sistemas expertos como laboratorios de aprendizaje, alineados con el enfoque constructivista, mientras que el surgimiento de los sistemas de tutoría inteligentes permitió integrar un modelo del estudiante y del dominio para personalizar la selección y secuencia de ejercicios, lo que enriqueció los sistemas conductistas de ejercicios (Wenger, 1987). Todos estos sistemas se basan en el paradigma de la inteligencia artificial simbólica y en representaciones de conocimiento encapsuladas en el sistema.

En las últimas décadas, los avances de la ingeniería de *software* en arquitecturas basadas en composición dinámica de servicios abrieron la puerta a portales de aprendizaje programables, así como a la creación y fácil adaptación de sistemas de gestión de aprendizaje personalizables. En este contexto, el auge de la web semántica (Berners-Lee *et al*., 2001) y el posterior énfasis en la web de datos abiertos enlazados (Allemang y Hendler, 2011) ofrecen la posibilidad de desarrollar plataformas orientadas a servicios semánticos (Carbonaro, 2020) y entornos de aprendizaje basados en representaciones de conocimiento abierto (García *et al*., 2013), para personalizar el aprendizaje y recomendar actividades y recursos (Figueroa *et al*., 2015), eventualmente mejoradas con técnicas de aprendizaje automático.

Desde el punto de vista educativo, los estudiantes de las nuevas generaciones están desafiando el modelo clásico de enseñanza dentro del aula, lo que obliga a las instituciones a innovar con modelos como *b-learning*, *e-learning* y *massive open online courses* (MOOC); a diseñar cursos y currículos basados en competencias; a integrar tecnologías, como motores de búsqueda web y redes

sociales, en sus actividades de aprendizaje; y a ayudar a cada alumno a encontrar su propio camino de aprendizaje. La web semántica ofrece grandes posibilidades para abordar estos desafíos (Devedzic, 2004). En el resto de la sección profundizaremos en estas posibilidades.

Uso de ontologías para diseñar cursos y programas

El diseño de un curso o programa académico parte de una identificación de objetivos y competencias, cuyo logro implica la apropiación de un conjunto de conocimientos. Tradicionalmente, los profesores organizan estos conocimientos en una estructura jerárquica de temas y subtemas y, a partir de ella, identifican las actividades y recursos de un programa. Las ontologías proporcionan una semántica y relaciones más completas entre conceptos dentro de un dominio de conocimiento, lo que permite al diseñador del curso idear diferentes estrategias de apropiación del contenido para un mismo objetivo. En la actualidad existen múltiples ontologías de dominio disponibles en la web validadas por grandes equipos de académicos (Quezada-Sarmiento *et al*., 2020; Tapia-León *et al*., 2019), así como herramientas para construir su propia ontología (por ejemplo, GMOT presentada en la siguiente sección). Las ontologías también se han utilizado como insumo para las actividades de aprendizaje propuestas a los estudiantes, con el fin de que comprendan y exploren el dominio de aprendizaje. Así mismo, cabe resaltar que el diseño de cursos basado en ontologías permite una mejor documentación y, por ende, sostenibilidad y mantenimiento de un curso.

Por último, un gran desafío para el diseñador de programas y cursos es la sostenibilidad y la posibilidad de reutilizar recursos de aprendizaje. Desde el trabajo de Dietze *et al*. (2013), la comunidad de aprendizaje basada en la web semántica ha estado activa proponiendo formas de conectar esta y los recursos de aprendizaje, no solo al diseñar cursos, sino también para la anotación, la clasificación, el descubrimiento, la personalización y la recomendación de recursos.

Además de las ontologías para el conocimiento de un dominio temático, se han desarrollado ontologías específicas para la educación superior[4]. La *ontología de Bolonia* define los términos de un esquema estándar para las universidades europeas involucradas en la Reforma de Bolonia, los cuales garantizan la comparabilidad en los estándares. La *estructura interna de la institución académica* (AIISO) proporciona un esquema para describir la estructura organizativa

4 Véase https://linkededucation.wordpress.com/data-models/schemas/

interna de una institución académica; está diseñada para trabajar junto con la ontología *AIISO-roles*, que describe los roles que desempeñan las personas dentro de una institución. También existe la ontología *lista de recursos*, que recoge cursos académicos y colecciones de referencias, como listas de lectura, marcadores y bibliografías.

Diseño de personalización basado en web semántica en cursos de educación superior

La personalización se refiere a una instrucción adaptada a las necesidades o preferencias de aprendizaje y a los intereses y competencias específicos de los diferentes alumnos. En un entorno totalmente personalizado, los objetivos y el contenido del aprendizaje, así como el método y el ritmo, pueden variar, logrando una personalización que abarque diferenciación e individualización. La llegada de los MOOC ha hecho que la personalización sea aún más necesaria que antes. El mismo curso puede ser seguido por miles de estudiantes en diversas partes del mundo, todos con diferentes antecedentes, conocimientos y culturas, lo que hace difícil, si no imposible, proporcionar un entorno de aprendizaje único apto para todos.

Las tecnologías de la web semántica se han utilizado ampliamente para anotar recursos con los conocimientos y las relaciones de las que tratan, para que sea posible recomendarlos a los estudiantes (Denaux *et al.*, 2005; Dolog *et al.*, 2003), personalizar las experiencias de aprendizaje y proporcionar recomendaciones (Jevsikova *et al.*, 2017). Una solución para la personalización es crear un camino de aprendizaje individualizado compuesto de recursos seleccionados, de acuerdo con los objetivos de aprendizaje del alumno. Otra solución se basa en la colaboración grupal en una comunidad de estudiantes que comparte un escenario central predefinido, el cual puede evolucionar y actualizar dinámicamente los perfiles de los estudiantes, recomendar recursos y actividades adaptados y ajustar el escenario inicial a perfiles individuales. Por supuesto, ambos enfoques se pueden combinar de muchas maneras.

Ayudar a los estudiantes a navegar en la web

Los estudiantes actuales son ciudadanos digitales cuya principal fuente de información es la web, tanto para tareas personales como de aprendizaje (Nadzir, 2015; Nikolopoulou y Gialamas, 2011). Si bien la web contiene mucha información útil, el número de documentos y datos falsos e inexactos ha ido creciendo de forma constante en las últimas décadas. Los estudiantes de educación

superior recurren a buscadores generalistas, como Google o Yahoo, como primera fuente de información y conocimiento para un objetivo de aprendizaje o tarea de investigación, no siempre recuperando información relevante, actualizada o incluso veraz. Si bien los repositorios de objetos de aprendizaje han intentado abordar este problema al seleccionar y anotar recursos de aprendizaje, no pueden competir con la impresionante cantidad de información en el resto de la web, y solo un pequeño número de estudiantes utiliza estos repositorios como su primera fuente de información. La búsqueda de información se ha convertido en una actividad central en el aprendizaje, hasta el punto de que la investigación reciente sobre "la búsqueda como aprendizaje" se centra en comprender y mejorar los procesos de aprendizaje que tienen lugar al realizar búsquedas en línea (Ghosh *et al*., 2018; Moraes *et al*., 2018; Tibau *et al*., 2018). Proporcionar herramientas de búsqueda basadas en la web semántica, como las presentadas en esta investigación, puede ayudar a los estudiantes a navegar por la web de manera más adecuada y eficiente.

La web semántica como soporte a la colaboración y a las comunidades de aprendizaje

La web semántica tiene un gran potencial para integrar y enriquecer las redes sociales. Esta evolución hacia la *web semántica social* apoya la colaboración para el aprendizaje en MOOC u otros tipos de entornos de aprendizaje en línea basados en el trabajo comunitario, en escuelas u organizaciones (Da Costa *et al*., 2017; Tiropanis *et al*., 2009). Las tecnologías semánticas contribuyen a las comunidades de aprendizaje al vincular documentos, datos y aplicaciones involucradas en una variedad de situaciones, rompiendo así el efecto silo en la web social. Por su parte, las comunidades en línea que utilizan *software* de la web social producen datos e información masivos, los cuales pueden procesarse mediante técnicas de análisis de aprendizaje para extraer nuevos conocimientos para la web semántica, con el fin de crear aplicaciones más inteligentes que las disponibles en la actualidad.

Herramientas de web semántica para gestión de recursos, referenciación de conocimientos/competencias, búsqueda y recomendación

El auge de la web semántica coincidió con avances importantes en ingeniería de *software*, en particular la arquitectura basada en modelos y el desarrollo de sistemas orientado a servicios web. Con estas técnicas se pretende encapsular

los componentes de un sistema, de forma que este se modele como un conjunto desacoplado de subsistemas (en la web), que se prestan servicios entre ellos y al todo. En este contexto, surge la arquitectura basada en ontologías, para modelar los componentes a partir de su semántica, y los sistemas basados en servicios de web semántica, en donde cada servicio está descrito con elementos de ontologías tecnológicas en términos de funcionalidad, requerimientos de configuración, etc., de modo que automáticamente se pueda validar y conectar al sistema.

El centro de investigación LICEF, de la TELUC, lleva varios decenios realizando investigación aplicada en inteligencia artificial para el mejoramiento de la educación, desde desarrollos fundamentados en sistemas expertos y tutores inteligentes, hasta llegar a *frameworks* o metasistemas basados en servicios de web semántica, donde estos servicios, los cuales es posible utilizar de forma autónoma, sustentan a su vez el diseño de cursos basado en ontologías y personalización.

En esta sección analizamos dos aplicaciones de las tecnologías de la web semántica (Domingue *et al*., 2011) desarrolladas en el centro LICEF, que ilustran el uso de la web semántica en dos niveles: (1) la ejecución de servicios de plataformas educativas y (2) el referenciamiento, búsqueda y recomendación de recursos basados en conocimientos y competencias; más importante aún, proporcionan herramientas a los diseñadores de aprendizaje para implementar ambientes de aprendizaje cognitivistas, constructivistas y socioconstructivistas.

TELOS, una plataforma basada en ontologías

Technology Enhanced Learning Operating System (TELOS) (Paquette *et al*., 2007), desarrollado en el LICEF, es un sistema de web semántica basado en una ontología técnica explícita, la cual estructura los objetos que el sistema va a procesar, actuando como su modelo ejecutable (Davies *et al*., 2003; Tetlow *et al*., 2006). La ejecución de los servicios de TELOS se realiza mediante consultas a la ontología técnica del sistema. Los escenarios multiactores proporcionan el mecanismo central de agregación del entorno de aprendizaje, que agrupa a los actores, las operaciones que llevan a cabo y los recursos que utilizan o producen. El editor de escenarios multiactor y su motor de ejecución son un componente central de TELOS; proporcionan apoyo para el diseño de entornos de aprendizaje.

En el siguiente enlace[5] se muestra la interfaz de usuario de escritorio de TELOS en un navegador web, con tres herramientas principales abiertas: el

5 Encuentre y escanee el QR respectivo en el anexo.

administrador de recursos, el editor de escenarios y el administrador de tareas. El editor GMOT OWL, un editor de perfiles de competencias y otras herramientas, también se puede iniciar desde el escritorio de TELOS: https://gobierno.uniandes.edu.co/wp-content/uploads/Uniandes-Fig-A.1-1.png

El administrador de recursos de TELOS sirve para integrar y gestionar los recursos que los actores utilizan o producen en el sistema, incluidos los alumnos, individualmente o en equipos, y los facilitadores, es decir, profesores, tutores, expertos en contenidos o diseñadores. Los recursos se clasifican en clases de la ontología técnica integrada en el sistema, para guiar la ejecución de sus servicios de forma adaptada a la especificidad de cada clase. Para recursos de tipo "Escenario", las funciones "Ver" y "Modificar" abren el editor de escenarios que se muestra en la segunda ventana de la imagen del enlace, mientras que la opción "Ejecutar" inicia el motor de inferencia que procesará el escenario y lo presentará a sus usuarios en el administrador de tareas. Para recursos de tipo "Usuarios TELOS", las funciones "Ver" y "Modificar" abren un navegador de usuario para ver o ingresar información personal, como correos electrónicos, fotografías, portafolios, etc. Esta herramienta está vinculada a un portafolio electrónico que presenta las competencias actuales de un usuario y algunas previdencias (evaluaciones, proyectos) de su adquisición. La información sobre los usuarios y otros recursos está disponible durante los procesos de ejecución del escenario. Los componentes de *software* se almacenan como operaciones en el administrador de recursos. Al seleccionar dichos recursos, se lanzan durante la ejecución del escenario para proporcionar una variedad de servicios web.

El editor de escenarios TELOS (Paquette, 2010) aporta un lenguaje de programación visual de alto nivel (MOT), que incluye símbolos conceptuales para representar documentos, herramientas, recursos semánticos, entre otros; símbolos de procedimiento, que representan procesos descomponibles en actividades realizadas por humanos o máquinas, posiblemente relacionadas con condiciones; y símbolos de actores, que representan usuarios, grupos, roles o agentes de *software*. Este editor se utiliza en TELOS a diferentes niveles. En un primer nivel permite describir y ejecutar un diseño instruccional (curso) que puede tener consideraciones valiosas, como abordar la diversidad cultural en un grupo de estudiantes (Savard *et al*., 2013) o personalizar características en un entorno de aprendizaje MOOC (Bejaoui *et al*., 2017). En un nivel superior también se usa para diseñar y establecer una configuración particular del sistema, es decir, un sistema de administración del aprendizaje (*learning management system*, LMS) particular.

Los alumnos y facilitadores utilizan el administrador de tareas TELOS para interactuar con algunos escenarios en tiempo de ejecución. Se abrirá automáticamente para sus usuarios cuando se inicie un escenario seleccionado en el

administrador de recursos. El administrador de tareas, guiado por la ontología técnica, presenta interfaces adaptadas potencialmente diferentes para cada participante en el escenario. Por ejemplo, el profesor podría ver todos los documentos y las barras de tareas de progreso de todos los alumnos involucrados en el escenario, mientras que los alumnos verían solo las tareas en las que están involucrados y los documentos o herramientas que se supone que deben usar. Esta flexibilidad del sistema es posible gracias a su diseño y ejecución basados en la web semántica.

Ejemplo de un escenario socioconstructivista

El escenario pedagógico presentado en los enlaces[6] https://gobierno.uniandes.edu.co/wp-content/uploads/Uniandes-Fig-A.1-1.png y https://gobierno.uniandes.edu.co/wp-content/uploads/Uniandes-Fig-A.2-1.png es un ejemplo simple de un escenario socioconstructivista en el que participan un profesor y dos equipos de alumnos. En el acto 1, los alumnos reciben una lista de conocimientos y competencias objetivo sobre los planetas del sistema solar y se organizan en dos equipos; cada alumno y cada equipo que los agrupa se registran en el administrador de recursos. En el acto 2, cada grupo recibe recursos, información sobre los planetas y objetivos particulares por discutir, junto con herramientas de comunicación y configuración de roles para la discusión. En el acto 3, que se activa luego de determinado tiempo, los grupos se disuelven y cada alumno debe construir soluciones individuales a algunos problemas relacionados con los planetas. En el acto 4 participarán en una evaluación grupal y un foro plenario.

Cuando se han realizado las elecciones pedagógicas en el editor de escenarios, los diseñadores del entorno deben decirle a TELOS qué representa cada ícono en el escenario gráfico, en términos de su ontología técnica, para que el sistema lo procese; esto se denomina *semántica de ejecución del ícono*. Una interfaz en el editor de escenarios proporciona un servicio a los diseñadores para establecer las propiedades semánticas de cada recurso en el escenario. Al usarlo, el diseñador puede decirle al sistema que un determinado ícono es una actividad que debe mostrarse en el administrador de tareas en tiempo de ejecución, que otro ícono es un usuario o un grupo, o que otro ícono es un documento de un determinado tipo. Es importante notar que en el tiempo de ejecución cada grupo y cada alumno podrá tener experiencias diferentes en función de sus roles en el grupo, tareas, instrucciones y recursos, y el profesor no necesitará gastar tiempo en explicar el proceso.

6 Encuentre y escanee los QR respectivos en el anexo.

Referenciamiento de recursos por conocimientos y competencias

Otro uso importante de la web semántica es el referenciamiento de recursos a través de ontologías de dominio aumentadas por una estructura de competencias. En TELOS, este referenciamiento sirve para referenciar no solo recursos, sino también actividades y usuarios, al igual que para informar a los estudiantes y diseñadores sobre el conocimiento y la competencia que incorpora un recurso; habilitar métodos de búsqueda de recursos basados en sus propiedades de conocimiento y competencia; informar a los agentes de recomendación para que puedan ayudar a los usuarios a realizar determinadas actividades, y recomendar recursos adecuados a sus conocimientos y competencias reales.

El proceso se desarrolla en tres pasos. Primero se selecciona una ontología de dominio existente o se construye directamente mediante el editor de ontología de TELOS. En segundo lugar, a través del editor de competencias de TELOS, se construye una estructura de competencias, asociando habilidades y niveles de desempeño a algunos elementos de conocimiento en la ontología del dominio. En tercer lugar, TELOS proporciona una herramienta de referenciamiento que permite asociar a un estudiante con un conocimiento y una competencia relacionada a este, indicando su nivel actual, y vincular a actividades y recursos con dos parejas de conocimiento/competencia, que indican los prerrequisitos para entender y aprovechar ese elemento y los objetivos que busca apoyar. El sistema podrá entonces validar y personalizar el aprendizaje en función de la relación entre los conocimientos y competencias del estudiante y de las actividades y recursos a su disposición, como se explica a continuación.

Búsqueda y recomendación basada en competencias

Para permitir la búsqueda de recursos basada en competencias y brindar recomendaciones a los usuarios, desarrollamos un algoritmo que compara dos competencias: $C_1=(K_1,S_1,P_1)$ y $C_2=(K_2,S_2,P_2)$, donde *K* es la parte de conocimiento, *S* es el nivel de habilidad y *P* es el nivel de desempeño, para evaluar la proximidad semántica o cercanía entre ellas, en función de las respectivas posiciones de sus partes de conocimiento en el gráfico de ontología del dominio y de los valores de los niveles de habilidad y desempeño (Paquette *et al.*, 2012). Se puede insertar un agente de recomendación que utilice este algoritmo en cualquier punto de un escenario. Es posible evaluar si la competencia real de un usuario está muy cerca, cerca o lejos del requisito previo o las competencias objetivo de un recurso, para recomendar el uso de este recurso o no. El agente también puede comparar las competencias reales de un usuario con las competencias de

otros usuarios, para recomendar emparejarlos en un equipo para determinadas actividades, así como evaluar si una competencia es más fuerte o más débil que otra, según los niveles de habilidad y desempeño, o determinar si la competencia es más específica o más general, de acuerdo con las posiciones de los componentes del conocimiento correspondientes en la ontología.

El algoritmo de comparación de conocimientos y competencias también se utiliza en TELOS para permitir la búsqueda de recursos basada en ontologías (documentos, tareas, usuarios, etc.). Permitimos buscar a partir de un identificador de conocimiento o de competencia los recursos anotados con él o recursos cercanos, y a partir de un recurso, recursos similares semánticamente.

Administración de recursos en la web de datos hilados

Después de una década de investigación y práctica en el campo de los repositorios de recursos educativos abiertos (REA), ha aparecido una serie de limitaciones para su uso generalizado; muchas de estas requieren de nuevos enfoques. En LICEF, nuestras herramientas para la gestión de recursos evolucionaron desde una herramienta de primera generación llamada Paloma (Paquette *et al.*, 2004), basada en la tradicional metatada de objetos de aprendizaje, la cual usa tecnología de bases de datos relacionales, hasta el uso de anotaciones basadas en ontologías dentro del sistema TELOS. Por último, pasamos a la herramienta Comète, que utiliza tecnologías semánticas para la web de datos abiertos vinculados (Heath y Bizer, 2011).

Este sistema de segunda generación se utiliza en las universidades de Quebec para la búsqueda y referencia de recursos educativos. El mayor problema técnico que se resolvió fue la falta de interoperabilidad entre los distintos repositorios de recursos que existen en todo el mundo. Algunos de estos utilizan un esquema propietario (como en TELOS), junto con estándares como Dublin Core (DC) o Learning Object Metadata (LOM); sin embargo, los repositorios LOM emplean una diversidad de perfiles de aplicaciones, cada uno de los cuales utiliza diferentes vocabularios controlados, por lo que existen importantes dificultades para buscar recursos en varios repositorios. Por ejemplo, una búsqueda de recursos en Francia o España para un curso en un determinado nivel educativo no necesariamente encontrará recursos en repositorios de otros países, donde este nivel educativo no existe o está etiquetado de manera diferente. Por otra parte, no siempre se utilizan los mismos términos en diferentes repositorios para conceptos de un dominio con el mismo significado o uno relacionado.

La solución a estos problemas de interoperabilidad es hacer referencia a recursos a través de tecnologías de web semántica. El estándar ISO/IEC 19788

(ISO-MLR, 2013) propone un esquema de referencia RDFS, en el que todos los repositorios pueden consultarse utilizando el lenguaje SPARQL. Esta propuesta tiene como objetivo proporcionar una compatibilidad óptima con repositorios estandarizados y, al mismo tiempo, evitar la proliferación de perfiles de aplicaciones no interoperables. Lo más importante es que permite la búsqueda dentro del grafo LOD, mediante conjuntos de datos ampliamente utilizados como DBpedia, Foaf, Geobase, etc., para hacer referencia a recursos con cualquier vocabulario LOD, incluidos DC y LOM.

Nuestro gestor de recursos Comète se basa en estos principios. Permite recolectar recursos educativos que constituyen el patrimonio de una organización, cualquiera que sea el esquema de metadatos que utilice. De igual manera, integra las descripciones de los recursos cosechados en un grafo homogéneo de tripletas RDF. En la figura 6.2 se muestran íconos interconectados para recursos, autores o contribuyentes, organizaciones a las que pertenecen y referencias de vocabulario/ontología para el contenido de los recursos.

Mediante diversas técnicas, el sistema intenta maximizar la coherencia interna del grafo. Su módulo "Identidad" se encarga de la importación de identidades, que representan personas u organizaciones, asegurándose de que cada una siga siendo única. El módulo "Vocabulario" implementa la gestión de vocabularios, tesauros y ontologías, y gestiona las correspondencias entre conceptos y propiedades en ontologías de diversas fuentes.

El uso de tecnologías de web semántica supera los problemas de interoperabilidad entre repositorios. No es necesario ceñirse a una especificación única como la LOM, con muchos campos por completar y que compite con otras normas; así, se hace referencia a los recursos con su contenido de conocimiento y varias otras propiedades. La web de datos hilados rompe el efecto silo y representa un gran repositorio, en el que es posible recolectar y buscar todo tipo de repositorios.

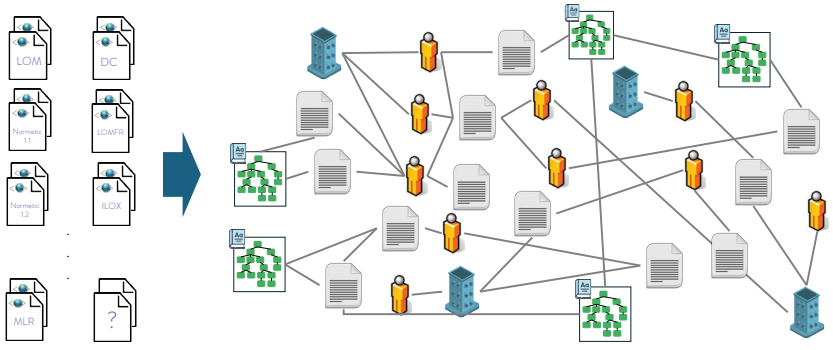


Figura 6.2. Integrar registros de metadatos en grafos RDF
Fuente: elaboración propia.

Beneficios educativos de los sistemas basados en web semántica

Plataformas como TELOS se constituyen en modelos de campus virtuales multinivel, los cuales permiten desde configurar plataformas para una institución, hasta diseñar y ejecutar módulos de cursos, al ofrecer herramientas visuales para el diseño de escenarios multiactores con actividades y recursos de múltiples tipos. Su enfoque basado en ontologías garantiza que el código final responda a los requerimientos iniciales.

Por su parte, la anotación semántica de recursos educativos integrada a sistemas de almacenamiento y búsqueda, como el sistema Comète, facilita la interoperabilidad y reutilización de repositorios y asegura una anotación de recursos basada en ontologías depuradas.

Estas tecnologías, ejemplificadas en estos dos sistemas, buscan reducir la interferencia tecnológica y, por ende, el tiempo y esfuerzo dedicado a montar sistemas de *e-learning*, lo que permite a los actores centrarse en la dimensión enseñanza-aprendizaje.

Herramientas de web semántica para aprendizaje a lo largo de la vida

Los retos del aprendizaje autónomo

Las nuevas generaciones están desafiando los paradigmas de la educación, y la educación superior no es una excepción. Cada vez más adultos se están convirtiendo en aprendices de por vida. Incluso mientras siguen programas académicos formales, esta nueva generación está ampliando sus conocimientos a través de procesos de adquisición de conocimientos fuera del aula, utilizando tecnologías tan simples como un motor de búsqueda genérico (Kurt y Gursel, 2018).

El paradigma del aprendizaje permanente apoyado en la tecnología está redefiniendo el enfoque clásico centrado en el tutor, pasando de modelos cerrados en los que los objetivos, contenidos y secuencias están predeterminados, hacia escenarios más abiertos y autodirigidos. En este formato, los alumnos son responsables de buscar, seleccionar y organizar los recursos más apropiados para lograr sus objetivos de aprendizaje. En los últimos años, la web ha asumido un papel destacado como proveedor de recursos de aprendizaje: presentaciones de Slideshare, videos de YouTube, preguntas/respuestas en foros, publicaciones de blogs y artículos de noticias son ejemplos de recursos web que se han utilizado con fines de aprendizaje. *Aprendizaje basado en recursos* (*resource-based learning*, RBL) es el nombre que se le da al aprendizaje autorregulado que utiliza

recursos que se encuentran en la web (Helen, 2014). En el aprendizaje permanente, los alumnos son empujados continuamente a un escenario de RBL. En tal entorno, los estudiantes autodirigidos enfrentan el desafío de buscar y seleccionar de manera efectiva contenido que tal vez no esté anotado educativamente, así como de organizar este contenido de una forma pedagógicamente sólida.

El desafío de la selección surge porque los estudiantes no siempre tienen suficientes habilidades de alfabetización para realizar una búsqueda eficaz y selectiva (Hill, 2012). Hay un gran volumen de recursos disponibles en la web, que puede abrumar a los estudiantes, en especial en dominios desconocidos. Los recursos de aprendizaje en la web por lo general no están bien indexados, debido a la falta de anotaciones de metadatos. Como resultado, la respuesta de los motores de búsqueda tradicionales puede no ser adecuada al objetivo de aprendizaje (Changuel *et al*., 2015). Si seleccionar los recursos adecuados es difícil para un estudiante autodirigido, es aún más difícil para él poder organizar los contenidos de aprendizaje recuperados de una manera útil para su propósito de aprendizaje (Scholl *et al*., 2007). Esto es un gran obstáculo porque, como afirma la teoría de la elaboración de Charles Reigeluth, aprender conceptos complejos requiere de la comprensión de otros más básicos (Talukdar y Cohen, 2012). El principio clave de esta teoría es que el contenido que se enseña debe organizarse desde lo más simple y luego aumentar el orden de complejidad, siguiendo las relaciones de requisitos previos entre los conceptos involucrados (Reigeluth y Darwazeh, 1982). A pesar de haber sido propuesta hace más de cuarenta años, esta teoría sigue siendo relevante; de hecho, la mayoría de los contenidos de los cursos MOOC están organizados en una secuencia que sigue este principio (Manrique, 2019; Manrique *et al*., 2018b).

Un enfoque basado en web semántica para apoyar el aprendizaje a lo largo de la vida

La Universidad de los Andes ha sido pionera en América Latina en investigación en tecnología para el mejoramiento de la educación; al igual que la TELUQ, en Canadá, ha integrado la inteligencia artificial desde sus comienzos a sus propuestas y trabaja en la actualidad de forma activa en la web semántica para educación. A continuación se presenta el trabajo realizado en los últimos años para apoyar el aprendizaje autónomo con sistemas basados en web semántica. Además de apoyarse en esta tecnología, nuestra propuesta usa técnicas de procesamiento de lenguaje natural y aprendizaje de máquina.

Con base en la gran cantidad de conocimiento multidominio, organizado, validado y en constante crecimiento en la nube LOD de KG, desarrollamos un

proceso automático que, a partir de un objetivo de aprendizaje establecido por un alumno autónomo en lenguaje natural, selecciona documentos web semánticamente pertinentes para ese objetivo y los organiza en una secuencia pedagógica, dada por un conjunto de posibles rutas de exploración de los recursos que respetan la teoría de la elaboración.

Los recursos de aprendizaje se seleccionan de un corpus de documentos, el cual puede ser tan amplio como toda la web o tan limitado como un repositorio de objetos de aprendizaje particular, como videos en la plataforma Coursera MOOC o artículos académicos de la base de datos CORE.

La solución propuesta sigue tres pasos: (1) anotamos semánticamente, es decir, en términos de conceptos y relaciones de un KG referencial, tanto el objetivo de aprendizaje como los recursos del corpus considerado, y con base en esta anotación seleccionamos los recursos más relevantes para el objetivo de aprendizaje previsto; (2) identificamos relaciones de prerrequisitos entre conceptos en el KG de referencia; y (3) a través de la representación semántica de los recursos seleccionados y de las relaciones de prerrequisitos, organizamos estos recursos en una secuencia coherente con la teoría de la elaboración, lo que significa que los recursos que tratan conceptos más básicos se colocan en las secuencias antes de los que requieren conceptos complejos.

La figura 6.3 presenta los tres componentes responsables de estos tres pasos. Cada uno de los componentes de nuestro enfoque ha sido desarrollado y evaluado de forma independiente en trabajos previos de diferentes autores. La representación semántica fue presentada y evaluada en Grévisse *et al.* (2018), Manrique *et al.* (2017) y Manrique y Mariño (2017). La estrategia de identificación de conceptos centrales mediante representaciones semánticas fue evaluada en Manrique *et al.* (2018a). La identificación del concepto de prerrequisito se presenta y evalúa en detalle en Manrique *et al.* (2019a, 2019b). Las estrategias de secuencia se presentan en Manrique (2019). En los siguientes apartados presentamos los aspectos más importantes de cada uno de los componentes.

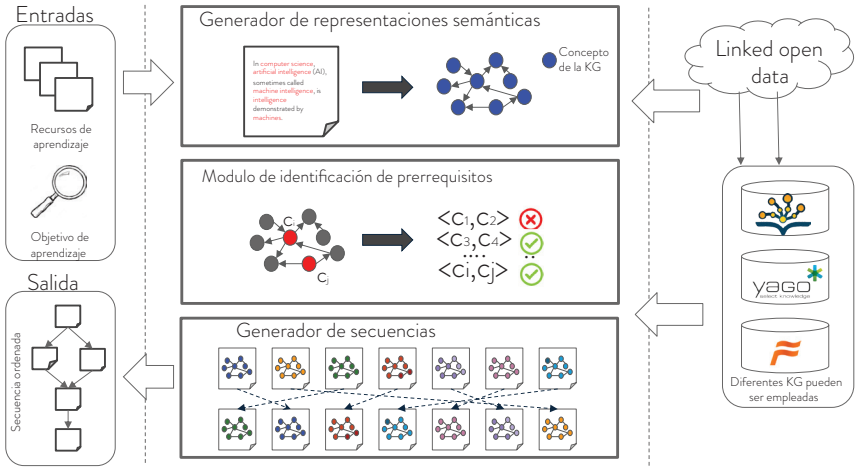


Figura 6.3. Componentes del sistema de apoyo al aprendizaje autónomo
Fuente: elaboración propia.

Construcción de la representación semántica de un documento e identificación de sus conceptos centrales

Nuestro primer componente es el constructor de representaciones semánticas. El objetivo de este componente es identificar los conceptos centrales abordados por cada recurso.

Varios trabajos previos han abordado el desafío de determinar automáticamente ya sea los conceptos principales de un recurso o la importancia de un concepto particular en un recurso (también llamada su *centralidad*) (Farhat *et al.*, 2015; Grévisse *et al.*, 2018; Sultan *et al.*, 2014). Algunos utilizan técnicas de aprendizaje automático basadas en características del texto, mientras que otros aprovechan ontologías de dominio creadas por expertos para guiar la identificación de conceptos. El trabajo propuesto por Krieger *et al.* (2015) es, hasta donde sabemos, el único que utiliza KG de LOD para identificar la relevancia de un recurso en un contexto de aprendizaje; sin embargo, no identifica los conceptos centrales del recurso.

Así, desarrollamos dos caminos diferentes para identificar los conceptos centrales abordados por cada recurso. El primer camino es la construcción de un grafo que represente la semántica del recurso, es decir, una estructura que presente los conceptos abordados, sus relaciones y la importancia de cada concepto. A esta estructura la llamamos *representación semántica del recurso*. El segundo utiliza técnicas de aprendizaje automático enriquecidas con características extraídas de la representación semántica, un recurso para determinar la centralidad de un concepto en el recurso.

Construcción de un grafo de conceptos y relaciones que represente la semántica de un recurso

La representación semántica de un recurso r_i es un grafo dirigido G_i que tiene conceptos como nodos y relaciones entre conceptos como aristas. Tanto los conceptos como las relaciones en la representación corresponden a entidades que se encuentran en el KG referencial. A lo largo del proyecto utilizamos DBpedia como KG. Los nodos y las aristas de la representación semántica del recurso tienen pesos asociados que corresponden a su importancia en el recurso.

El KG consta de un conjunto de tripletas formadas por un identificador (*uniform resource identifier*, URI) de concepto o entidad, un URI de relación o propiedad, y un URI de otro concepto o un valor literal. El módulo de anotación de conceptos busca menciones de conceptos en el texto (anotaciones) y las vincula a conceptos en el KG. Para esta tarea se pueden usar diversos servicios de vinculación de entidades y desambiguación del sentido de las palabras, como DBpedia Spotlight[7], Aylien[8] o Babelfy[9]. El resultado de este paso es una lista de URI con los identificadores de esos conceptos en el KG.

Se utiliza una estrategia de expansión para enriquecer la representación con conceptos que no se mencionan explícitamente en el texto o no fueron identificados por el servicio de anotación. Ampliamos el conjunto de anotaciones (conceptos que se encuentran en el texto), de acuerdo con los enlaces taxonómicos y de propiedad de cada concepto mencionado en el KG. Un vínculo taxonómico conecta un concepto con sus categorías o una categoría con una más amplia, y un vínculo de propiedad conecta dos conceptos que están en una relación, por ejemplo, los conceptos "persona" y "país" a través de la relación "lugar de nacimiento".

La importancia de cada concepto en la representación semántica y, por tanto, en el recurso relacionado se evalúa mediante diferentes funciones de ponderación, que aprovechan la estructura de la representación basada en grafos. Se definieron tres funciones de ponderación: frecuencia de conceptos, puntuación de conectividad semántica y medidas de centralidad. Para cada estrategia de ponderación, se genera una lista ordenada por peso de los conceptos.

Como se hace con los recursos candidatos, todo este proceso también se sigue con el texto que establece el objetivo de aprendizaje previsto, para el cual también se extraen un grafo ponderado y conceptos básicos.

7 Véase https://github.com/dbpedia-spotlight/dbpedia-spotlight

8 Véase https://aylien.com

9 Véase http://babelfy.org

Técnicas de aprendizaje automático basadas en características semánticas para encontrar los conceptos centrales de un recurso

Varias investigaciones sobre la identificación de la importancia de un concepto en un documento utilizan enfoques de aprendizaje automático supervisado (Sultan *et al*., 2014). En estos sistemas, las características se especifican en términos de texto, como la posición relativa de un concepto en una oración o en el título del documento. Al igual que en estas investigaciones, decidimos probar un enfoque de aprendizaje automático supervisado, para identificar la esencia de un concepto en un documento. La principal innovación de nuestro enfoque es que incluimos características relacionadas con la representación semántica de los documentos, extraídas de la representación semántica generada antes. Más precisamente, dada una pareja de recurso de aprendizaje y concepto (c, r_i), asignamos la puntuación de centralidad mediante un modelo de regresión supervisado, al que llamamos $CoCoDisK_{superv}$. Para cada par concepto-recursos de aprendizaje (c, r_i), calculamos cuatro tipos de características: basadas en texto, en el grafo, en similitud semántica y en complejidad. Solo consideramos conceptos en la representación semántica G_i del recurso r_i (Manrique, 2019).

Evaluación

Para evaluar nuestras dos propuestas, un grupo de expertos anotó 192 videoconferencias MOOC en tres idiomas diferentes: inglés, español y francés. El conjunto de datos final contiene 419 anotaciones de conceptos centrales de las videoconferencias. Como líneas de base, utilizamos SWAT[10] y TextRazor[11] , dos sistemas de última generación del área relacionada con el análisis de entidades salientes (Ponza *et al*., 2017). Esta área examina entidades importantes en documentos, donde una entidad es una página de Wikipedia que puede relacionarse fácilmente con un concepto de DBpedia. En general, los enfoques supervisados basados en el aprendizaje automático dan mejores resultados que nuestro enfoque no supervisado: el primer camino de generación de un grafo semántico ponderado; sin embargo, nuestro enfoque supervisado enriquecido con características semánticas supera todas las líneas de base (Manrique, 2019). Vale la pena señalar que todos los sistemas funcionan mejor con recursos en inglés que con recursos en español y francés, probablemente debido al bajo rendimiento de los servicios de etiquetado, como DBpedia Spotlight, en estos idiomas. Por

10 Véase https://services.d4science.org/web/tagme/swat-api

11 Véase https://www.textrazor.com

último, si bien los enfoques supervisados ofrecen mejores resultados que nuestra propuesta no supervisada, este primer enfoque no requiere de un paso de entrenamiento, por lo que podría preferirse cuando el tiempo es un problema o no es factible dicho entrenamiento.

Identificación de relaciones de prerrequisito entre conceptos mediante LOD

La identificación automática de relaciones de prerrequisito entre conceptos se considera una de las piedras angulares de las aplicaciones educativas en línea modernas y a gran escala (Gasparetti *et al*., 2017; Pan *et al*., 2017; Talukdar y Cohen, 2012). Recientemente, ha habido un interés creciente en enfoques automáticos para identificar prerrequisitos (Liang *et al*., 2015; Pan *et al*., 2017), organizar recursos de aprendizaje (Manrique *et al*., 2018b) y generar automáticamente listas de lectura (Fabbri *et al*., 2018). La mayoría de estos enfoques aprovechan las técnicas de procesamiento del lenguaje natural y las estrategias de aprendizaje automático, con el objetivo de extraer conexiones latentes entre conceptos en grandes corpus de documentos para encontrar dependencias de prerrequisitos. A diferencia de los enfoques anteriores, nuestra propuesta utiliza KG abierto como fuente principal para identificar prerrequisitos. Dado un concepto objetivo *c*, se busca identificar sus prerrequisitos en el espacio conceptual del KG. Como un KG puede tener millones de conceptos, un primer paso consiste en recuperar el conjunto de conceptos candidatos que eventualmente podrían ser prerrequisitos de un concepto *c*. Este conjunto se construye, por un lado, al analizar los vínculos jerárquicos e incluir todos los conceptos que comparten una categoría común con *c* y, por otro lado, al integrar todos los conceptos encontrados a través de una ruta no jerárquica de longitud dada.

Este primer conjunto de candidatos se reduce luego mediante un proceso de poda. Nuestra estrategia de poda de conceptos se basa en una medida simple que analiza las referencias entre conceptos. Esta medida, llamada RefD, fue propuesta por Liang *et al*. (2015) y originalmente se calcula para evaluar el grado en que un concepto c_a requiere un concepto c_b como requisito. La noción principal detrás de RefD es que si la mayoría de los conceptos relacionados de c_a se refieren a c_b, pero pocos conceptos relacionados de c_b se refieren a c_a, entonces es más probable que c_b sea un requisito de c_a. En nuestro enfoque, RefD se modificó ligeramente para aplicarlo a KG. La versión modificada de RefD, llamada SemRefD, tiene en cuenta rutas semánticas en el KG (Manrique *et al*., 2019a). Para ser incluido en el conjunto de posibles prerrequisitos, un concepto debe superar un límite configurable (θ) para la función SemRefD.

Una vez que se obtiene un conjunto candidato, se evalúa la relación de prerrequisitos entre todos los posibles pares de conceptos del conjunto candidato final y el concepto objetivo. La evaluación se realiza mediante un modelo supervisado (Manrique *et al*., 2018b).

Evaluación

Para evaluar nuestra propuesta, seguimos todo el proceso con quince conceptos de las áreas de ciencias de la computación y matemáticas, que produjeron 582 parejas con posibles prerrequisitos. Estas parejas se evaluaron con el sistema supervisado, cuyo resultado fue confrontado con el etiquetado manual por parte de expertos. La evaluación, que se detalla en Manrique *et al*. (2019b), mostró que al priorizar los vínculos horizontales directos sobre los jerárquicos, se logra una precisión de entre 83 % y 92,9 %, dependiendo del parámetro configurable del método de poda. Un valor (θ) de 0,2 parece ser apropiado para regular el equilibrio entre la precisión y el número de conceptos previos correctos identificados. Finalmente, vale la pena señalar que el proceso depende en gran medida del KG elegido, en nuestro caso DBpedia. Por lo tanto, los nuevos conceptos con tendencia de mala representación en el grafo podrían dar malos resultados.

Generación automática de secuencias pedagógicamente sólidas de recursos para lograr un objetivo de aprendizaje

Nuestro último módulo es el generador de secuencia. Este orquesta los cálculos de las relaciones de prerrequisitos y las representaciones semánticas de los recursos de aprendizaje, para producir una secuencia pedagógicamente sólida (coherente con la teoría de la elaboración) de recursos de aprendizaje.

El problema de generar una secuencia puede entenderse como producir un orden parcial sobre un conjunto finito de elementos. En nuestro caso, el problema es que el orden se basa en relaciones de prerrequisitos por pares. Esto implica que la directiva de ordenamiento no relaciona de forma directa los elementos (es decir, los recursos de aprendizaje), sino los conceptos en su representación. Además, es posible ampliar la relación de prerrequisitos entre conceptos para que prevalezcan las relaciones entre recursos de aprendizaje. Por lo tanto, las estrategias de secuenciación presentadas en esta sección se basan en relaciones de prerrequisitos entre conceptos y relaciones de precedencia entre recursos.

Así, diseñamos e implementamos tres estrategias diferentes (Manrique, 2019). La primera, llamada *método de clasificación simple* (SRM), se basa en la idea de clasificar cada recurso de aprendizaje r_i. Si el número de recursos de

aprendizaje que abordan conceptos identificados como prerrequisitos de los conceptos que se encuentran en r_i es alto, r_i se clasificará alto para indicar que debe ir al final de la secuencia. Esta estrategia es simple, pero puede resultar costosa, ya que para cada par de recursos, cada par de conceptos básicos se comparan por pares. Para reducir la cantidad de cálculos de prerrequisitos, propusimos una segunda estrategia, el *método bridge ensemble* (BEM). Esta se basa en el supuesto de que cuando se introduce un nuevo concepto en un recurso de aprendizaje, se espera que el recurso mencione sus conceptos previos más importantes. Los conceptos previos son, por tanto, puentes hacia la explicación de conceptos más complejos. Los conceptos puente se introducen en un recurso de aprendizaje anterior, pero reaparecen en un recurso de aprendizaje posterior cuando se introducen algunos conceptos nuevos. Esto implica que los conceptos de puente son los conceptos comunes entre los recursos de aprendizaje.

En la figura 6.4 se presentan tres recursos de aprendizaje (L_1; L_2; L_3), cada uno representado por tres conceptos. Las flechas indican una relación de prerrequisito entre los conceptos (por ejemplo, c_3 es un prerrequisito de c_4 y c_5). Bajo la noción de que para introducir un nuevo concepto se deben mencionar sus conceptos previos, debe haber recursos que compartan conceptos comunes en su representación, conceptos *puente*. En la figura 6.4, c_3 es un concepto puente entre L_1 y L_2. Tenemos el diagrama de Venn entre L_1 y L_2. El análisis de prerrequisitos entre los conceptos puente y no puente muestra una dirección común de la relación de prerrequisitos. En este caso, está claro que es más apropiado asignar a L_1 una clasificación superior que a L_2. Ahora, al considerar el diagrama de Venn entre L_1 y L_3, en el que el concepto puente es c_1, se vuelve a calcular el análisis de prerrequisitos entre los conceptos puente y no puente. A diferencia del caso anterior, no existe una dirección clara de la relación de prerrequisitos, lo que indica que L_1 y L_3 podrían tener clasificaciones similares. Para finalizar, el diagrama de Venn entre L_2 y L_3 muestra que no existen conceptos puente, por lo que su análisis no contribuye realmente al proceso de clasificación. Con esta estrategia, el número de cálculos se reduce de forma considerable.

Nuestra última estrategia se llama *estrategia de descubrimiento de unidades* (SUD). Esta se basa en el hecho de que los cursos suelen estar estructurados en unidades, y los recursos de aprendizaje dentro de una unidad comparten conceptos comunes. Para imitar esta estrategia de organización, diseñamos una estrategia automática que agrupa recursos con representaciones semánticas similares en unidades basadas en un algoritmo de agrupamiento jerárquico aglomerativo (AGC) (Gulagiz, 2017). Tras agrupar los recursos de aprendizaje en unidades, se aplican las estrategias anteriores (SRM o BEM) tanto para ordenar las unidades tratándolas como recursos como para ordenar los recursos dentro de cada unidad.

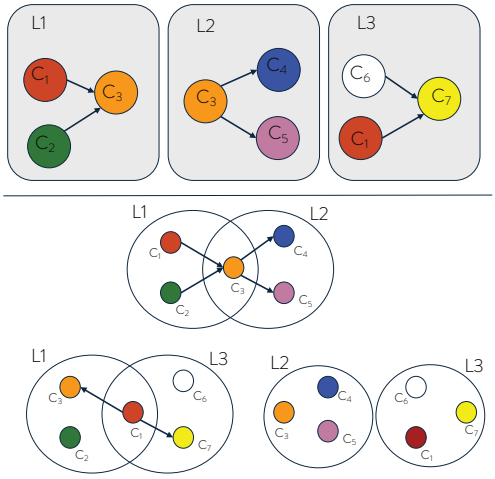


Figura 6.4. Principio del concepto puente
Las flechas indican una relación de prerrequisito entre conceptos. c_3 y c_1 son conceptos puente. No existen conceptos puente entre L_2 y L_3.
Fuente: elaboración propia.

Evaluación

Para evaluar nuestra propuesta, generamos un conjunto de datos con recursos (transcripciones de videoconferencias y documentos de texto) de cincuenta MOOC de la plataforma Coursera de informática (26), ingeniería eléctrica (9), matemáticas (4), física y química (4), y siete cursos de otras áreas diferentes. Para ser incluido, un curso debía tener al menos 50 evaluaciones y un promedio de 4 sobre una escala de 5, según los sitios MOOC-list[12] y Classcentral[13]. La mayoría de los cursos están en inglés, y algunos en francés o español. No se incluyeron los recursos etiquetados como opcionales. El orden de estudio de los recursos de aprendizaje sugerido por el profesor que diseñó el MOOC se utilizó como secuencia correcta para evaluar nuestros resultados. El problema de comparar dos secuencias se conoce en la literatura como *comparación de listas clasificadas*; entre las diferentes métricas empleadas para esta tarea, la distancia de rango de Kendall Tau y la regla de Spearman son estándares actuales (Fagin *et al.*, 2003; Kumar y Vassilvitskii, 2010).

Encontramos cuatro trabajos en la literatura referentes al problema de la secuenciación automática de recursos de aprendizaje (Changuel *et al.*, 2015;

12 Véase https://www.mooc-list.com

13 Véase https://www.classcentral.com/

Gasparetti *et al.*, 2017; Shen *et al.*, 2015; Siehndel *et al.*, 2014). En los dos más cercanos a nuestro trabajo (Changuel *et al.*, 2015; Gasparetti *et al.*, 2017) no se proporcionó el código ni información para replicar el experimento, por lo que optamos por implementar las propuestas de Shen *et al.* (2015) y Siehndel *et al.* (2014) y utilizarlos como líneas de base. En Shen *et al.* (2015), los recursos se representan como una bolsa ponderada de palabras, en la que el peso viene dado por una medida semántica calculada en WordNet. La estrategia propuesta por Shen *et al.* (2015) requiere la anotación de los recursos con conceptos de Wikipedia. Usamos TextRazor como servicio de anotaciones. Cada recurso de aprendizaje está representado por un conjunto de características que se extraen del conjunto respectivo de páginas de conceptos de Wikipedia. En todos los casos, nuestras tres estrategias, incluso en sus peores resultados (peor calibración de los parámetros), superan las líneas de base seleccionadas y la diferencia es estadísticamente significativa ($\rho < 0{,}05$) al utilizar una prueba *t* de dos colas sobre la medida Kendall Tau.

Encontramos que nuestro enfoque es apropiado para la generación de secuencias de aprendizaje. Además, los algoritmos y las estrategias propuestas se pueden utilizar también para crear motores de búsqueda centrados en los procesos de aprendizaje, clasificar recursos y apoyar los procesos de diseño de cursos.

Potencial y límites de un enfoque de web semántica para organizar automáticamente los recursos web en secuencias de aprendizaje para estudiantes autónomos

Los resultados presentados aquí muestran el gran potencial de las aplicaciones de web semántica basadas en KG en la educación superior. El sistema presentado permite entregar el control del aprendizaje al estudiante autónomo, al tiempo que lo guía por una secuencia pedagógicamente sólida de recursos pertinentes. El corpus de donde se obtienen estos recursos puede ser tan amplio como la web misma, lo que permite personalizar aprendizajes con objetivos muy específicos, conocimientos muy recientes o interdisciplinarios, que rara vez se encuentran en cursos tradicionales. Por su parte, los procesos de anotación automática de la semántica y conceptos centrales de un recurso resuelven uno de los problemas históricos de los repositorios de objetos de aprendizaje, tradicionalmente anotados de forma manual: la escalabilidad de dichas anotaciones. Además, nuestras evaluaciones demuestran que la calidad de los resultados con estas técnicas es mayor que la obtenida con técnicas de aprendizaje automático.

No obstante, estas nuevas oportunidades todavía tienen algunas limitaciones. Nuestra evaluación mostró que la calidad de los resultados no era uniforme

en todas las disciplinas e idiomas, siendo mejor en inglés y en cursos como los de matemáticas, en los que la apropiación de algunos conceptos depende en gran medida de conceptos previos. Gracias al crecimiento constante del LOD, podemos esperar que estos problemas se resuelvan en el futuro próximo.

Desde una perspectiva más amplia, una experiencia de aprendizaje es más que una secuencia adecuada de recursos de aprendizaje relevantes. Tener en cuenta el estilo y las preferencias de aprendizaje del alumno, así como estrategias pedagógicas distintas a la teoría de la elaboración, debería ser un desafío para las nuevas investigaciones.

Conclusión

En los últimos diez años, las tecnologías de la web semántica han pasado de investigaciones a aplicaciones reales, y el conocimiento representado en ontologías y KG ha crecido de unos pocos a más de mil conjuntos de datos altamente interconectados, algunos con millones de instancias y cientos de miles de enlaces. Esperamos que la reflexión sobre las oportunidades para la educación superior basada en la web semántica sirva como hoja de ruta para futuras investigaciones y desarrollos en el campo. Nuestras propias innovaciones presentadas son ejemplos concretos de materialización de estas oportunidades (Manrique *et al.*, 2018a, 2018b, 2019a, 2019b).

Referencias

Allemang, D. y Hendler, J. (2011). *Semantic web for the working ontologist* (2.ª ed.). Morgan Kaufmann.

Bejaoui, R., Paquette, G., Basque, J. y Henri, F. (2017). Cadre d'analyse de la personnalisation de l'apprentissage dans les cours en ligne ouverts et massifs (clom). *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, *24* (2), 37-63.

Berners-Lee, T., Hendler, J. y Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Carbonaro, A. (2020). Concept integration to develop next generation of technology- enhanced learning systems. En M. Rehm, J. Saldien y S. Manca (Eds.), *Project and design literacy as cornerstones of smart education* (pp. 121-129). Springer Singapore.

Castells, P. (2003). *La web semántica*. Escuela Politécnica Superior, Universidad Autónoma de Madrid.

Changuel, S., Labroche, N. y Bouchon-Meunier, B. (2015). Resources sequencing using automatic prerequisite-outcome annotation. *ACM Transactions on Intelligent Systems Technology*, *6*(1), 1-30. https://doi.org/10.1145/2505349

Da Costa, A., De Carvalho, C. y Ferreira, D. (2017). *A collaborative learning environment based on the semantic web principles*. International Association for Development of the Information Society.

Davies, J., Fensel, D. y Van Harmelen, F. (2003). *Towards the semantic web: Ontology-driven knowledge management*. Wiley. https://books.google.com.co/books?id=kREOBAAAQBAJ

Denaux, R., Aroyo, L. y Dimitrova, V. (2005). An approach for ontology-based elicitation of user models to enable personalization on the semantic web. En *Special interest tracks and posters of the 14th international conference on world wide web* (pp. 1170-1171). Association for Computing Machinery. http://doi.acm.org/10.1145/1062745.1062923

Devedzic, V. (2004). Education and the semantic web. *International Journal of Artificial Intelligence in Education*, *14*(2), 165-191.

Dietze, S., Sánchez-Alonso, S., Ebner, H., Yu, H., Giordano, D., Marenzi, I. y Nunes, B. (2013). Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program*, *47*(1), 60-91. https://doi.org/10.1108/00330331211296312

Dolog, P., Henze, N., Nejdl, W. y Sintek, M. (2003). Towards the adaptive semantic web. En F. Bry, N. Henze y J. Maluszynski (Eds.), *Principles and practice of semantic web reasoning* (pp. 51-68). Springer Berlin Heidelberg.

Domingue, J., Fensel, D. y Hendler, J. A. (Eds.). (2011). *Handbook of semantic web technologies*. Springer. https://doi.org/10.1007/978-3-540- 92913-0

Fabbri, A. R., Li, I., Trairatvorakul, P., He, Y., Ting, W.T., Tung, R., Westerfield, C. y Radev, D. R. (2018). Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 611-620). Association for Computational Linguistics.

Fagin, R., Kumar, R. y Sivakumar, D. (2003). Comparing top k lists. En *Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 28-36). Society for Industrial and Applied Mathematics. http://dl.acm.org/citation.cfm?id=644108.644113

Farhat, R., Jebali, B. y Jemni, M. (2015). Ontology based semantic metadata extraction system for learning objects. En G. Chen, V. Kumar, Kinshuk,

R. Huang y S.C. Kong (Eds.), *Emerging issues in smart learning* (pp. 247-250). Springer Berlin Heidelberg.

Figueroa, C., Vagliano, I., Rocha, O. y Morisio, M. (2015). A systematic literature review of linked data-based recommender systems. *Concurrency Computation*, *27*(17), 4659-4684. https://doi.org/10.1002/cpe.3449

García, I., Benavides, C., Alaiz Moreton, H. y Alonso, A. (2013, 08). A study of the use of ontologies for building computeraided control engineering self-learning educational software. *Journal of Science Education and Technology*, *22*. https://doi.org/10.1007/s10956-012-9416-6

Gasparetti, F., Medio, C. D., Limongelli, C., Sciarrone, F. y Temperini, M. (2017). Prerequisites between learning objects: Automatic extraction based on a machine learning approach. *Telematics and Informatics*, *35*(3), 595-610.

Ghosh, S., Rath, M. y Shah, C. (2018). Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. En *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 22-31). Association for Computing Machinery. http://doi.acm.org/10.1145/3176349.3176386

Grévisse, C., Manrique, R., Mariño, O. y Rothkugel, S. (2018). Knowledge graph-based teacher support for learning material authoring. En J. E. Serrano C. y J. C. Martínez-Santos (Eds.), *Advances in Computing. CCC 2018* (pp. 177-191). Springer International Publishing.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/knac.1993.1008

Gulagiz, F. (2017). Comparison of hierarchical and non-hierarchical clustering algorithms. *International Journal of Computer Engineering and Information Technology*, *9*(1), 6-14.

Heath, T. y Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1-136. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Helen, M. (2014). *Personalized recommender systems for resource-based learning hybrid graph-based recommender systems for folksonomies* [tesis doctoral]. Technischen Universitat Darmstadt.

Hill, J. R. (2012). Resource-based learning. En N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 2850-2852). Springer. https://doi.org/10.1007/978-1-4419-1428-6

International Organization for Standardization (ISO) e International Electrotechnical Commission (IEC). (2013). *ISO-IEC 19788. Information technology. Learning, education and training. Metadata for learning resources multipart standard* [computer software manual]. ISO-IEC.

Jevsikova, T., Berniukevicius, A. y Kurilovas, E. (2017). Application of resource description framework to personalise learning: Systematic review and methodology. *Informatics in Education*, *16*(1), 61-82.

Krieger, K., Schneider, J., Nywelt, C. y Rösner, D. (2015). Creating semantic fingerprints for web documents. En *Proceedings of the 5th international conference on web intelligence, mining and semantics* (pp. 1-6). Association for Computing Machinery.

Kumar, R. y Vassilvitskii, S. (2010). Generalized distances between rankings. En *Proceedings of the 19th international conference on world wide web* (pp. 571-580). Association for Computing Machinery.

Kurt, A. y Gursel, B. (2018). Analysis of students' online information searching strategies, exposure to internet information pollution and cognitive absorption levels based on various variables. *Malaysian Online Journal of Educational Technology*, *6*, 18-29.

Le Moigne, J.-L. (2001). Pourquoi je suis un constructiviste non repentant. *Revue du MAUSS*, *17*(1), 197-223. https://doi.org/10.3917/rdm.017.0197

Liang, C., Wu, Z., Huang, W. y Giles, C. L. (2015). Measuring prerequisite relations among concepts. En *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1668-1674). Association for Computational Linguistics.

Manrique, R. (2019). *Towards automatic learning resources organization via knowledge graphs* [tesis doctoral]. Universidad de los Andes.

Manrique, R., Grévisse, C., Mariño, O. y Rothkugel, S. (2018a). Knowledge graph-based core concept identification in learning resources. En R. Ichise, F. Lecue, T. Kawamura, D. Zhao, S. Muggleton y K. Kozaki (Eds.), *Semantic technology: 8th Joint International Conference, JIST 2018* (pp. 36-51). Springer.

Manrique, R., Herazo, O. y Mariño, O. (2017). Exploring the use of linked open data for user research interest modeling. En A. Solano y H. Ordoñez (Eds.), *Advances in computing* (pp. 3-16). Springer International Publishing.

Manrique, R. y Mariño, O. (2017). How does the size of a document affect linked open data user modeling strategies? En *Proceedings of the international conference on web intelligence* (pp. 1246-1252). Association for Computing Machinery. http://doi.acm.org/10.1145/3106426.3109440

Manrique, R., Pereira, B. y Mariño, O. (2019a). Exploring knowledge graphs for the identification of concept prerequisites. *Smart Learning Environments*, *6*(1). https://doi.org/10.1186/s40561-019-0104-3

Manrique, R., Pereira, B., Mariño, O., Cardozo, N. y Wolfgand, S. (2019b). Towards the identification of concept prerequisites via knowledge graphs. En *IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (pp. 332-336). Institute of Electrical and Electronics Engineers (IEEE).

Manrique, R., Sosa, J., Marino, O., Nunes, B.P. y Cardozo, N. (2018b). Investigating learning resources precedence relations via concept prerequisite learning. En *IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 198- 205). Institute of Electrical and Electronics Engineers (IEEE).

Moraes, F., Putra, S. R. y Hauff, C. (2018). Contrasting search as a learning activity with instructor-designed learning. En *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 167-176). Association for Computing Machinery. http://doi.acm.org/10.1145/3269206.3271676

Musto, C., Lops, P., De Gemmis, M. y Semeraro, G. (2017). Semantics-aware recommender systems exploiting linked open data and graph-based features. *Knowledge-Based Systems*, *136*. https://doi.org/https://doi.org/10.1016/j.knosys.2017.08.015

Nadzir, M. (2015). Identifying the information-seeking behaviours among school of computing undergraduate students. *Journal of Theoretical and Applied Information Technology*, *74*(2), 149-154.

Newell, A. y Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Nikolopoulou, K. y Gialamas, V. (2011). Undergraduate students' information search practices. *Themes in Science and Technology Education*, *4*(1), 21-32.

Ontotext. (2017). *What are linked data and linked open data?* https://ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

Pan, L., Li, C., Li, J. y Tang, J. (2017). Prerequisite relation learning for concepts in moocs. En *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1447-1456). Association for Computational Linguistics.

Paquette, G. (2010). Ontology-based educational modelling making IMS-LD visual. *Technology, Instruction, Cognition and Learning*, *7*(3-4), 263-286.

Paquette, G., Lundgren-Cayrol, K., Miara, A. y Guérette, L. (2004). The Explora-2 learning object manager. En R. McGreal (Ed.), *Online education using learning objects* (pp. 254-268). Routledge/Falmer.

Paquette, G., Rogozan, D. y Mariño, O. (2012). Competency comparison relations for recommendation in technology enhanced learning scenarios. https://cinfonia.uniandes.edu.co/publications/competency-comparison-relations-for-recommendation-in-technology-enhanced-learning-scenarios/

Paquette, G., Rosca, I., Mihaila, S. y Masmoudi, A. (2007). Telos: A service-oriented framework to support learning and knowledge management. En S. Pierre (Ed.), *E-learning networked environments and architectures: A knowledge processing perspective* (pp. 79-109). Springer London.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, *8*(3), 489-508. https://doi.org/10.3233/ SW-160218

Piaget, J. (1936). *Le probléme biologique de l'intelligence*. https://www.fondationjeanpiaget.ch/fjp/site/textes/VE/JP36_NdI_avpropos_intro.pdf

Ponza, M., Ferragina, P. y Piccinno, F. (2017). Document aboutness via sophisticated syntactic and semantic features. En F. Frasincar, A. Ittoo, L. M. Nguyen y E. Métais (Eds.), *Natural language processing and information systems* (pp. 441-453). Springer International Publishing.

Quezada-Sarmiento, P. A., Enciso, L., Conde, L., Mayorga-Díaz, M. P., Guaigua-Vizcaino, M. E., Hernández, W. y Washizaki, H. (2020). Body of knowledge model and linked data applied in development of higher education curriculum. En K. Arai y S. Kapoor (Eds.), *Advances in computer vision* (pp. 758-773). Springer International Publishing.

Reigeluth, C. M. y Darwazeh, A. (1982). The elaboration theory's procedure for designing instruction. *Journal of instructional development*, *5*(3), 22-32. https://doi.org/10.1007/BF02905492

Savard, I., Bourdeau, J. y Paquette, G. (2013). An ontology and a method to support instructional design integrating cultural variables. En *Proceedings of the Workshop on Culturally-aware Technology-Enhanced Learning (Cultel).*

Scholl, P., Mann, D., Rensing, C. y Steinmetz, R. (2007). *Support of acquisition and organization of knowledge artifacts in informal learning contexts*. European Distance and E-Learning Network (EDEN).

Shen, S., Lee, H., Li, S., Zue, V. y Lee, L. (2015). Structuring lectures in massive open online courses (MOOCS) for efficient learning by linking similar sections and predicting prerequisites. En *Interspeech 2015, 16th Annual Conference of the International Speech Communication Association* (pp. 1363-1367). International Speech Communication Association (ISCA).

Siehndel, P., Kawase, R., Nunes, B. P. y Herder, E. (2014). Towards automatic building of learning pathways. En *Proceedings of the 10th international conference on web information systems and technologies* (pp. 270-277). Springer.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, *24*, 86-97.

Sultan, M. A., Bethard, S. y Sumner, T. (2014). Towards automatic identification of core concepts in educational resources. En *Proceedings of the 14th ACM/IEEE-CS joint conference on digital libraries* (pp. 379-388). IEEE Press.

Talukdar, P. P. y Cohen, W. W. (2012). Crowdsourced comprehension: Predicting prerequisite structure in Wikipedia. En *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 307-315). Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2390384.2390423

Tapia-León, M., Aveiga, C., Chicaiza, J. y Suárez-Figueroa, M.C. (2019). Ontological model for the semantic description of syllabuses. En *Proceedings of the 9th international conference on information communication and management* (pp. 175-180). Association for Computing Machinery. http://doi.acm.org/10.1145/3357419.3357442

Tetlow, P., Pan, J.Z., Oberle, D., Wallace, E., Uschold, M. y Kendall, E. (2006). *Ontology driven architectures and potential uses of the semantic web in systems and software engineering*. W3C. http://www.w3.org/2001/sw/BestPractices/SE/ODA/

Tibau, M., Siqueira, S., Pereira Nunes, B., Bortoluzzi, M., Marenzi, I. y Kemkes, P. (2018). Investigating users' decision-making process while searching online and their shortcuts towards understanding. En G. Hancke, M. Spaniol, K. Osathanunkul, S. Unankard y R. Klamma (Eds.), *Advances in web-based learningicwl 2018* (pp. 54-64). Springer International Publishing.

Tiropanis, T., Davis, H., Millard, D. y Weal, M. (2009). Semantic technologies for learning and teaching in the web 2.0 era. *IEEE Intelligent Systems*, *24*(6), 49-53. https://doi.org/10.1109/MIS.2009.121

Vygotsky, L., Cole, M., John-Steiner, V., Scribner, S. y Souberman, E. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Morgan Kaufmann Publishers Inc.

Anexo

Escritorio TELOS con tres herramientas importantes abiertas:

(Escanée el código para ver la figura o visite https://gobierno.uniandes.edu.co/wp-content/uploads/Uniandes-Fig-A.1-1.png)

Escenario socioconstructivista en TELOS:

(Escanée el código para ver la figura o visite https://gobierno.uniandes.edu.co/wp-content/uploads/Uniandes-Fig-A.2-1.png)

ROLES Y EFECTOS DE LA GEOAI EN EL ENTENDIMIENTO DE TERRITORIOS Y COMUNIDADES DEL SUR

Ana María Bustamante Duarte, Diego Pajarito Grajales, Manuel Portela, Leonardo Parra Agudelo

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.07

Introducción

La adopción de la inteligencia artificial (IA) ha crecido deprisa en los últimos años en casi todos los campos de la ciencia y las humanidades. Su uso se ha ampliado para intentar responder problemas complejos, como el cambio climático (Huntingford *et al*., 2019), la migración (Madianou *et al*., 2020), el cuidado de la salud (Kitchin, 2020) o la planificación urbana (Taylor *et al*., 2016); problemas profundamente entrelazados con aspectos geoespaciales por su relación con los territorios, las comunidades, las instituciones (De Albuquerque *et al*., 2016) y todo aquello *más que humano* (Puig de la Bellacasa, 2017). Sin embargo, estos retos están siendo abordados por geografías producidas a través de y por lo digital (Ash *et al*., 2018; Datta, 2024), incluidas técnicas como el aprendizaje automático y la IA. Esto se ha denominado GeoAI.

En la actualidad, el desarrollo de la GeoAI se encuentra limitado a las ciencias de la computación, las ciencias de la geoinformación y la geoinformática (Mortaheb y Jankowski, 2023; Hu *et al*., 2019; Janowicz *et al*., 2020). Como resultado, la comprensión y la reflexión crítica y transdisciplinar (así como sus posibles retos e impactos) del componente geoespacial se vuelve primordial para varios de estos desarrollos y aplicaciones de IA. Como explicitan Janowicz *et al*. (2020), la práctica y el uso de IA en el ámbito geoespacial no es nuevo y se remonta a algunas décadas.

Hoy en día, a la GeoAI se le reconocen potenciales usos para mejorar la eficiencia de los servicios urbanos y la calidad de vida de los residentes, al igual que para responder a los retos sociales, económicos y ecológicos de la producción de datos, conocimientos e información geoespacial sobre las dinámicas

humanas que ocurren en los territorios (Mortaheb y Jankowski, 2022). No obstante, las tecnologías y sus prácticas son las que actualmente están marcando las pautas que definen las políticas públicas y las regulaciones. Nuestra perspectiva en este capítulo entiende la GeoAI no solo como una solución tecnológica a los problemas geoespaciales, sino como una sociotecnología que aún es emergente, está fragmentada en múltiples prácticas a diversos niveles (Wynne, 1988) e impacta múltiples aspectos.

En este contexto, es clave recalcar que los problemas actuales que han sido investigados en la IA también se encuentran en la GeoAI. Por ejemplo, la reproducción de sesgos, ya sean humanos o producto de la automatización, los cuales están arraigados en los parámetros y datos utilizados para el entrenamiento de los sistemas de IA durante sus fases de desarrollo y que se revelan durante su implementación. Algunos casos en los que se refleja esto presentan sesgos embebidos en los sistemas de clasificación y las categorizaciones que se generan en estos sistemas, que pueden reforzar la discriminación existente en la sociedad y aumentar las inequidades sociales (Ferrer *et al.*, 2021; Buolamwini y Gebru, 2018). Este fenómeno es complejo, ya que clasificaciones y categorías responden a una visión concreta, donde se cifran y se normalizan las experiencias ocurridas en las comunidades y sus territorios, por lo general marginalizadas (Johnson, 2014).

Desde los aspectos geográficos también se incorporan sesgos adicionales a los ya presentes en la IA. Explícitos o implícitos (Ferrer *et al.*, 2021), algunos de los sesgos están asociados al uso de datos de entrenamiento de geografías del norte global, con representaciones excesivas de espacios o elementos geográficos irrelevantes para la mayoría del planeta, o con una limitada o inexistente representatividad de espacios geográficos de países del sur; estos se pueden amplificar con la agregación de diversos tipos de datos. Las huellas digitales de las geografías que se identifican o se generan a través del uso de estos sistemas de GeoAI conducen a tratamientos diferenciales, en muchos casos discriminatorios. Esto ocurre no solo en el espacio físico, sino también en el digital, lo que produce, potencialmente, *periferias informacionales* (Datta, 2024) u otro tipo de espacios y dinámicas de reconocimiento, inventario y control territorial, en su mayoría para fines militares y político-económicos, o de usos para hipervigilancia y ordenamiento (Anderson, 1991; Ferrer *et al.*, 2021) que podrían llegar a ser discriminatorios.

Sin embargo, al mismo tiempo, aquello que no es documentado en geovisualizaciones suele ser invisibilizado al no formar parte de las narrativas o discursos construidos a partir de las representaciones que ofrecen estos sistemas. Esta dualidad *visible-invisible* ha llevado a conflictos sociales que han sentado

la base para el reconocimiento de otras formas de entender y representar el mundo, como los conocimientos y las prácticas de tipo ancestrales, emergentes, orgánicas y situadas de contextos específicos. Estas formas usan herramientas y mecanismos diversos de georeferenciación o representación cartográfica participativos o directamente liderados por las comunidades, para hacer sentido de diferentes realidades y cuestionar el uso estándar y priorizado que se le ha dado a las formas de representación euclidianas y positivistas que responden a dinámicas específicas de poder.

En este capítulo consideramos críticamente los datos geoespaciales explícitos[1], sus categorías generadas desde aproximaciones institucionales y su representación en términos físicos en los mapas. Estos datos e información, aunque tienen diversos beneficios en términos, por ejemplo, de apoyos a soluciones técnicas, no necesariamente corresponden o estiman la totalidad de las dinámicas socioespaciales de algunas comunidades precarizadas. De esta manera, esperamos complejizar la lectura sobre el aspecto geoespacial en las GeoAI implementadas en los territorios del sur, a partir de la reflexión sobre proyectos que utilizan estos sistemas y tienen aproximaciones más participativas. Buscamos discutir alrededor de estos casos, enfocados en el uso de datos geoespaciales implícitos y participativos como una de las formas de responder a los retos del uso actual de la GeoAI. Por último, cuestionamos si es necesario representar geográficamente todos los aspectos de estos territorios/comunidades para plantear políticas públicas, estrategias y acciones para su mejoramiento.

Sobre clasificación y categorías en la GeoAI

La construcción de categorías como manera de ver nuestras realidades, en general, hace que las definamos como estructuras socioculturales que responden a condiciones espaciotemporales específicas. El proceso epistemológico de organizar el conocimiento, lo conocido, lo no conocido, en esquemas de clasificación corresponde a formas de estructurar la realidad para diseñar, conducir y evaluar cómo esta opera basada en valores, presuntamente comunes entre las

1 Por *datos geoespaciales explícitos* entendemos, por ejemplo, datos de localización, imágenes satelitales georeferenciadas, algunos datos cualitativos de percepción, etc., pero que tienen orígenes institucionales o de *arriba hacia abajo*; por otro lado, los *datos implícitos* son aquellos que cuantifican o cualifican descripciones geoespaciales (p. ej., localizaciones de espacios o prácticas no institucionalizados/oficiales, narrar espacios en texto, audio o dibujos) y que en ocasiones no utilizan datos con referencias geográficas directas.

comunidades, las personas expertas o aquellas provenientes de otros lugares de poder (Alaimo y Kallinikos, 2021).

Así, las categorías se perciben como formas de encontrar similitudes entre agentes, eventos o acciones que no son idénticos, para ser clasificados como elementos equivalentes y agrupados basados en estas equivalencias (Rosch, 1975). Adicional a esto, las clasificaciones y las categorías que acompasan a otras (p. ej., sistema de jerarquías) referencian el mundo que se captura, analiza y representa a través de estos sistemas y definen el balance entre los niveles de abstracción y de detalle que son necesarios para operar cada uno (Alaimo y Kallinikos, 2021). Por ejemplo, los datos deberían representar las características principales de los agentes, pero con un nivel de abstracción suficiente para no sobrecargar el sistema, sin dejar de representar aspectos que son vitales para el entendimiento de las realidades, sobre todo si son sistemas que ayudan a darles forma a estas últimas (Alaimo y Kallinikos, 2021).

La mayoría de estos sistemas usan categorías normativas que pueden llevar a visiones parciales de los datos. Son empleados con frecuencia por las instituciones y sus sistemas, o por grupos socioeconómicamente dominantes, y reflejan la normalización de sus juicios, lo que revela conformidad con un fenómeno conocido como *estándar de normalidad* (Johnson, 2014).

Además, si los datos son agregados, como ocurre en la mayoría de los procesos actuales relacionados con GeoAI, hay varios elementos positivos y negativos que se amplifican con respecto a los grupos sujetos del sistema. Una de esas amplificaciones está relacionada con los sesgos, lo que causa una mayor discriminación de grupos sociales ya marginalizados (Zehlike *et al*., 2017, Buolamwini y Gebru, 2018, Ferrer *et al*., 2021). Potencialmente, esta situación socava las bases de la equidad y los valores democráticos que forman parte de las sociedades actuales o va en contra de otras formas de organización, como las ancestrales. Los dos puntos anteriores cobran aún más importancia en sistemas de GeoAI, o de geotecnología en general, relacionados con el desarrollo de políticas públicas y el entendimiento de los territorios.

Por ejemplo, el uso de plataformas digitales como Google Earth y Google Maps reproducen maneras de ver los territorios (Farman, 2010) que corresponden al *statu quo* del momento o a lo que es normativo; sin embargo, han abierto el acceso al procesamiento de la información cartográfica sobre diferentes territorios, comunidades y dinámicas sociales. Por ejemplo, en el caso palestino, Google Maps reproduce formas de ver el territorio donde se revela la ausencia de datos geográficos relevantes para sus habitantes y solo incluye algunos nombres de calles, de negocios autorreportados e hitos urbanos limitados (Goodfriend, 2021). Los datos geoespaciales y los sistemas de clasificación aquí usados,

intencionalmente, reproducen dinámicas de ideales sociopolíticos y de hipervigilancia de otros países, y muestran sus maneras de ver y controlar este territorio. Este fenómeno sobre lo que es representado —(hiper/in)visibilizado—[2] se complejiza cuando los datos y los sistemas de clasificación y representación deben ser entendidos, hasta un nivel funcional, por las personas que los usan o son sujetos de estas tecnologías.

La falta de claridad y, en algunos casos, transparencia en el funcionamiento de los sistemas limita su entendimiento funcional e impone agendas externas sobre los territorios y poblaciones, lo que restringe las posibilidades para que las personas puedan desarrollar estrategias de contestación.

La GeoAI en el sur global

El uso de plataformas y herramientas de IA para ofrecer servicios digitales son el nuevo estándar en la gran mayoría del mundo. Estos servicios se han vuelto clave para el desarrollo de infraestructura urbana digital, así como para el procesamiento, acceso y circulación de datos e información a diversas escalas territoriales en un gran número de países (Gutiérrez y Muñoz-Cadena, 2023), incluidos varios en Latinoamérica. Esto ha llevado a impulsar fuertemente iniciativas de digitalización y, en general, estrategias de datificación para su funcionamiento (Schou y Hjelhot, 2019; Broomfield y Reuter, 2022; Datta, 2024). Queremos resaltar que, en la mayoría de los casos, los procesos de datificación son esenciales para el desarrollo e implementación de las tecnologías digitales propuestas (plataformas, servicios, etc.); de este modo, en este capítulo vamos a hablar de *digitalización/datificación*, dado que un proceso suele conllevar el otro y hoy son prácticamente inseparables.

El uso de la digitalización/datificación está motivado en especial por la idea de que promueve el desarrollo económico y formas de gobernanza más eficientes, seguras y democráticas (Ricaurte *et al.*, 2024). Así mismo, quienes apoyan su uso argumentan, con poca reflexión crítica, que estos procesos permiten asignar responsabilidades a diferentes agentes de los servicios públicos y contribuyen a la democratización de estos servicios, a través de la reducción de los conflictos que aparecen en algunas de las fases del procesamiento de datos

2 La *(hiper/in)visibilización* es un concepto formulado desde la justicia de datos. Plantea la discusión entre legibilidad de personas, grupos o fenómenos y el derecho a la justa representación en los datos, aunque esto no permita el accionar sobre aspectos clave o, por el contrario, exponga a las personas a agendas de explotación (véase Jameson *et al.*, 2019; Heeks *et al.*, 2020; Behrendt y Sheller, 2023).

e información (Hoefsloot y Gateri, 2024). Además, en términos de gestión de territorios, se enfatiza en su uso por su amplia cobertura, su estandarización y su reporte en tiempo real, al igual que por la automatización de ciertos procesos de decisión, que apoyan la presunta eficiencia y eficacia de estos sistemas dentro de los procesos de gobernanza.

En parte, debido a lo anterior, estos sistemas de GeoAI han sido usados en las ciudades latinoamericanas en la gestión de referencias geográficas de áreas de asentamientos precarizados (Abascal *et al*., 2022; Thomson *et al*., 2020), la coordinación de la movilidad (flujos de tráfico tanto de transporte público como privado), el manejo de agua y su infraestructura asociada (Hoefsloot *et al*., 2023), la gestión de la seguridad y el acceso y provisión de servicios sociales. Así, el énfasis en la digitalización/datificación de los territorios y sus comunidades toma fuerza al seguir la premisa de informar y guiar las decisiones y acciones de los Gobiernos (Pelizza y Kuhlmann, 2017; Kitchin *et al*., 2016), lo que transforma sus modelos y formas de gobernanza y los mueve hacia la gubernamentalidad algorítmica (Ricaurte *et al*., 2024). Sin embargo, algunas reflexiones sobre estas tecnologías también apuntan a que fortalecen la función histórica de la cartografía, como una supuesta herramienta que mejora la legibilidad de aquello que los Estados, sobre todo occidentales, consideran suyos, pero que les permite distinguirse de la otredad (Oluoch *et al*., 2022). Otras reflexiones cuestionan aspectos de justicia en estos sistemas de GeoAI asociados a la (hiper/in)visibilidad de las personas, así como la agencia para relacionarse con las tecnologías y los datos de estas de las comunidades sujeto de estas y para contrarrestar los elementos de sesgos y discriminación que generan (Taylor, 2017).

Weiskopf y Hansen (2022) definen que la gubernamentalidad se enfoca en las prácticas de gobernanza de y para las personas que no están digitalizadas en los registros clásicos de modelos políticos y económicos liberales y neoliberales. De esta manera, la gubernamentalidad algorítmica se enfoca en las prácticas contemporáneas en las que se usan tecnologías digitales y, en muchos casos, automatizadas, las cuales procesan datos digitales de las personas que habitan los territorios donde estos modelos de gubernamentalidad son instaurados (Weiskopf y Hansen, 2022; Barry, 2019). Por lo general, estos datos salen de las actividades "en línea" de la población residente, lo que lleva a formas de *data gaze* y permite el perfilamiento continuo (Beer, 2019) y el análisis predictivo de sus futuras posibilidades, (re)acciones, etc. (Flyverbom y Gartsten, 2022).

Este tipo de prácticas de gubernamentalidad algorítmica tiene grandes retos en los países del sur, incluidos países latinoamericanos, debido al reforzamiento de las dinámicas sociopolíticas y económicas de valores y prácticas provenientes del norte (Datta, 2024; Ricaurte, 2023; Ricaurte *et al*., 2024) y el afianzamiento

de aproximaciones tecnosolucionistas, que supuestamente resuelven problemas complejos de nuestras sociedades latinoamericanas (Ricaurte *et al*., 2024). Así mismo, Ricaurte *et al.* (2024) resaltan que estas prácticas llevan también a la implantación de dinámicas neocoloniales de poder, en tanto se identifica su rol en la (re)afirmación de poder de algunos países del norte global sobre otros y su utilización como medios de los Gobiernos para automatizar asimetrías sociales y de control social. Esto se agrava cuando en esta realidad la gestión de infraestructuras y servicios digitales relevantes, sobre todo urbanas, se encuentra en su mayoría controlada por el Estado o empresas (Hoefsloot y Gateri, 2024; Carr y Hesse, 2023).

En este escenario, los procesos están manejados y, en muchos casos, dirigidos por infraestructuras, servicios y plataformas digitales, donde las automatizaciones de las categorizaciones y de los mecanismos de entrada y salida de datos dan forma continuamente al mundo que conocemos, determinando los componentes humanos, geoespaciales y más que humanos, que son (hiper/in)visibles (Taylor, 2017; Hoefsloot *et al*., 2022), al igual que cómo se ven, gestionan y controlan. Estos sistemas de clasificación y sus categorías afectan lo que se representa, analiza, reporta y decide, a partir de las implementaciones más frecuentes de la GeoAI, relacionadas con temas como la georeferenciación de (1) edificios e infraestructura, (2) actividades y aspectos sociales y humanos, (3) análisis de usos de tierra y (4) manejo del ambiente y desastres (Song *et al*., 2023; Sawhney, 2023).

En este sentido, se plantean cuestionamientos críticos desde la justicia socioespacial y de datos relacionados con el rol que estas tecnologías podrían tener como herramientas de respuesta a crisis socioespaciales, que en algunos casos han sido atendidas en los territorios y sus comunidades, pero que en otros han sido amplificadas a través del uso de tecnologías.

Exploraciones de la GeoAI en el sur: de la implementación y sus pasos previos

En este capítulo consideramos algunos aspectos de GeoAI que son críticos y los examinamos desde casos de gestión territorial basados en experiencias de comunidades urbanas y rurales, donde se vuelve clave la discusión sobre la gestión y los efectos de la información que se recolecta y visualiza de estas comunidades en los sistemas digitales geoespaciales. Los casos exploran aproximaciones de GeoAI que van desde la georeferenciación automatizada de asentamientos, hasta la discusión sobre categorías y sistemas de clasificación en los datos que se utilizan para entrenar estos sistemas para el uso público.

Caso 1. Proyecto de georeferenciación automática de asentamientos precarizados

Contexto

Entre los múltiples intentos por generar datos que describan mejor las condiciones de vida en las ciudades, analizamos un proyecto de georeferenciación automática de asentamientos precarizados, en el cual uno de los autores del capítulo forma parte del equipo. Este proyecto, basado en África occidental y África oriental (Lagos y Kano, en Nigeria, y Nairobi, en Kenia), se ha puesto como objetivo establecer una metodología dialógica que permita la generación y actualización periódica de dichos datos. Para ello, el sistema integra las cuatro principales alternativas de producción de datos[3] y las combina con datos geoespaciales originados en métodos más manuales y humanos (levantamientos de campo, cartografía, encuestas periódicas y censos) y aquellos que resultan de la producción automatizada de datos (p. ej., modelos computacionales de datos sintéticos, procesamiento de imágenes de sensores remotos y técnicas basadas en aprendizaje de máquina) (Thomson *et al*., 2020).

Dicha combinación de métodos se justifica debido a las limitaciones de las cuatro técnicas iniciales, como los altos requerimientos técnicos que dificultan su realización de manera periódica; los costos elevados de implementación (equipos y salarios); la baja velocidad en la producción de datos, usualmente relacionada con las limitaciones de acceso por condiciones geográficas a zonas remotas o poblaciones alejadas de los centros poblados principales; y las producciones automatizadas de sesgos y errores que deben corregirse con visitas a campo sin validación (Thomson *et al*., 2020). Este proyecto es, en parte, el resultado de los más de veinte años que lleva la georeferenciación de áreas precarizadas utilizando datos recolectados a partir de sensores de muy alta resolución, como imágenes satelitales y drones, mediante diferentes técnicas como el análisis basado en objetos (OBIA, por sus siglas en inglés) y el aprendizaje automático (Kuffer *et al*., 2016), que han motivado el desarrollo de algunas técnicas más contemporáneas dentro de las GeoAI, como el uso de redes neuronales (Abascal *et al*., 2022).

La aproximación dialógica para las herramientas de GeoAI aquí discutidas involucra desde el inicio a todas las partes interesadas, incluidas comunidades,

3 Las cuatro técnicas principales de generación de datos de asentamientos precarizados son: (1) enumeración o censo de hogares en los asentamientos; (2) interpretación humana/visual de imágenes satelitales; (3) clasificación de imágenes satelitales basada en aprendizaje de máquina e IA; y (4) cartografía tradicional basada en trabajo de campo (Thomson *et al*., 2020).

autoridades e investigadores, para que establezcan conjuntamente las características de recolección y representación, así como los usos de los datos que se van a generar, para lo cual la reflexión colectiva sobre las categorías de los datos y la información es esencial. Así, se promueve que la representación cartográfica generada y sus herramientas relacionadas de GeoAI, afinadas y contextualizadas al caso, sean relevantes para las partes y se basen en las realidades del territorio y las necesidades de las comunidades que lo habitan, y no solo en las maneras de verlas de *arriba hacia abajo* que tienen las instituciones desde sus datos y sus formas de usarlos. El sistema tiene varios ciclos de cocreación entre distintos agentes. En un primer ciclo de cocreación se priorizan datos relevantes para la comunidad (formatos, tipos, categorías, viabilidad de generación de estos, etc.). En el segundo ciclo se revisan los resultados del mapeo de datos secundarios y terciarios, al igual que su calidad y representatividad. Estos comentarios y modificaciones son implementados, y aquellos que acarrean impactos negativos son analizados y se desarrollan estrategias para contrarrestarlos, minimizarlos o simplemente omitirlos, para así proceder a desarrollar la generalización del proceso y hacerlo escalable, reproducible, entendible y transferible a otras realidades que compartan similitudes contextuales.

La combinación de estos métodos busca, por un lado, proponer estrategias para responder a diferentes limitaciones que estas aproximaciones de manera aislada presentan, como inconsistencias en la referenciación cartográfica, indicadores complejos, restricciones en las características de las variables, altos costos de implementación y mantenimiento, entre otros. Por otro lado, su integración en una aproximación más holística intenta que las representaciones cartográficas explícitas, a partir de la extracción automatizada de características geográficas de imágenes satelitales, se combinen con datos implícitos, con el fin de no (hiper/in)visibilizar los territorios y sus comunidades. Lo anterior cuestiona las lecturas y categorías institucionales sobre estos territorios, a partir de sistemas como, por ejemplo, verlos desde su estatus de asentamiento de desarrollo informal, en vías de legalización o formalizado, o desde la condición de sus viviendas o calidad de hábitat, tipos de actividades económicas institucionalizadas y tipos de infraestructura pública institucional.

La diversificación de estas formas de ver los territorios/comunidades permite disminuir los sesgos y la continuidad de ciclos de precarización, al contrarrestar, por ejemplo, *narrativas de déficit* (D'Ignazio y Klein, 2020) sobre estos. De igual modo, permite identificar y disminuir el impacto de los aspectos considerados desde la política pública que no corresponden con sus realidades, sino a una inercia histórica de las presuntas necesidades y dinámicas de estas áreas y sus poblaciones, la cual proviene desde las instituciones y sus intereses,

debido a agendas políticas, oportunidades de financiación, lugares de mejoras de indicadores urbanos, entre otros. Sin embargo, la implementación de estas aproximaciones más participativas en los sistemas de GeoAI, como parte de las tareas de modelamiento, incluidas las de entrenamiento y validación de los modelos, se encuentran con la realidad que constituye la escasez de datos de referencia, sobre todo institucionales, y la limitada cobertura (espaciotemporal) de los que están disponibles.

Reflexiones desde los sistemas de clasificación y las categorías

El sistema del caso de georeferenciación automatizado aquí presentado ha encontrado retos importantes para la generación de conocimiento geoespacial de asentamientos precarizados, pues la representación de algunos elementos físicos relevantes para las comunidades riñe con muchas de las categorías oficiales. Por ejemplo, para una de las ciudades localizadas en África occidental, se plantea un énfasis en la generación de información de acceso a los servicios de salud por parte de las madres gestantes que habitan asentamientos precarizados.

Así, se plantea una primera aproximación desde los conjuntos de datos abiertos o institucionales disponibles; después, se incorporan datos generados por la comunidad, mediante OpenStreetMap alineado con maneras participativas de las comunidades de responder y manejar su propia visibilidad y responder a sesgos. A partir de allí, se realiza una estimación inicial y de referencia de diferentes niveles de accesibilidad a infraestructura y servicios de salud disponibles. Los resultados revelan que varios de estos no incluyen todas las prácticas, los relatos o la información compartida por las personas que habitan estos territorios. En particular, aquellos datos e información sobre los espacios y tipos de servicios tradicionales o comunitarios a los que se accede durante la gestación, incluso otros temas de salud, como las parteras, curanderos u otras personas que llevan a cabo prácticas tradicionales no normativas asociadas al bienestar y la salud. Al no contar con datos geoespaciales explícitos de referencia sobre estas prácticas, estos espacios y servicios, los cuales se encuentran por fuera de la regulación y el reconocimiento institucional, son invisibles dentro las cartografías automatizadas que se generan oficialmente y que son utilizadas para gestionar y manejar las decisiones sobre estos territorios urbanos. Esta situación es similar en otros sectores, como el de acceso a la educación, el saneamiento básico, el agua potable, o la percepción de seguridad o criminalidad.

Asumir que es viable modelar bajo paradigmas establecidos de GeoAI implica resultados, lecturas y aproximaciones incompletas, debido a que las categorías definidas desde las instituciones son normativas, en varios casos

estándar, y tienen por origen información explícita a nivel geoespacial. Dichas aproximaciones suelen no dar cuenta de la densidad y complejidad de la vida que existe fuera de esos marcos preestablecidos y que es más evidente en nuestros países del sur. Para responder a esto, se deben incorporar procesos de cocreación de visiones compartidas y situadas en el territorio. Esto permitiría, por un lado, que la agencia sobre los niveles de visibilidad cartográfica y, en general institucional, de las comunidades sujeto de estos sistemas de georeferenciación automatizada, sus prácticas y espacios, no solo sea respetada, sino activamente incluida en la configuración de la información geoespacial de los territorios urbanos que habitan.

Por otro lado, cuando las comunidades deciden a la par de las instituciones sobre estos niveles de visibilidad, se espera que se abra espacio para que la definición y caracterización de los datos básicos que alimentan estos sistemas de GeoAI puedan responder, en efecto, a perspectivas de datos y categorías que están situadas y sean más cercanas a sus realidades. Sin embargo, es clave resaltar que, al plantear visibilizar estas prácticas y espacios no normativos, se vuelve esencial garantizar procesos y espacios de información y exploración con las comunidades sobre los potenciales impactos no deseados (p. ej., intentos de las instituciones por regularizar/formalizar estas actividades) de la (hiper/in)visibilidad automatizada en las cartografías institucionales de estas prácticas. Esta reflexión final sobre este proyecto nos lleva directamente a nuestro segundo caso.

Caso 2. Proyecto de generación de datos para apoyar comunidades impactadas por inundaciones

Contexto

El registro de inundaciones y otras afectaciones por eventos climáticos extremos ha sido tradicionalmente una práctica de las entidades oficiales. En varios casos, el conocimiento geográfico permite estimar los niveles de vulnerabilidad o riesgo de las comunidades que habitan estos territorios. Estos cálculos y estimaciones son producidos por los Gobiernos nacionales, mientras las acciones de prevención y atención de desastres son delegadas a autoridades del nivel local o municipal. Este arreglo convencional de responsabilidades suele excluir el conocimiento local y las experiencias propias de las poblaciones que han sufrido los desastres y que, en varios casos, ya se han adaptado al riesgo o lo gestionan fuera de la institucionalidad.

La brecha existente entre las instituciones y las comunidades se acentúa en los territorios usualmente referidos como asentamientos precarizados, aquellas comunidades urbanas periféricas o en la ruralidad ubicadas en zonas en su

mayoría autoconstruidas. No es coincidencia que estas brechas tengan una relación alta con las desigualdades territoriales presentes en las ciudades del sur (De Andrade *et al*., 2021) o en zonas rurales con condiciones particulares de riesgo, como aquellas que se encuentran cerca de volcanes (Pardo *et al*., 2021). Como describen De Andrade *et al*. (2021), la generación voluntaria de datos geográficos sobre impactos de eventos climáticos, en especial aquellos creados a través de redes sociales, se concentra en zonas de la ciudad con altos ingresos, actividad económica formal o infraestructura consolidada (que también comprende la infraestructura de conectividad a internet). Así, se encuentra una baja representación de información en zonas de bajos ingresos, a pesar de contar con evidencia técnica de los impactos negativos que producen las inundaciones. Es necesario observar cuáles de los condicionantes tienen mayor prevalencia para esta baja representatividad, incluidos los costos de conectividad, el acceso limitado a dispositivos móviles, el menor interés en la producción de información, etc.

Son múltiples los casos en los que la organización comunitaria ha entrado en diálogo con las autoridades y ha servido de complemento o fuente de información para los estudios oficiales, varios de los cuales utilizan GeoAI como herramienta y parten del análisis de las condiciones físicas del territorio y datos explícitos (p. ej., pendientes de los terrenos, composición química de los suelos, arreglos físicos del relieve, cobertura y uso del suelo, entre otros). A través de estos sistemas, se estiman niveles de vulnerabilidad ante situaciones que impliquen consecuencias graves para la población, como eventos climáticos, geológicos y de otros tipos; sin embargo, persiste el problema subyacente asociado a definiciones comunes o criterios aceptados entre disciplinas, instituciones, Gobiernos y diferentes tipos de organizaciones.

El proyecto Datos a Prueba de Agua (Waterproofing Data, en su nombre original en inglés) permite realizar campañas de sensibilización, diálogo y producción de datos explícitos e implícitos, principalmente con comunidades, sobre todo en escuelas, acerca de lluvia e inundaciones en cinco estados de Brasil, bajo el liderazgo del Centro Nacional de Monitoreo de Riesgo de Desastres Naturales (Cemaden). Dichas campañas buscan responder a la necesidad de mejorar los sistemas de prevención del riesgo y las herramientas de GeoAI que los acompañan, para abrir espacios y crear estrategias de diálogo entre los múltiples participantes, con el fin de determinar las bases conceptuales y metodológicas del monitoreo de niveles de lluvia, construir los instrumentos de medición, registrar los eventos de inundación, y analizar cualitativamente los eventos históricos de inundación y las formas de representación de esta información. Los resultados de estas campañas son socializados con ambos grupos

de participantes, para que, de manera colectiva, se defina la forma en que pueden utilizarse dentro de los procesos institucionales de generación de alertas y de gestión del riesgo, que en la actualidad tiene aproximaciones de *arriba hacia abajo* en el manejo de desastres.

Reflexiones desde los sistemas de clasificación y las categorías

Las realidades locales de algunas regiones de países del sur colisionan con definiciones sobre cómo entender fenómenos específicos. En particular, las aproximaciones de fenómenos naturales que se consideran riesgos para las poblaciones, como lo pueden ser las inundaciones, los deslizamientos de tierra o las erupciones volcánicas, muchas veces parten, como se mencionó, de definiciones de riesgo con aproximaciones de *arriba hacia abajo*, donde los fenómenos naturales se estudian como objetos independientes para reducir la incertidumbre asociada a procesos que podrían tener consecuencias vitales para entidades humanas y más que humanas. Por ejemplo, Pardo *et al*. (2021) señalan que, en el caso de los impactos de las erupciones volcánicas, este tipo de aproximaciones falla en reconocer a los volcanes como parte esencial de la vida en sus laderas y vecindad, y se ignora el conocimiento de quienes allí habitan sobre su propio entorno.

Este ejemplo nos permite remarcar la importancia de los ejercicios de recolección de datos dialógicos, sistémicos, colaborativos y, en algunos casos, implícitos, como los del caso presentado, al ser elementos clave y estructurales para las aproximaciones de gestión de riesgo, así como para el desarrollo de las herramientas de GeoAI que las acompañan, dado que permiten una mayor alineación entre las estrategias institucionales y, a nivel de tecnología, entre las variables, las categorías y la manera de implementarlos (p. ej., priorización de eventos, condiciones, etc.); en efecto, es importante que representen la realidad vital de la población y no vayan en contra de los procesos de renovación y coexistencia entre todos los elementos humanos y más que humanos que entrecruzan de forma dinámica, en respuesta a las realidades de habitar estos territorios en riesgo.

En el panel Ciencia Ciudadana y Producción de Conocimiento Geográfico Local, que se llevó a cabo en abril del 2024 en el marco del proyecto Datos a Prueba de Agua y donde participamos dos de los autores del presente capítulo, una de las panelistas dio uno de los ejemplos más claros sobre esto. Nos relató cómo un grupo de personas expertas en gestión de riesgo llegaron a una comunidad, la cual había sido identificada por tener un nivel de riesgo muy alto por inundaciones, a apoyarlos en el desarrollo de estrategias para responder más adecuadamente a estos fenómenos. Al trabajar con los habitantes, se dieron cuenta

que las estrategias de gestión que iban a apoyar (resultado del análisis de datos institucionales y explícitos) no debían enfocarse en ese fenómeno de manera exclusiva, sino en otros con menor ocurrencia, pero que tenían mayor impacto en la calidad de vida y el bienestar de la comunidad. Las observaciones de campo y los datos implícitos recolectados revelaron que, debido a la alta cantidad de inundaciones en este territorio, la comunidad ya había desarrollado maneras de adaptarse a los impactos de este fenómeno, por lo que su resiliencia a estos eventos había aumentado y ya no eran, desde lo local, el mayor de sus riesgos.

En este contexto aparecieron prioridades asociadas a riesgos de deslizamiento, crecientes súbitas del río o caída de rocas y lodos. Sin embargo, estos son datos e información que los modelos de análisis de riesgo o de vulnerabilidad no consideraron por diversas razones. Entre estas se encuentran: (1) el uso de fuentes de datos institucionales con escalas inadecuadas (basados en datos generales que no capturan el nivel de detalle de la zona de estudio), donde una gran parte de los datos son geoespacialmente explícitos; y (2) se utilizan sistemas de clasificación y categorías de datos que hipervisibilizan elementos explícitos y tangibles de los fenómenos, como los temas topográficos, de pendientes, frecuencia e intensidad de lluvias, etc., que no son necesariamente aquellos que se manifiestan en el territorio, y que son percibidos por sus habitantes en el día a día y narran las vivencias de las comunidades.

En este tipo de circunstancias, la propuesta participativa de Datos a Prueba de Agua para la recolección de datos —en su etapa inicial, hace énfasis en la cocreación de datos que alimentan procesos de GeoAI analíticos relacionados a riesgos— permite que las escuelas se constituyan en el lugar de encuentro y diálogo entre los sistemas educativos locales, las comunidades y los equipos institucionales de gestión del riesgo. Allí, estudiantes, profesores, oficiales de Defensa Civil y personas de la comunidad narran sus experiencias relacionadas con los fenómenos naturales y sus riesgos, y construyen conceptos y categorías asociadas a la inundación, los niveles críticos de lluvia, los criterios para la generación de alertas climáticas, entre otros. Así, se producen datos complementarios para el sistema analítico de gestión del riesgo (p. ej., áreas con percepción del riesgo alta, eventos de inundación de escala local, registros de niveles de lluvia generados por las comunidades), el cual se espera que complemente los sistemas y métodos oficiales de análisis y gestión de geoinformación aceptados por los equipos técnicos y las autoridades oficiales.

Al coconstruir conceptos y criterios que antes se veían como elementos técnicos que solo podían ser formulados por personas expertas —donde las comunidades no eran consideradas como tal—, estas apuestas activamente participativas apoyan la formación y creación de diversas partes y elementos de los

GeoAI esenciales para la gestión del riesgo, contribuyendo a la democratización de la implementación de las GeoAI en el sur.

Reflexiones generales sobre las GeoAI en el sur

Un elemento común en los casos presentados en este capítulo es el cuestionamiento sobre la percepción de si todo aquello que utiliza datos digitalizados y sistemas digitales geoespaciales para su gestión muestra o representa realidades tangibles y ciertas de los territorios y las comunidades. Sobre este punto es esencial reflexionar acerca de cómo en la actualidad frente a la digitalización/datificación de los territorios no se habla solo de "la instrumentalización de espacios físicos con sensores", sino que amerita que se amplíe el entendimiento de estos procesos y sus tecnologías digitales, en este caso las GeoAI, como elementos sociotecnológicos. Esto incluye también el entendimiento de estas integraciones de tecnologías de la información y las estrategias de cuantificación de personas en la gobernanza en otros contextos sociofisicos. Las reflexiones sobre lo que es (hiper/in)visible (D'Ignazio y Klein, 2020; Ricaurte, 2019; Taylor, 2017) a través de estas geotecnologías, en los usos descritos y otros, cobra relevancia por diversas razones.

Por un lado, la existencia de dinámicas socioespaciales y culturales no necesariamente se considera dentro de los sistemas de GeoAI, porque estas ocurren de maneras no hegemónicas y, por ende, son entendibles y visibles solo a través de este tipo de datos no explícitos y de sistemas de clasificación y categorías que los integren y no los excluyan. Lo anterior es evidente en los casos presentados de georeferenciación automatizada de asentamientos informales o en el proyecto Datos a Prueba de Agua. Por otro lado, la manera en la que los datos geoespaciales —tanto explícitos como implícitos— se relacionan con la reproducción de prácticas espaciales de poder (Mattern, 2018) plantea preguntas sobre la necesidad de (hiper/in)visibilidad de ciertos grupos en los procesamientos de estas geotecnologías.

En el primero, estas preguntas surgen al hacer énfasis en prácticas locales para promover el bienestar y el cuidado a la salud, como lo son curanderos, parteras, etc.; estos espacios, servicios y personas suelen hacerse evidentes en los datos implícitos (p. ej., memorias, narrativas, entre otros) de las comunidades y no en los explícitos, por lo que no se visibilizan cuando se hacen cartografías (automatizadas) de infraestructuras de salud en estos territorios. En el segundo, se plantea enfatizar en la cocreación con comunidades de conceptos (Bryan, 2011) y criterios de riesgo, para que estos reflejen las realidades de sus territorios y no sean el resultado de análisis realizados en otros espacios e

instituciones que, al catalogarlos como de alto o bajo riesgo, terminan teniendo poder e influencia sobre la gestión de la vida en estos territorios, sin ser estas sus necesidades y realidades actuales.

Lo anterior nos lleva a sugerir que, para que un sistema de GeoAI pueda, en efecto, responder a realidades humanas y más que humanas, es clave reconocer que las dinámicas humanas y naturales, las múltiples escalas espaciales y temporales, y la coexistencia de la pluralidad y la diversidad deben considerarse parte de estos sistemas, con una sensibilidad situada en las realidades de los territorios en los que se adelantan proyectos asociados a tecnologías que, en muchos casos, hacen abstracciones conectadas a agendas que no se originan en los lugares de estudio, desarrollo o implementación. Adicionalmente, los procesos de gubernamentalidad deberían situarse con precisión en los territorios y las comunidades en clave local, que sirvan para promover la autorregulación y autocontrol mientras se apalancan procesos de autodeterminación. Dentro de estos procesos los sistemas de GeoAI deberían servir como apoyo desde esta visión construida de *abajo hacia arriba*.

Las implicaciones que tiene la imposición de puntos de vista que no están situados en los territorios dificultan el entendimiento, la apropiación y el uso de sistemas particulares; en general, la agencia de las comunidades de dichos territorios. Las dinámicas humanas y más que humanas no son todas traducibles a datos. En algunos casos, las escalas espaciales y temporales se pueden cuantificar, como es evidente en los casos aquí presentados, así como en algunos referenciados desde la literatura (Hoefsloot y Gateri, 2024; Pardo *et al*., 2021), pero las implicaciones del conocimiento local sobre cómo habitar estos territorios o la forma de responder a situaciones de riesgo están más afincadas en procesos intangibles, que se refieren a visiones locales sobre el territorio y a cómo se habita. Así mismo, la pluralidad y la diversidad, cuando lo humano y lo más que humano se encuentran, superan las limitaciones de capacidad del entendimiento humano. En un sentido relacional, para que un sistema GeoAI pueda responder a la complejidad que proponen múltiples sistemas en interacción, a lo largo de tiempos que cruzan escalas temporales y espaciales, debería definirse desde el diálogo y la búsqueda conjunta, mientras se reconocen las limitaciones de sistemas que principalmente operan desde lo cuantificable. Como se ha mencionado al inicio de este capítulo, ningún sistema está libre de sesgos ni es posible cuantificar el mundo en su totalidad. Por esto, se deben priorizar aquellas experiencias y la información de carácter local que tenga una relación directa con la vida y el ecosistema local en el cual generarán un impacto.

Teniendo en cuenta estas reflexiones, surgen varias preguntas: ¿la respuesta está en que estos sistemas de georeferenciación automatizada (p. ej., en el caso

de la georeferenciación de asentamientos precarizados) incluyan estos datos geoespaciales implícitos sobre servicios y espacios infraestructurales no normativos? ¿Cómo se pueden crear nuevas categorías relevantes localmente dentro de los datos y los sistemas de GeoAI institucionales, de forma que se incluya la diversidad? O si, en vez de automatizar la inclusión de este tipo de prácticas en estos tipos de GeoAI, ¿deberían, en general, ser mapeadas de manera automática o más bien se quedan para ser representadas solo con procesos de visibilización que sean contextuales, participativos y locales, que no supongan un riesgo para las comunidades sujetas de datos en diferentes frentes? ¿Cuáles son los procesos clave para identificar los componentes locales y comunitarios que se pueden incluir en estos sistemas y cuáles deben excluirse?

Nuestras reflexiones nos llevan a concluir que el proceso de carácter sistémico que considera sistemas ecológicos y de suelos, culturales y sociales, el cual se comienza a explorar en ambos casos, se direcciona al desarrollo de una visión transdisciplinaria colectiva. El desarrollo de estrategias para la implementación de GeoAI y sus sistemas y plataformas debe responder a aproximaciones que sean integradoras, diversas y contextuales, para que la toma de decisiones esté situada en la realidad de nuestros territorios y comunidades del sur. Estas estrategias más participativas asociadas a las GeoAI deberían permitir la interconexión de múltiples escalas sociales, espaciales y temporales, además de promover la coexistencia de la pluralidad y la diversidad.

Referencias

Anderson, B. (1991). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso Books.

Abascal, A., Rodríguez-Carreño, I., Vanhuysse, S., Georganos, S., Sliuzas, R., Wolff, E. y Kuffer, M. (2022). Identifying degrees of deprivation from space using deep learning and morphological spatial analysis of deprived urban areas. *Computers, Environment and Urban Systems*, *95*. https://doi.org/10.1016/j.compenvurbsys.2022.101820

Alaimo, C. y Kallinikos, J. (2021). Managing by data: Algorithmic categories and organizing. *Organization Studies*, *42*(9), 1385-1407. https://doi.org/10.1177/0170840620934062

Ash, J., Kitchin, R. y Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, *42*(1), 25-43. https://doi.org/10.1177/0309132516664800

Barry, L. (2019). The rationality of the digital governmentality. *Journal of Cultural Research*, *23*(4), 365-380. https://doi.org/10.1080/14797585.2020.1714878

Beer, D. (2019). *The data gaze: Capitalism, power and perception*. Sage.

Behrendt, F. y Sheller, M. (2024). Mobility data justice. *Mobilities*, *19*(1), 151-169.

Broomfield, H. y Reutter, L. (2022). In search of the citizen in the datafication of public administration. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221089302

Bryan, J. (2011). Walking the line: Participatory mapping, indigenous rights, and neoliberalism. *Geoforum*, *42*(1), 40-50. https://doi.org/10.1016/j.geoforum.2010.09.001

Buolamwini, J. y Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Carr, C. y Hesse, M. (2023). Is tech-led urbanism sociopathic? On the invisible and visible urban agendas driven by Big Tech. https://hdl.handle.net/10993/58933

Datta, A. (2024). The informational periphery: Territory, logistics and people in the margins of a digital age. *Asian Geographer*, *41*(2), 125-142. https://doi.org/10.1080/10225706.2023.2253233

de Albuquerque, J. P., Eckle, M., Herfort, B. y Zipf, A. (2016). Crowdsourcing geographic information for disaster management and improving urban resilience: an overview of recent developments and lessons learned. En *European handbook of crowdsourced geographic information* (pp. 309-321). https://doi.org/10.5334/bax.w

de Andrade, S. C., Porto de Albuquerque, J., Restrepo-Estrada, C., Westerholt, R., Rodríguez, C. A. M., Mendiondo, E. M. y Delbem, A. C. B. (2021). The effect of intra-urban mobility flows on the spatial heterogeneity of social media activity: Investigating the response to rainfall events. *International Journal of Geographical Information Science*, *36*(6), 1140-1165. https://doi.org/10.1080/13658816.2021.1957898

D'Ignazio, C. y Klein, L. (2020). 2. Collect, analyze, imagine, teach. *Data Feminism*. https://data-feminism.mitpress.mit.edu/pub/ei7cogfn

Farman, J. (2010). Mapping the digital empire: Google Earth and the process of postmodern cartography. *New Media & Society*, *12*(6), 869-888. https://doi.org/10.1177/1461444809350900

Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M. y Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293

Flyverbom, M. y Garsten, C. (2021). Anticipation and organization: Seeing, knowing and governing futures. *Organization Theory*, *2*(3), 1-25. https://doi.org/10.1177/26317877211020325

Goodfriend, S. (2021). A street view of occupation: Getting around Hebron on Google Maps. *Visual Anthropology Review*, *37*(2), 217-476. https://anthrosource.onlinelibrary.wiley.com/doi/epdf/10.1111/var.12247

Gutiérrez, J. D. y Muñoz-Cadena, S. (2023). Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia. *GIGAPP Estudios Working Papers*, *10*(267-272), 365-395.

Heeks, R., Evans, J. Z., Graham, M. y Taylor, L. (2020). The urban data justice case study collection. *Digital Development Working Paper 88*. http://dx.doi.org/10.2139/ssrn.3705563

Hoefsloot, F. I. y Gateri, C. (2024). Contestation, negotiation, and experimentation: The liminality of land administration platforms in Kenya. *Environment and Planning D: Society and Space*, *42*(5-6). https://doi.org/10.1177/02637758241254943

Hoefsloot, F. I., Jiménez, A., Martínez, J., Miranda Sara, L. y Pfeffer, K. (2022). Eliciting design principles using a data justice framework for participatory urban water governance observatories. *Information Technology for Development*, *28*(3), 617-638. https://doi.org/10.1080/02681102.2022.2091505

Hoefsloot, F. I., Richter, C., Martínez, J. y Pfeffer, K. (2023). The datafication of water infrastructure and its implications for (il)legible water consumers. *Urban Geography*, *44*(4), 729-751. https://doi.org/10.1080/02723638.2021.2019499

Hu, Y., Gao, S., Newsam, S. y Lunga, D. (2019). GeoAI 2018 workshop report the 2nd ACM SIGSPATIAL international workshop on GeoAI: AI for geographic knowledge discovery Seattle, WA, USA-November 6, 2018. *SIGSPATIAL special*, *10*(3), 16. https://doi.org/10.1145/3307599.3307609

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T. y Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12). https://doi.org/10.1088/1748-9326/ab4e55

Jameson, S., Richter, C. y Taylor, L. (2019). People's strategies for perceived surveillance in Amsterdam Smart City. *Urban Geography*, *40*(10), 1467-1484.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y. y Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, *34*(4), 625-636. https://doi.org/10.1080/13658816.2019.1684500

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*, 263-274. https://doi.org/10.1177/1461444809350900

Kitchin, R. (2020). *Using digital technologies to tackle the spread of the coronavirus: Panacea or folly. The Programmable City Working Paper 44*. https://progcity.maynoothuniversity.ie/wp-content/uploads/2020/04/Digital-tech-spread-of-coronavirus-Rob-Kitchin-PC-WP44.pdf

Kitchin, R., Maalsen, S. y McArdle, G. (2016). The praxis and politics of building urban dashboards. *Geoforum*, *77*, 93-101. https://doi.org/10.1016/j.geoforum.2016.10.006

Kuffer, M., Pfeffer, K. y Sliuzas, R. (2016). Slums from space: 15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6), 455. https://doi.org/10.3390/rs8060455

Madianou, M., Dencik, L., Aradau, C., Taylor, L., Metcalfe, P. y Perret, S. (2020). The biometric lives of migrants: Borders, discrimination and (in) justice. *AoiR: Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11137

Mattern, S. (2018, noviembre). Maintenance and care. *Places*. https://placesjournal.org/article/maintenance-and-care/?cn-reloaded=1

Mortaheb, R. y Jankowski, P. (2023). Smart city re-imagined: City planning and GeoAI in the age of big data. *Journal of Urban Management*, *12*(1), 4-15. https://doi.org/10.1016/j.jum.2022.08.001

Oluoch, I., Kuffer, M. y Nagenborg, M. (2022). In-between the lines and pixels: Cartography's transition from tool of the state to humanitarian mapping of deprived urban areas. *Digital Society*, *1*. https://doi.org/10.1007/s44206-022-00008-0

Pardo, N., Espinosa, M. L., González-Arango, C., Cabrera, M. A., Salazar, S., Archila, S. *et al.* (2021). Worlding resilience in the Doña Juana volcano-páramo, Northern Andes (Colombia): A transdisciplinary view. *Natural Hazards*, *107*, 1845-1880. https://doi.org/10.1007/s11069-021-04662-4

Pelizza, A. y Kuhlmann, S. (2017). *Mining governance mechanisms: Innovation policy, practice, and theory facing algorithmic decision-making*. Springer.

Puig de la Bellacasa, M. (2017). *Matters of care: Speculative ethics in more than human worlds*. University of Minnesota Press.

Ricaurte, P., Gómez-Cruz, E. y Siles, I. (2024). Algorithmic governmentality in Latin America: Sociotechnical imaginaries, neocolonial soft power, and authoritarianism. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241229697

Ricaurte, P. (2023). AI for/by the majority world: From technologies of dispossession to technologies of radical care. En *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 3-4). https://doi.org/10.1145/3600211.3607544

Ricaurte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, *20*(4), 350-365. https://doi.org/10.1177/1527476419831640

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*, 532-547. https://doi.org/10.1016/0010-0285(75)90021-3

Schou, J. y Hjelholt, M. (2019). Digital state spaces: State rescaling and advanced digitalization. *Territory, Politics, Governance*, *7*(4), 438-454. https://doi.org/10.1080/21622671.2018.1532809

Sawhney, N. (2023). Contestations in urban mobility: Rights, risks, and responsibilities for urban AI. AI & Society, *38*(3), 1083-1098. https://doi.org/10.1007/s00146-022-01502-2

Song, Y., Kalacska, M., Gašparović, M., Yao, J. y Najibi, N. (2023). Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping. *International Journal of Applied Earth Observation and Geoinformation*, *120*. https://doi.org/10.1016/j.jag.2023.103300

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717736335

Taylor, L., Richter, C., Jameson, S. y Pérez de Pulgar, C. (2016). Customers, users or citizens? Inclusion, spatial data and governance in the smart city. http://dx.doi.org/10.2139/ssrn.2792565

Thomson, D. R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H. *et al.* (2020). Need for an integrated deprived area "slum" mapping system (IDEAMAPS) in low-and middle-income countries (LMICS). *Social Sciences*, *9*(5), 80. https://doi.org/10.3390/socsci9050080

Weiskopf, R. y Hansen, H. K. (2022). Algorithmic governmentality and the space of ethics: Examples from "People Analytics". *Human Relations*, *76*(3), 483-506. https://doi.org/10.1177/00187267221075346

Wynne, B. (1988). Unruly technology: Practical rules, impractical discourses and public understanding. *Social Studies of Science*, *18*(1), 147-167. https://doi.org/10.1177/030631288018001006

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M. y Baeza-Yates, R. (2017). FA*IR: A fair top-k ranking algorithm. En *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1569-1578). https://doi.org/10.1145/3132847.3132938

INTELIGENCIA ARTIFICIAL PARA LA PREVENCIÓN DEL ABUSO SEXUAL EN LÍNEA DE LA INFANCIA Y LA ADOLESCENCIA EN COLOMBIA*

Pablo Andrés Arbeláez, Lina María Saldarriaga, Viviana Quintero, Ángela Castillo, Juanita Puentes, Yuly Calderón, Wilmar Osejo, Alejandro Castañeda, Carolina Paz, Laura Hernández, Diana Agudelo

* Este estudio fue financiado por Safe Online y el Tech Coalition. Véase https://safeonline.global/

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.08

Introducción

La explotación sexual de niñas, niños y adolescentes en línea (OCSEA, por sus siglas en inglés) es un delito en el que se usan el internet y las tecnologías digitales para abusar y explotar sexualmente a personas menores de dieciocho años (WeProtect, 2023). Este fenómeno incluye la producción, la distribución y el consumo de material de explotación sexual de niñas, niños y adolescentes (CSAM, por sus siglas en inglés), el acoso sexual (*grooming*), la transmisión en vivo de abuso sexual, la extorsión sexual, entre otros.

En los últimos años se ha observado un incremento alarmante en el volumen de OCSEA. De acuerdo con WeProtect (2023), desde el 2019 los reportes mundiales sobre CSAM han aumentado cerca de un 87 % y los delitos de captación de menores en internet se incrementaron un 80 % entre el 2018 y el 2022. Asimismo, Van der Bruggen y Blokland (2020) indican que el CSAM se ha convertido en uno de los tipos de contenido más populares y de mayor difusión en la *dark web*. Aunque apenas el 2 % de los sitios ocultos en la *dark web* está relacionado con CSAM, aproximadamente el 80 % del tráfico se dirige a estos sitios, lo que indica que este contenido es solicitado y visitado con frecuencia. El estudio *Scale of harm* (International Justice Mission y University of Nottingham Rights Lab, 2023) señala que uno de cada cien menores de dieciocho años ha sido víctima de transmisión en directo de abusos sexuales en Filipinas, país que se considera el punto crítico a nivel global para este delito.

Entre el 2020 y el 2022, la Internet Watch Foundation (IWF, 2022) reportó un crecimiento del 360 % en los casos de imágenes sexuales "autogeneradas" de niñas y niños entre siete y diez años en Reino Unido. Para el 2023, la IWF

también reportó un aumento del 65 % en este mismo país (104 282 casos), en comparación con el 2022 (63 057 casos) (IWF, s. f.). Por su parte, el número de denuncias recibidas por el National Center for Missing and Exploited Children (NCMEC, 2024) en relación con extorsión sexual financiera dirigida a niñas, niños y adolescentes también se incrementó un 300 % entre el 2021 y el 2023. Recientemente, esta misma organización alertó que los casos de *grooming* están presentando un crecimiento exponencial en el mundo. En el 2021, a nivel mundial, CyberTipline procesó 44 155 casos, en el 2022 fueron 80 524 y en el 2023 recibieron 186 819 reportes de situaciones *grooming* en línea que involucraron a menores de dieciocho años.

En el contexto latinoamericano, los pocos datos que existen también muestran un panorama similar. Solo entre el 2021 y el 2022 la línea de reporte Te Protejo de Colombia procesó cerca de 1633 casos relacionados con sextorsión, *sexting*, *grooming* y ciberacoso sexual en Colombia. La diferencia que existe entre la manera en que los países del norte y el sur global responden al OCSEA es abismal. Mientras que países de América del Norte, Europa, Australia e incluso algunos de Asia cuentan con recursos técnicos, legales y económicos para combatir este delito, los países ubicados en el sur global, en especial en América Latina, carecen de legislaciones robustas y tienen un acceso limitado a tecnología y recursos económicos que les permitan proteger los derechos de niñas, niños y adolescentes; esto se agrava por el hecho de que en estos países se reportan con más frecuencia víctimas de OCSEA (WeProtect, 2023).

Los desafíos que enfrenta Latinoamérica frente a este fenómeno son diversos. La variabilidad en los datos de los reportes, el volumen de información en las conversaciones donde ocurren abusos, la falta de personal capacitado para procesar dicha información y la forma en que se clasifican los delitos son solo algunas de las dificultades a las que se enfrentan las autoridades a diario al abordar casos de OCSEA en la región. Pero tal vez uno de los mayores retos en la prevención y combate de estos delitos es el impacto perjudicial a la que están expuestos analistas y autoridades. Según Ahern *et al*. (2016), la exposición constante a este tipo de material puede afectar gravemente la salud mental y bienestar integral de las personas.

El uso de la inteligencia artificial en la lucha contra el OCSEA y el CSAM

La investigación aplicada sobre el uso de inteligencia artificial (IA) para combatir y prevenir el OCSEA y el CSAM es un campo relativamente reciente. En el mundo existen numerosos estudios e iniciativas centrados en el análisis de videos e

imágenes de abuso, los cuales desarrollan soluciones innovadoras para el manejo de este material (Gangwar *et al*., 2021; Guerra y Westlake, 2021); sin embargo, los estudios y las herramientas dedicados al análisis de texto son escasos, en especial cuando se trata de textos en idiomas distintos al inglés (Ngo *et al*., 2023).

Los pocos estudios enfocados en análisis de texto han utilizado modelos de *machine learning* y *deep learning* para identificar contenido relacionado con abuso, *bullying*, *grooming* o sextorsión en plataformas como Facebook, X (antes Twitter), YouTube e incluso en algunos videojuegos. Los datos utilizados para entrenar estos modelos provenían, en algunos casos, de actores clave del entorno digital, como plataformas tecnológicas o autoridades, mientras que en otros fueron generados de forma directa por los investigadores.

Uno de los ejemplos más exitosos de uso de IA para abordar este tipo de problemáticas para las líneas de reporte como Te Protejo es el proyecto Augmented Visual Intelligence and Targeted Online Research (AviaTor), impulsado por la International Association of Internet Hotlines (INHOPE). Esta plataforma apoya a las líneas de reporte y a las autoridades en el procesamiento, evaluación y priorización de informes de CSAM. Este proyecto ha mostrado avances prometedores al utilizar técnicas avanzadas de visión por computadora e IA para analizar y extraer información relevante de imágenes y videos, lo que permite priorizar las denuncias y realizar análisis de manera más eficiente (INHOPE, 2023).

Otra de las iniciativas que existen a nivel global para analizar conversaciones entre agresores y menores de edad es DRAGON-Shield. Esta plataforma, desarrollada por la Universidad de Swansea, ofrece dos herramientas: un detector de *grooming* en línea para las autoridades (DRAGON-Spotter) y una plataforma de capacitación en prevención de *grooming* para cuidadores de niños, niñas y adolescentes (DRAGON-Shield). El DRAGON-Spotter combina lingüística e IA para detectar contenido de *grooming* y las tácticas de lenguaje que los agresores utilizan para abusar de niñas, niños y adolescentes. La herramienta analiza texto extraído de conversaciones entre agresores y víctimas, el cual ha sido recopilado por las autoridades en casos de OCSEA. Al igual que el proyecto AviaTor, DRAGON-Shield y DRAGON-Spotter han mostrado resultados muy prometedores en el análisis de interacciones entre abusadores y víctimas (Lorenzo-Dus *et al*., 2023).

Aunque estos ejemplos representan avances importantes en el diseño de herramientas de IA para ayudar a las líneas de reporte y a las autoridades a recopilar datos y procesar informes de manera más eficiente y segura, es importante notar que la mayoría se enfoca en el análisis de fotos y videos de contenido sexual, dejando de lado el procesamiento del texto. Además, ninguna de estas herramientas ha sido desarrollada para el contexto latinoamericano, donde el

español es el idioma principal en el que ocurre el abuso. Por último, una de las mayores limitaciones es que todas estas herramientas se centran en el análisis de la relación entre víctimas y abusadores, sin abordar la fuente del abuso: las conversaciones e interacciones entre los agresores.

El estudio

Este estudio, desarrollado entre el 2021 y el 2023 por medio de una alianza entre el Departamento de Psicología y el Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA) de la Universidad de los Andes, Aulas en Paz y Te Protejo, planteó desarrollar dos modelos de IA para el procesamiento de información relacionada con OCSEA. El primer modelo tuvo por objetivo optimizar el análisis de los reportes sobre sextorsión, *sexting*, *grooming* y ciberacoso sexual recibidos en la línea de reporte Te Protejo en Colombia. Esta herramienta analiza y clasifica los reportes, identificando patrones en los textos recibidos por los analistas. Además de mejorar la comprensión, el análisis y la derivación de los casos, el modelo buscó reducir el riesgo para la salud mental de los analistas, limitando su exposición al contenido dañino de estos. El clasificador desarrollado logró analizar y caracterizar la información contenida en los reportes a través de criterios como la gravedad del delito y el tipo de daño causado por la situación (WeProtect, 2021).

El segundo modelo que se construyó fue un prototipo de un sistema de alerta sobre OCSEA. Esta herramienta analiza información de foros relacionados con esta temática encontrados en la web profunda a través de la plataforma Atlas de Web-IQ. Esta examina el texto de las conversaciones en estos foros e identifica patrones relacionados con los temas, su objetivo, su función, los sesgos cognitivos utilizados por los agresores y el posible daño que podrían causar a las víctimas.

Metodología

Fuentes de información

La información utilizada para el modelamiento de las dos herramientas de IA provino de dos fuentes. En primer lugar, se tomaron los reportes recibidos y procesados a través de la línea de reporte Te Protejo Colombia. Esta es una plataforma digital administrada por la organización de la sociedad civil Red PaPaz, la cual permite a cualquier ciudadano en Colombia identificar e informar sobre situaciones que afectan los derechos de niñas, niños y adolescentes.

Entre estas situaciones, Te Protejo recibe reportes sobre la divulgación, compra, venta o intercambio de CSAM, así como situaciones de OCSEA hacia personas menores de dieciocho años. El conjunto de datos que se analizó abarca casos desde enero del 2021 hasta diciembre del 2022. Esto permitió generar una base de 1196 reportes, donde se incluyeron casos relacionados con *grooming*, extorsión sexual, divulgación no consentida de contenido sexual y otras formas de hostigamiento a personas menores de dieciocho años en internet.

La segunda fuente de información provino de los datos extraídos de la plataforma Atlas de Web-IQ. Esta herramienta de IA de fuentes abiertas ofrece acceso seguro a un conjunto estructurado de datos de foros en la web profunda, chats y otros sitios donde se discuten temas relacionados con la OCSEA. A través de ella, se realizaron búsquedas con palabras clave asociadas a la OCSEA y la región, lo que permitió identificar foros en los que se mencionaban situaciones de abuso y explotación sexual en entornos digitales en español. Estas palabras clave fueron recopiladas por las analistas de Te Protejo y del equipo de investigación, lo que dio como resultado más de 2800 entradas, que representan hilos de conversaciones en los que, entre otras cosas, se encontraron descripciones de situaciones reales o imaginarias de abuso, guías para realizar o replicar actos de explotación, y solicitudes de intercambio de CSAM. La información obtenida a partir de las búsquedas corresponde a conversaciones que tuvieron lugar entre el 2018 y el 2022.

Manejo ético y responsable de la información

Uno de los aspectos más importantes en este proyecto fue el manejo ético y responsable de la información y el cuidado de la salud mental y el bienestar del equipo. Desde el inicio se implementó una política de salvaguardia que sigue los estándares establecidos por la Universidad de los Andes y la American Psychological Association (APA).

De manera adicional, frente a la información obtenida de Te Protejo y Atlas, se firmaron acuerdos de confidencialidad para el manejo de la información. Estos garantizaron el manejo seguro de los datos en servidores encriptados y limitaron la exposición de los investigadores a información potencialmente dañina para su bienestar y salud mental.

Sistema de anotación

Uno de los aspectos clave en el desarrollo de modelos de IA para abordar el OCSEA es la creación de un sistema de anotación que permita entrenarlos en el lenguaje,

los temas y las formas de interacción propias de este fenómeno. Este sistema de anotación es un conjunto de categorías y criterios diseñados para etiquetar, clasificar y organizar los datos recolectados. En este proyecto, se crearon dos sistemas de anotación: uno para los reportes de Te Protejo y otro para las conversaciones encontradas en Atlas; aunque ambos emplearon categorías distintas, se basaron en criterios similares.

Criterios principales para el desarrollo de los sistemas de anotación

Temática relevante: el sistema debe ser capaz de reconocer descripciones y ejemplos concretos de situaciones en las que ocurre OCSEA. Estas categorías incluyen tipos específicos de interacciones sospechosas, comportamientos inapropiados o contenido que indique abuso.

Precisión: las clasificaciones deben ser precisas, con descripciones claras de cada categoría, para garantizar que sea posible determinar con exactitud si un caso cumple con las características señaladas.

Adaptabilidad: los sistemas deben ser capaces de ajustarse y evolucionar con el tiempo. Conforme se identifican nuevos patrones de comportamiento o tipos de abuso, se incorporan nuevas categorías y se ajustan los criterios de clasificación, con el fin de asegurar su relevancia y efectividad.

A medida que los modelos se entrenaron y se pusieron a prueba, el sistema de anotación fue ajustado para reflejar mejor la complejidad de la información analizada. Por ejemplo, en los primeros entrenamientos se descubrió que un solo reporte recibido por Te Protejo contenía datos sobre varios delitos de forma simultánea, lo que implicaba la necesidad de enviarlo a diferentes autoridades competentes. En el caso de los foros de chat que se obtuvieron mediante Atlas, las primeras iteraciones del entrenamiento del modelo revelaron una enorme variabilidad en el tipo y la función de la información encontrada. Esta abarcaba desde relatos imaginarios hasta instrucciones para llevar a cabo abusos, así como detallados tutoriales técnicos para aprender a ocultar material de CSAM.

Las modificaciones permitieron que el sistema de anotación diseñado para la base de datos de Te Protejo clasificara los casos como varios delitos simultáneos, en múltiples categorías de riesgo, y evaluara la gravedad de la situación mediante la clasificación de daños propuesta por WeProtect (2022) y el grado de criminalidad según el índice de Wolak *et al.* (2010). El segundo sistema, desarrollado a partir de información obtenida de Atlas sobre conversaciones entre agresores, le permitió al modelo comprender de mejor manera las características del discurso en los foros. Esta nueva clasificación incluyó categorías como el

contexto de las conversaciones (si eran relatos reales, ficticios u opiniones) y los temas generales tratados (instrucciones para el uso de plataformas tecnológicas para agredir, estrategias para evadir la ley, métodos de captación de víctimas, entre otros). También incorporó la clasificación de los participantes dentro de la tipología de agresores propuesta por Tener *et al.* (2015), al igual que las distorsiones cognitivas y las estrategias de desconexión moral identificadas en las conversaciones (Steel *et al.*, 2020; Bandura, 1999).

El desarrollo de estos sistemas de anotación más abarcadores y dinámicos resultó ser una de las innovaciones más importantes del proyecto, ya que permitió que la IA empezara a comprender la complejidad del OCSEA, así como la lógica usada por los analistas de este tipo de información, lo que mejoró significativamente la capacidad de la herramienta para interpretar y predecir el conjunto de datos.

Procedimiento

Preprocesamiento de los datos

Los datos utilizados para desarrollar los modelos del clasificador y el prototipo de sistema de alerta provienen de reportes que incluyen información personal, como números de teléfono, identificaciones, correos electrónicos y URL. Para proteger la privacidad y el anonimato de las personas, se eliminaron de manera sistemática todos los datos sensibles. Es fundamental resaltar que la información personal de los reportes o conversaciones no fue empleada en el entrenamiento del modelo. Este enfoque asegura un análisis ético y responsable, al respetar siempre los derechos de privacidad de los individuos.

Además, la colaboración entre los investigadores de *machine learning* y los expertos en OCSEA fue clave para el procesamiento de los datos. Debido a la naturaleza sensible de la información, los investigadores de *machine learning* no tuvieron acceso directo a los datos, por lo que se implementó un proceso continuo de validación entre ambos equipos. En principio, los investigadores encargados del desarrollo del método trabajaron con entradas que consistían en una sola fila de una hoja de cálculo; a partir de esta entrada, generaban una predicción inicial basada en datos en bruto. Luego, los expertos en OCSEA, quienes eran los únicos con acceso autorizado a los reportes y las conversaciones, revisaban los casos en los que fallaba el modelo de procesamiento de lenguaje natural (*natural language processing*, NLP), realizando un análisis exhaustivo de cada uno. En concreto, la revisión consistía en validar la precisión con la que el modelo predecía la existencia o no de una categoría de anotación en un reporte o en una conversación. Una vez los expertos en OCSEA determinaban si el

modelo había anotado correcta o incorrectamente la información, los investigadores de *machine learning* ajustaban la estrategia para mejorar el rendimiento del modelo. Este proceso iterativo permitió optimizar de manera continua el sistema predictivo, subrayando así la importancia de contar con un equipo multidisciplinario.

Análisis descriptivos

La figura 8.1 muestra las categorías finales de clasificación en las distintas categorías del sistema de anotación. La dimensión *asunto* incluye un total de 994 instancias de datos distribuidas en 8 clases; entre estas, sextorsión tiene el mayor número de instancias, con un total de 299. En la dimensión *grado de criminalidad* hay 943 instancias de datos, con más de la mitad pertenecientes a intención de daño; la clase menos representada en esta dimensión es fin comercial, con solo 21 instancias de datos. Por último, la dimensión *daño* consta de un total de 702 instancias de datos distribuidas en cuatro clases.

La figura 8.2 presenta la distribución de los datos relacionados con las diferentes categorías de conversaciones entre agresores obtenidas de Atlas. Las dimensiones propuestas para este análisis son: asunto, tipo de agresor, contexto y distorsiones cognitivas del agresor. Es importante resaltar la disparidad en la cantidad de clases y el desbalance de datos en cada dimensión, similar a lo que se observa en la base de datos de reportes. En particular, las últimas tres dimensiones (tipo de agresor, contexto y distorsiones cognitivas) tienen menos categorías, lo que hace aún más crucial contar con un mayor número de ejemplos. La falta de datos dificulta establecer correlaciones que permitan identificar una misma conversación en diferentes categorías, lo que complica la predicción del modelo. Por ello, es esencial disponer de datos equilibrados y suficientes en todas las dimensiones, para mejorar la precisión y efectividad del análisis.

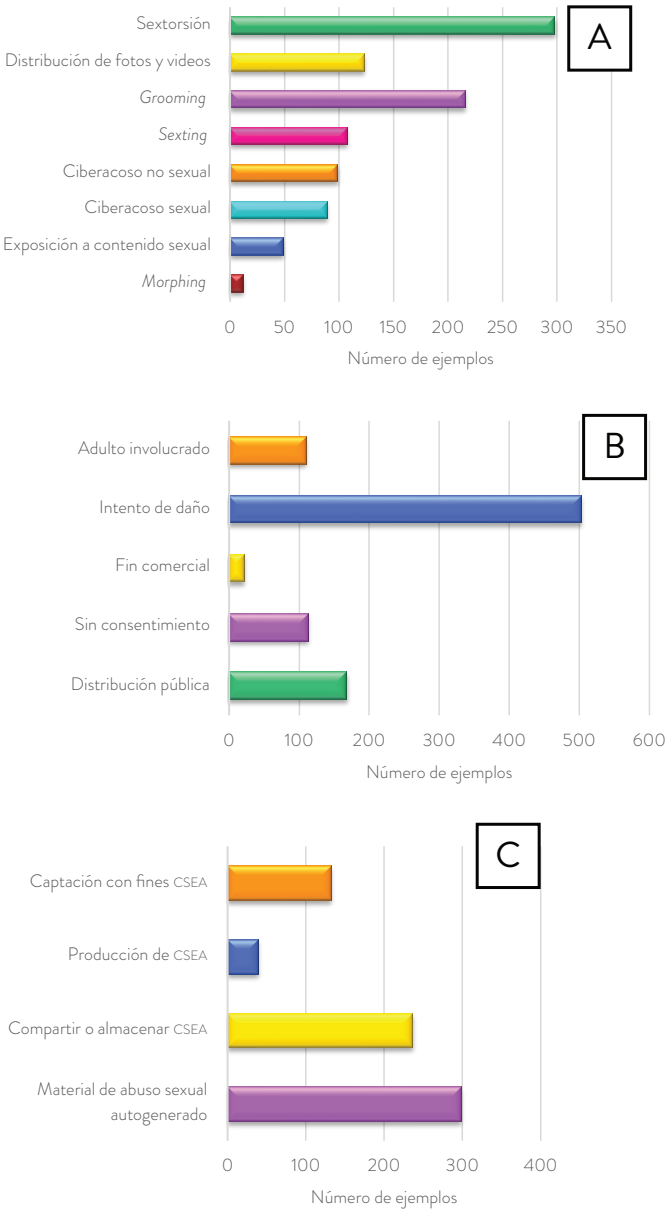


Figura 8.1. Distribución de instancias de reportes de la línea de reportes de Te Protejo en las dimensiones: (a) asunto, (b) grado de criminalidad y (c) daño
Fuente: elaboración propia.

Figura 8.2. Distribución de las instancias de las categorías en las conversaciones entre agresores para las dimensiones (a) asunto, (b) tipo de agresor, (c) contexto y (d) distorsiones cognitivas del agresor

Fuente: elaboración propia.

Resultados

Herramienta 1: clasificador para la línea de reportes Te Protejo

El objetivo principal de este modelo fue categorizar de manera efectiva y precisa los reportes recibidos por la línea Te Protejo en las categorías mencionadas. Para lograr esto, el modelo empleó una variedad de características para la clasificación, teniendo en cuenta las correlaciones observadas entre las diferentes categorías.

Correlaciones

En este contexto, las correlaciones se establecen en la medida en que dos categorías de reportes tienden a aparecer juntas. Una alta correlación entre dos categorías indica que es común que ambos tipos de reportes ocurran de forma simultánea, mientras que una baja correlación sugiere que es raro que coincidan. Estas correlaciones son esenciales para identificar patrones y relaciones recurrentes, lo que mejora la precisión de la clasificación. Una alta correlación puede ser aprovechada por el modelo para optimizar la categorización; entre tanto, una baja correlación proporciona diferentes matices que el modelo debe considerar. A continuación, se presentan ejemplos ilustrativos de correlaciones altas y bajas entre categorías. Un ejemplo claro de correlación alta puede ser este reporte:

> Una persona que conocí en línea se ganó mi confianza poco a poco. Yo creí que era importante y accedí a mandarle fotos mías con contenido explícito. Luego me chantajeó diciéndome que, si no le enviaba más fotos, las iba a difundir en internet.

Aunque este ejemplo es ilustrativo y no corresponde a un caso real, ayuda a comprender cómo se manifiesta el problema. Este tipo de situación se clasifica como un caso de sextorsión y *grooming*. En el conjunto de datos del estudio se han identificado 43 casos similares, lo que demuestra una correlación significativa entre *grooming* y sextorsión, y revela la estrecha relación entre ambas categorías.

Un ejemplo de una correlación baja puede ser: "Un individuo me pidió amistad a través de una plataforma de juegos. Luego me persuadió para que le enviara fotos. Luego me di cuenta de que mis imágenes fueron compartidas en un foro público". De nuevo, este es un ejemplo ilustrativo diseñado para ayudar a comprender el problema, no corresponde a un caso real. Este reporte se categoriza bajo *grooming* y distribución de fotos y videos; sin embargo, tales instancias son raras en nuestro conjunto de datos, con solo un caso reportado que muestra esta correlación.

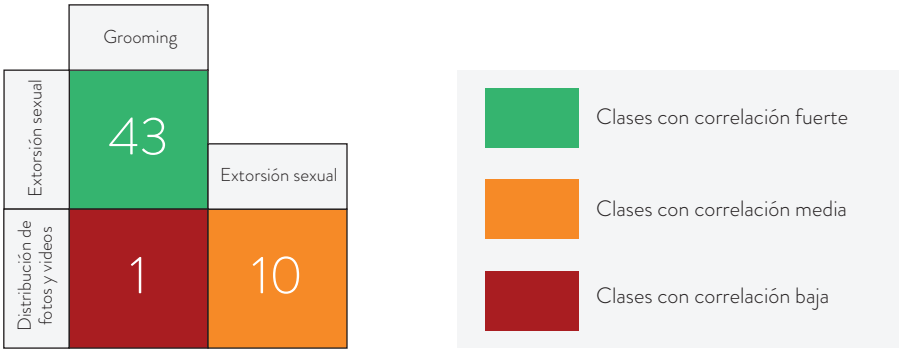


Figura 8.3. Matriz de correlación que ilustra las relaciones entre diferentes categorías de los reportes

Nota: la celda verde destaca las correlaciones altas, por ejemplo, la que se da entre *grooming* y sextorsión. Esto proporciona una mejor comprensión de cómo se da el fenómeno, para la clasificación precisa de reportes.
Fuente: elaboración propia.

La matriz de correlación (figura 8.3), obtenida a partir de estos ejemplos, ofrece perspectivas valiosas sobre las relaciones entre las distintas categorías de los reportes. Una correlación alta, como la que se observa entre *grooming* y sextorsión, sugiere que la presencia de una categoría a menudo coincide con la otra. Por el contrario, las categorías que muestran correlaciones bajas, como *grooming* y distribución de fotos y videos, tienen una menor probabilidad de ocurrir juntas. Estas correlaciones se emplean como características clave en nuestro método, lo que nos permite identificar y explotar estas relaciones para lograr una clasificación precisa de los reportes.

Modelo

La figura 8.4 ilustra la arquitectura general del modelo. Las herramientas desarrolladas utilizan una arquitectura basada en *bidirectional encoder representations from transformers* (BERT), creada por Devlin *et al.* (2019), una tecnología avanzada de IA para el NLP. Se seleccionó BERT debido a su capacidad para capturar el contexto bidireccional de las palabras, pues considera tanto las anteriores como las posteriores, lo que mejora la precisión en tareas como análisis de sentimientos, clasificación de texto y extracción de información. Además, BERT ofrece modelos preentrenados en varios idiomas; en este caso, utilizamos un modelo multilingüe entrenado con grandes corpus de datos no etiquetados, como Wikipedia y libros. Esto le permite aprender representaciones lingüísticas ricas y transferir ese conocimiento a tareas específicas, incluso con pocos datos etiquetados.

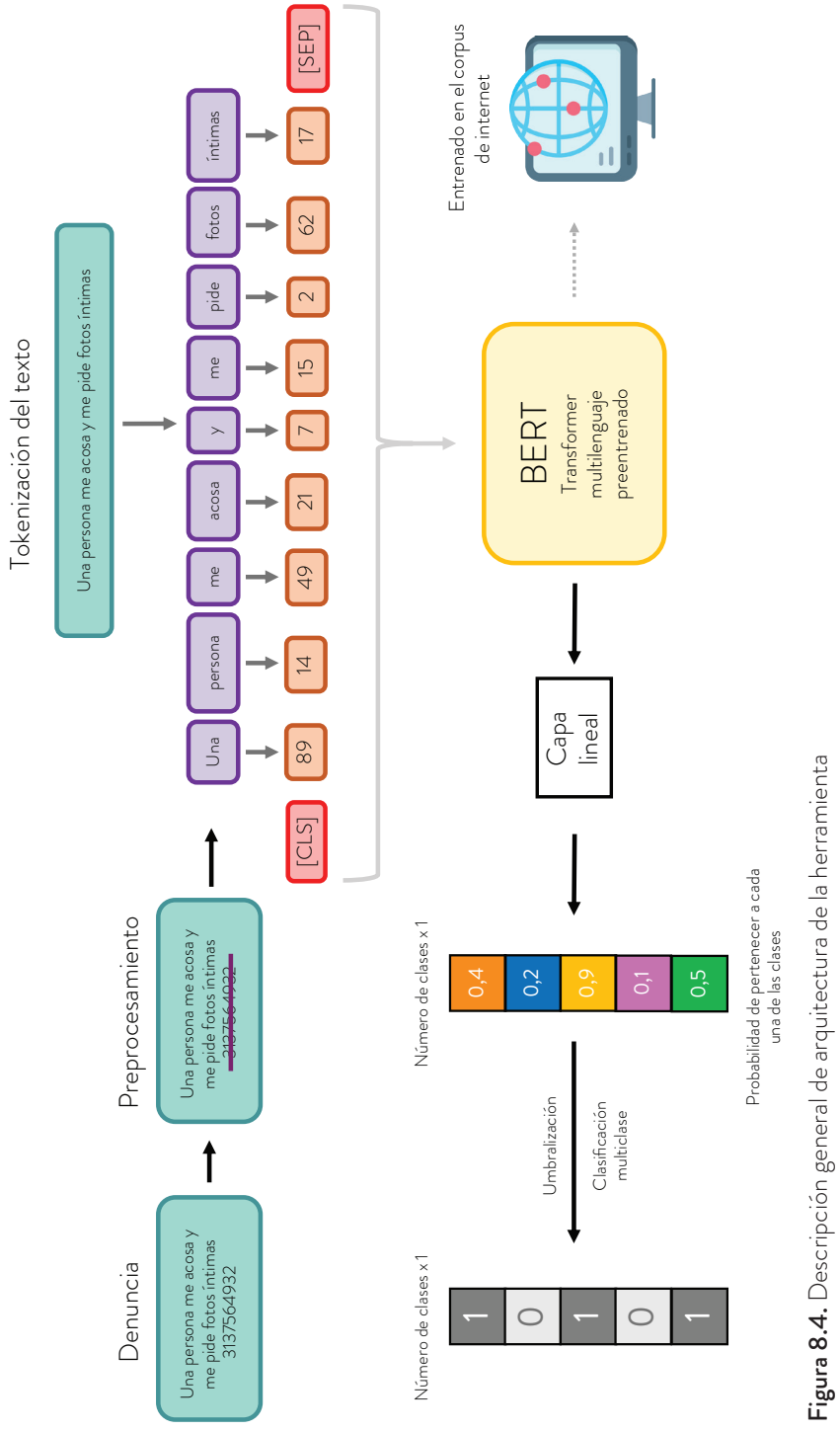


Figura 8.4. Descripción general de arquitectura de la herramienta

Nota: el método utiliza un modelo basado en BERT (Devlin *et al.*, 2019) para producir una predicción a partir de los reportes preprocesados. Luego, proporcionamos una predicción multiclase, según las categorías en cada dimensión.

Fuente: elaboración propia.

BERT es una herramienta altamente versátil y aplicable a diferentes tipos de datos, dada su capacidad de afinarse de forma sencilla para tareas particulares. Al ser un modelo basado en transformadores, también permite capturar relaciones de autoatención, lo que mejora su capacidad para comprender la dependencia entre diferentes partes del texto. Por último, la aplicación bien documentada de BERT en bibliotecas como Hugging Face's Transformers facilita su integración, lo que produce una implementación eficiente y efectiva, aun en proyectos con datos limitados.

Si bien la arquitectura de BERT utilizada en este modelo es similar a la original, las actualizaciones constantes de los modelos preentrenados, que incluyen el uso de conjuntos de datos más grandes y diversos, han mejorado significativamente su rendimiento. En particular, se usó el modelo BERT multilingüe, cuya última actualización fue en el 2023, optimizado para trabajar con múltiples lenguas, incluida el español, lo que permitió mejorar la precisión y cobertura en el procesamiento de nuestros datos.

Las entradas del modelo son reportes preprocesados, para eliminar cualquier dato personal sensible. A continuación, el texto de los reportes se convierte en pequeños fragmentos llamados *tokens*, lo que implica dividir el texto en palabras o partes de palabras, y convertir esos fragmentos en números que nuestro modelo puede entender. El componente principal de la herramienta es el codificador BERT multilingüe, que toma los *tokens* y los analiza, con el objetivo de entender el contexto y el significado de cada palabra en su entorno específico. Es similar a un lector muy inteligente que capta el significado profundo de las palabras, según cómo se usan en una oración.

Una vez que el codificador BERT ha procesado los *tokens*, los resultados pasan a través de varias capas de clasificación que determinan a qué categoría pertenece cada reporte, con base en diferentes dimensiones como el tipo de delito, el grado de criminalidad y el daño causado. De este modo, se clasifican los reportes en categorías como *grooming*, sextorsión y ciberacoso, evaluamos la severidad y la intención del acto delictivo, y analizamos el impacto y el daño producido por el incidente denunciado. Por último, se comparan las predicciones del modelo con las anotaciones de expertos en OCSEA para calcular las métricas de rendimiento, asegurando así la precisión y fiabilidad de la clasificación.

Aumento de datos

Dada la sensibilidad de los reportes, hubo limitaciones en la disponibilidad de datos para su análisis. Para mitigar esta restricción, empleamos la técnica de aumento de datos, que se utiliza en el aprendizaje automático y la IA para incrementar la cantidad y la diversidad de datos disponibles para entrenar modelos, en especial cuando los datos originales son limitados. Esta técnica consiste en generar nuevas muestras de datos alterando las existentes de manera controlada y realista. La figura 8.5 muestra la metodología de aumento de datos, en la que se modifican los reportes originales eliminando palabras con probabilidades variables, lo que crea un conjunto de datos de entrenamiento aumentado y diversificado (véase la figura 8.4). Para cada instancia de denuncia aumentada, dividimos el reporte original en palabras y aplicamos una probabilidad aleatoria, con el fin de decidir si eliminar cada una. Las palabras restantes se combinan para formar un nuevo ejemplo, lo que enriquece nuestra base de datos y amplía la variedad de ejemplos disponibles para nuestro modelo.

Resultados para la herramienta 1

En la tabla 8.1 se presentan los resultados obtenidos del primer modelo en la tarea de clasificación a través de cada dimensión evaluada: asunto, grado de criminalidad y daño. La métrica utilizada para evaluar el desempeño de los modelos es la *precisión promedio*, que mide la eficacia con la que nuestro modelo clasifica según las anotaciones de los expertos. Una métrica de precisión promedio más alta indica un mejor rendimiento del modelo en términos de clasificación.

Antes de describir los resultados, es crucial considerar la probabilidad de clasificar correctamente un ejemplo en todas las clases de manera aleatoria. A medida que aumenta el número de clases, la probabilidad de éxito por azar disminuye, lo que refleja la creciente complejidad de la tarea de clasificación. Esta disminución subraya la dificultad inherente de lograr una alta precisión en un escenario con múltiples clases.

En la tabla 8.1 se muestran dos conjuntos de resultados: aquellos obtenidos por un modelo basado en azar y los logrados por el modelo propuesto. Los resultados del modelo basado en azar se presentan en la primera fila; estos revelan bajas tasas de precisión promedio en todas las dimensiones: asunto (0,39 %), grado de criminalidad (3,125 %) y daño (6,25 %). Estos valores son significativamente bajos, como era de esperarse en un modelo que no utiliza información adicional para la clasificación. En contraste, nuestro modelo muestra un desempeño superior en todas las dimensiones: asunto (38,6 %), grado de criminalidad (41,7 %) y daño (45,6 %). Estos resultados indican que nuestro modelo es capaz

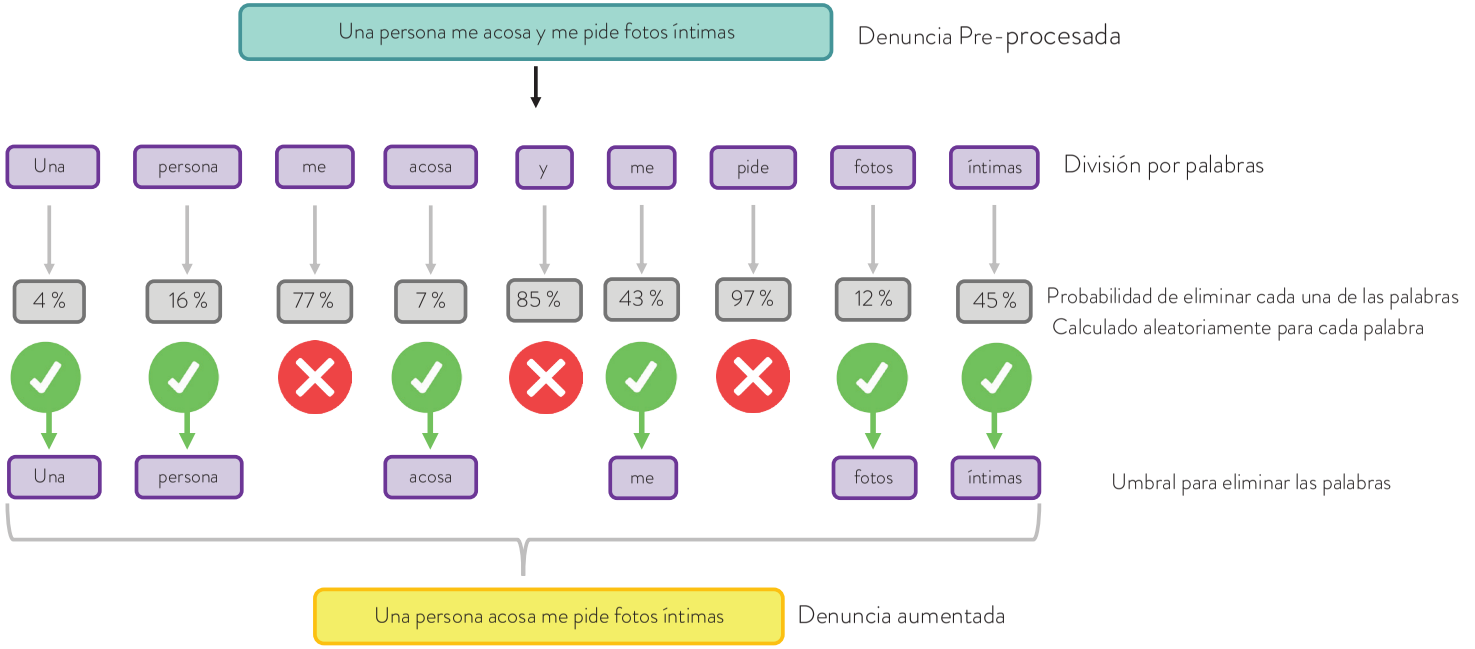


Figura 8.5. Metodología de aumento de datos

Nota: los reportes originales se modificaron al eliminar palabras con probabilidades variables, lo que genera un conjunto aumentado y diverso de datos de entrenamiento.

Fuente: elaboración propia.

Tabla 8.1. Resultados de la tarea de clasificación en cada dimensión evaluada
Se muestra el porcentaje de precisión promedio en cada caso.

Dimensión	Asunto	Grado de criminalidad	Daño
Azar	0,39	3,125	6,25
Nuestro modelo (sin aumento de datos)	38,2	39,7	42,9
Nuestro modelo (con aumento de datos)	**38,6**	**41,7**	**45,6**

Fuente: elaboración propia.

de capturar y modelar de manera efectiva las características subyacentes en los datos de OCSEA, y supera significativamente la clasificación aleatoria. Así mismo, demuestran que nuestro enfoque puede modelar los diversos comportamientos observados en casos de OCSEA de manera efectiva; además, destacan la importancia de disponer de más datos, ya que un mayor volumen de datos, por lo general, conduce a un mejor rendimiento del modelo. Por lo tanto, la adquisición continua de datos adicionales es esencial para mejorar la precisión y efectividad de los modelos, lo cual es fundamental para tener un mayor impacto en los esfuerzos de protección infantil.

Herramienta 2: prototipo de sistema de alertas: propósito, desarrollo técnico y resultados

Al igual que con la primera herramienta, el objetivo principal de este segundo modelo era categorizar de manera efectiva las conversaciones entre agresores relacionadas con OCSEA encontradas en la *deep web*; sin embargo, a diferencia del primer modelo, este no tenía como propósito principal clasificar las conversaciones para redirigirlas a las autoridades. En su lugar, se trataba de un ejercicio de análisis riguroso de un tipo de información que, debido a la dificultad de acceso a los datos y a su naturaleza nociva y traumática, ha sido poco estudiado hasta ahora.

En cada caso, se analizó el tema de la conversación, el tipo de agresor según su nivel de experiencia y las distorsiones cognitivas que mostraba en sus interacciones. Estas categorías permitieron clasificar las conversaciones y comprender la intención de cada uno de ellos. Como resultado, se desarrolló un prototipo de sistema de alerta que podría ser de gran utilidad para las autoridades de justicia y protección infantil.

Correlaciones

Para lograr una clasificación precisa, el modelo emplea una variedad de características diseñadas específicamente para este propósito, considerando las correlaciones observadas entre las diferentes categorías de contenido presentes en dichas conversaciones. Estas correlaciones miden la frecuencia con que dos categorías de contenido aparecen juntas, lo que contribuye a identificar patrones y mejorar la precisión de la clasificación. Una alta correlación puede ser aprovechada por el modelo para mejorar la categorización, mientras que una baja correlación proporciona diferentes matices que el modelo debe considerar.

Modelo

La arquitectura general de este modelo, al igual que la del primero, se basa en BERT (Devlin *et al*., 2019), una tecnología avanzada para el NLP, en su última actualización del 2023. En este caso, en lugar de analizar reportes, la nueva herramienta se centró en examinar conversaciones. El texto fue convertido en *tokens*, los cuales, a través de códigos numéricos, sirvieron para entrenar el modelo. El codificador BERT multilingüe procesa estos *tokens* y analiza el contexto y el significado de cada palabra dentro de su entorno, de manera similar a un analista que comprende a fondo una conversación e interpreta su significado.

Una vez que se procesan los *tokens*, los resultados pasan por sistemas de clasificación que determinan a qué categoría pertenece cada conversación, con base en criterios como el contexto, el tipo de agresor, su nivel de experiencia y las distorsiones cognitivas detectadas. Esto permite clasificar las conversaciones en categorías relevantes, evaluar la severidad e intención de las acciones y analizar su impacto. Por último, se compararon las predicciones del modelo y las anotaciones de expertos en OCSEA, con el objetivo de garantizar la precisión y fiabilidad de la clasificación.

Resultados para la herramienta 2

Los resultados de esta herramienta muestran la capacidad predictiva del modelo de lenguaje para identificar cada una de las dimensiones propuestas para el análisis de estas conversaciones. En particular, la dimensión que estudia el tipo de agresor presenta el mejor desempeño promedio relativo al azar, lo que indica que las conversaciones contienen información relevante que revela características importantes de los agresores. Esta dimensión categoriza a los agresores como expertos (o no) y evalúa si el tono es cínico, y si su enfoque es afectivo o sexual. De estas categorías, el modelo identifica con mayor confianza y precisión

las conversaciones que indican si el agresor es experto o no. Este análisis es interesante, porque permite identificar a los agresores que están compartiendo su conocimiento para entrenar a otros, lo cual es crucial para las intervenciones preventivas y la implementación de medidas de seguridad (tabla 8.2).

Tabla 8.2. Resultados de la tarea de clasificación en cada dimensión evaluada: asunto, tipo de agresor y distorsiones
Se muestra el porcentaje de precisión promedio en cada caso.

Dimensión	Asunto	Contexto	Tipo de agresor	Distorsiones
Azar	0,09	12,5	6,25	12,5
Nuestro método	34,3	47,7	41	34,6

Fuente: elaboración propia.

Por su parte, como se observa en la tabla 8.2, la dimensión de distorsiones es la que menor desempeño promedio presenta, pues es la que tiene menor cantidad de ejemplos. Esta observación subraya la importancia crucial de contar con datos anotados de calidad para llevar a cabo este estudio de manera exitosa; sin una cantidad suficiente de ejemplos bien anotados, es difícil para los modelos de NLP aprender y generalizar correctamente sobre las distintas dimensiones analizadas. Además, es esencial tener personas entrenadas en el proceso de anotación, ya que la precisión y la consistencia de estas anotaciones impactan de forma directa en la eficacia del modelo. Anotadores capacitados pueden identificar y categorizar con precisión las características complejas de las conversaciones, lo que asegura que la base de datos utilizada para entrenar los modelos de NLP sea robusta y fiable. Este nivel de detalle y exactitud es fundamental para desarrollar herramientas predictivas que no solo clasifiquen los datos de forma correcta, sino que también ofrezcan análisis valiosos para la identificación y prevención de comportamientos delictivos en línea.

Discusión

El desarrollo de los dos modelos de IA presentados en este estudio representa un avance significativo en la detección y el análisis de situaciones relacionadas con OCSEA. El primer modelo, centrado en los reportes recibidos por la línea Te Protejo, permitió no solo una categorización más eficiente y precisa de los reportes, sino también una disminución en la exposición directa de los analistas a contenido altamente perturbador. Esto marca un hito importante en el uso

de la IA como una herramienta complementaria para proteger la salud mental de los profesionales involucrados en la prevención y respuesta a delitos de OCSEA.

El segundo modelo, orientado al análisis de conversaciones en la *deep web*, contribuye de manera novedosa al entendimiento del discurso entre agresores, una dimensión del OCSEA que ha sido históricamente poco estudiada, debido a las dificultades para acceder a este tipo de datos. La capacidad de identificar patrones recurrentes, distorsiones cognitivas y tipologías de agresores dentro de estos foros ofrece un potencial invaluable para el diseño de intervenciones preventivas y medidas legales más efectivas; sin embargo, la implementación de este tipo de herramientas enfrenta varios retos, entre ellos la limitación de datos disponibles y la necesidad de mejorar los sistemas de anotación para aumentar la precisión de los modelos.

Algunas reflexiones sobre la inteligencia artificial y el OCSEA

Alrededor del mundo, existen numerosas líneas de reporte (*hotlines*) que forman parte de INHOPE. Estas actúan como canales directos para que los ciudadanos puedan reportar de manera anónima y segura contenidos relacionados, principalmente, con CSAM. Su función es notificar a los proveedores de servicios de internet, con el fin de asegurar la rápida eliminación de contenidos ilegales y, además, denunciar los casos ante las autoridades, garantizando la protección y justicia para niñas, niños y adolescentes. Por lo general, estas líneas de reporte están conformadas por analistas que evalúan la ilegalidad del contenido de acuerdo con las directrices y la legislación internacional establecidas por la Organización Internacional de Policía Criminal (Interpol), además de cumplir con la normativa específica de cada país.

Las líneas de reporte pueden estar a cargo de instituciones gubernamentales, organizaciones sin fines de lucro, departamentos de policía o asociaciones de proveedores de servicios de internet. En Colombia, Te Protejo es la línea nacional de reporte gestionada por Red PaPaz, la cual permite a los ciudadanos informar sobre cualquier situación que ponga en riesgo o vulnere los derechos de niñas, niños y adolescentes. Esta plataforma, completamente digital, está disponible a través de su página web y en una aplicación móvil. Aunque Te Protejo cuenta con un sistema propio para procesar reportes, apoyado por protocolos y rutas de atención que garantizan una gestión adecuada de los casos, el alto volumen de información, la necesidad de proteger a sus analistas y los recursos limitados reducen su capacidad de respuesta.

Por esta razón, el desarrollo de una herramienta como el clasificador supone un avance significativo para esta línea. Por un lado, permite una mayor

objetividad y especialización en el análisis de los casos, ya que minimiza los sesgos cognitivos que podrían influir en la evaluación de los reportes. Esto garantiza que todos los casos sean tratados de manera justa y basada en criterios objetivos, lo cual es esencial para la protección de los derechos de las víctimas. Además, los clasificadores mejoran la capacidad de priorización de las líneas de reporte, lo que posibilita identificar los casos con mayor riesgo y urgencia para dar una respuesta más rápida y adecuada, conforme a los protocolos de las autoridades.

Otra ventaja clave es la reducción significativa en los tiempos de exposición a contenido violento por parte de los analistas. Al automatizar el proceso de clasificación y análisis preliminar, la herramienta ayuda a disminuir la carga emocional y el tiempo que los analistas de Te Protejo deben dedicar a la revisión de reportes relacionados con ciberacoso y violencia sexual.

En el mundo, en especial en Latinoamérica, líneas de reporte como Te Protejo desempeñan un papel fundamental en la lucha contra la violencia sexual en entornos digitales, al asegurar que niñas, niños y adolescentes tengan un medio seguro y anónimo para reportar contenidos ilegales. Estas iniciativas no solo facilitan la eliminación rápida de dicho contenido de internet, sino que también garantizan que los reportes sean gestionados de manera efectiva y oportuna por las autoridades nacionales e internacionales.

Ahora, sin duda, uno de los aspectos más inexplorados en la prevención y manejo del OCSEA es el comportamiento de los agresores. Por esta razón, la información contenida en los foros de la red profunda que fueron analizados en este estudio se convierte en una herramienta clave para comprender y combatir este delito. El prototipo del sistema de alerta desarrollado arrojó resultados muy prometedores. A pesar de contar con una cantidad limitada de datos, logró identificar patrones y características comunes entre los agresores que abusan de niñas, niños y adolescentes en Colombia.

Esta información es crucial para las autoridades en Latinoamérica, ya que proporciona una visión más completa sobre cómo se desarrolla el delito en diferentes países de habla hispana y ofrece insumos valiosos para diseñar políticas públicas orientadas a la prevención. Al entender mejor las intenciones y los comportamientos de los agresores, se pueden formular estrategias más efectivas para prevenir el OCSEA desde su raíz.

Otra de las grandes contribuciones hechas por este ejercicio de investigación fue la creación de un marco metodológico avanzado para el análisis de datos relacionados con el OCSEA en el contexto latinoamericano, y específicamente en español. Este proyecto demuestra la viabilidad y eficacia de utilizar sistemas de anotación e IA que den cuenta de cómo se presenta el fenómeno en

países de habla hispana. También hace eco a la importancia de crear procesos específicos que mejoren la precisión y eficiencia en el análisis de datos complejos y sensibles.

El desarrollo de herramientas basadas en IA para clasificar y analizar reportes también abre nuevas vías para la investigación. Los académicos pueden utilizar estos sistemas para desarrollar nuevos estudios y generar modelos predictivos más sofisticados. Asimismo, estos avances contribuyen a acelerar la creación de mejores sistemas de respuesta para la protección de niñas, niños y adolescentes en contextos como el latinoamericano. Además, presenta resultados promisorios para optimizar los recursos técnicos y humanos disponibles.

Al crear modelos de IA capaces de detectar y analizar casos de OCSEA de manera más precisa y rápida, la capacidad de las familias y comunidades para identificar y responder a situaciones de riesgo también puede mejorar significativamente. Estas herramientas permiten avanzar en el desarrollo de sistemas de alerta temprana de base tecnológica para problemas sociales; en el contexto específico de OCSEA, esto podría prevenir que los riesgos se conviertan en delitos consumados.

El conocimiento y las herramientas generadas mediante este proyecto pueden utilizarse para diseñar e implementar estrategias preventivas y programas educativos dirigidos a madres, padres, educadores y líderes comunitarios. Es posible enfocar estas iniciativas para enseñar a las familias a identificar señales de riesgo y actuar de manera efectiva ante posibles casos de OCSEA. A través de la colaboración con organizaciones locales, escuelas y otras entidades comunitarias, es posible difundir esta información para fortalecer la capacidad de las comunidades en la protección de la niñez.

Por último, el proyecto priorizó la importancia de proteger a los investigadores y analistas de la exposición directa a contenido sensible, al establecer protocolos y sistemas que aseguran su bienestar psicológico y emocional. Este enfoque integral no solo mejora la calidad del análisis de datos, sino que también garantiza que quienes trabajamos en la protección en línea de niñas, niños y adolescentes podamos hacerlo de manera segura y sostenible.

Limitaciones

Pese a los avances significativos logrados con los nuevos sistemas de anotación desarrollados para el proyecto, existen limitaciones que vale la pena mencionar. Una de las principales dificultades radica en la complejidad inherente del fenómeno de la OCSEA. El hecho de que los casos reflejen una variedad de riesgos y características que se superponen hace que la clasificación en múltiples categorías llegue a resultar ambigua. Aunque el nuevo sistema permite clasificar

los casos en más de una categoría, es importante seguir trabajando en clasificaciones que permitan reflejar la complejidad del fenómeno y la naturaleza cambiante de los datos disponibles.

Otro desafío significativo es la precisión de las categorías de anotación. Por ejemplo, las categorizaciones en Atlas incluyen contextos variados y temas generales de conversación, y esa amplitud puede llevar a dificultades en la discriminación precisa entre categorías y, así, afectar la capacidad de la IA para aprender patrones específicos y aplicar el conocimiento de manera efectiva. Por esto, resultó de gran importancia incluir la retroalimentación e iteración para los casos fallidos del modelo de NLP. Esta retroalimentación no solo le permite al equipo de desarrollo elaborar estrategias para que el modelo aprenda mejor, sino que contribuye a afinar las definiciones y categorías de análisis.

Por otra parte, la tipología de agresores y las distorsiones cognitivas propuestas por los expertos son conceptos altamente especializados y en evolución. La clasificación de los participantes en la base de datos Atlas, basada en estas tipologías y distorsiones cognitivas, puede presentar limitaciones causadas por la subjetividad en la interpretación y por la posible evolución de las tácticas de agresión reflejadas en el conjunto de datos. En este sentido, será fundamental garantizar la actualización permanente de las categorías y enfoques, para mantenerse al día con las tendencias emergentes y los últimos hallazgos en la investigación del fenómeno de OCSEA.

Vale la pena aclarar que la implementación de este sistema de anotación también depende en gran medida de la experiencia y precisión de los analistas encargados de categorizar los datos. Aunque incluir a analistas expertos mejoró la calidad de la información, la intervención humana introduce un riesgo de sesgo y variabilidad en la anotación, lo cual afecta la consistencia y la fiabilidad de los datos.

Por otra parte, a pesar de que la protección de los investigadores frente a la exposición directa a los datos sensibles es fundamental, también podría limitar la comprensión profunda y detallada del contenido por parte del equipo. Esta separación puede hacer que algunos matices y contextos importantes sean pasados por alto, lo que afecta la calidad de la interpretación y el análisis realizado por la IA. Dado lo anterior, en fases futuras de esta investigación se buscará involucrar a más analistas en diferentes contextos de América Latina, con el objetivo de balancear el sesgo y ampliar la perspectiva sobre el fenómeno y su complejidad.

Por último, es importante señalar que, aunque las correlaciones observadas en ambos modelos ayudan a mejorar la capacidad predictiva de las herramientas, la disponibilidad y la calidad de los datos sigue siendo un factor limitante.

En especial en América Latina, donde el acceso a tecnologías y recursos es restringido, las herramientas de IA deben contemplarse como una oportunidad para compensar algunas de estas carencias, pero no como una solución completa. La colaboración continua entre expertos en la temática y científicos de datos es crucial para afinar estos modelos y maximizar su impacto en contextos locales específicos.

Conclusiones

Este estudio pone de manifiesto el potencial de la IA para abordar desafíos complejos en la lucha contra el OCSEA, en particular en entornos de datos sensibles y de difícil acceso como la *deep web*. Los resultados de ambos modelos muestran que la IA no solo puede facilitar la clasificación y análisis de incidentes de abuso, sino también reducir el impacto psicológico sobre los profesionales que trabajan directamente con estos casos; sin embargo, el éxito de estas herramientas depende de varios factores, entre ellos la cantidad y calidad de los datos disponibles, y de un sistema de anotación bien estructurado y alineado con las particularidades del fenómeno en contextos latinoamericanos.

La colaboración interdisciplinaria se destaca como un elemento clave para mejorar la precisión de los modelos y diseñar sistemas más robustos que se adapten a las dinámicas locales. Los próximos pasos deben enfocarse en la ampliación de las bases de datos, el perfeccionamiento de los modelos de lenguaje y la implementación de estas herramientas en entornos operativos reales. Asimismo, es esencial que las autoridades y las organizaciones locales comprendan y adopten estas tecnologías como parte de una estrategia más amplia de prevención y lucha contra el OCSEA.

El uso de IA ofrece una promesa significativa para fortalecer la protección de los derechos de niñas, niños y adolescentes frente a la explotación sexual en línea, en especial en regiones como América Latina, donde los recursos son limitados. Este estudio sienta las bases para el desarrollo de futuras aplicaciones de IA orientadas a combatir estos delitos de manera ética, eficiente y segura.

Referencias

Ahern, E. C., Sadler, L. H., Lamb, M. E. y Gariglietti, G. M. (2016). Wellbeing of professionals working with suspected victims of child sexual exploitation. *Child Abuse Review*, *26*(2), 130-140. https://doi.org/10.1002/car.2439

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality And Social Psychology Review*, *3*(3), 193-209. https://doi.org/10.1207/s15327957pspr0303_3

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* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Gangwar, A., González-Castro, V., Alegre, E. y Fidalgo, E. (2021) AttM-CNN: Attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images. *Neurocomputing*, *445*, 81-104. https://doi.org/10.1016/j.neucom.2021.02.056

Guerra, E. y Westlake, B. G. (2021). Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites. *Child Abuse & Neglect*, *122*, 105336. https://doi.org/10.1016/j.chiabu.2021.105336

International Association of Internet Hotlines (INHOPE). (2023). *Save time, save lives: Annual report 2023*. https://aviatorproject.com/media/pages/articles/aviator-annual-report-2023/7ec012052c-1719996764/aviator-annual-report-2023.pdf

International Justice Mission y University of Nottingham Rights Lab. (2023). *Scale of harm research method, findings, and recommendations: Estimating the prevalence of trafficking to produce child sexual exploitation material in the Philippines*. International Justice Mission. https://ijmstoragelive.blob.core.windows.net/ijmna/documents/studies/IJM_Scale_of_Harm_2023_Full_Report_5f292593a9.pdf

Internet Watch Foundation (IWF). (2022, 8 de agosto). *20 000 reports of coerced "self-generated" sexual abuse imagery seen in first half of 2022 show 7- to 10-year-olds.* https://www.iwf.org.uk/news-media/news/20-000-reports-of-coerced-self-generated-sexual-abuse-imagery-seen-in-first-half-of-2022-show-7-to-10-year-olds/

Internet Watch Foundation (IWF). (s. f.). *"Self-generated" child sexual abuse*. https://www.iwf.org.uk/annual-report-2023/trends-and-data/self-generated-child-sex-abuse/

Lorenzo-Dus, N., Mullineux-Morgan, R., Perkins, K., Lacey, S. y Evans, C. (2023). *Developing and evaluating DRAGON Shield. DRAGON: Developing Resistance Against Grooming Online*. Swansea University. https://www.swansea.ac.uk/media/FINAL-DRAGON-Shield-June-2023-report.pdf

National Center for Missing and Exploited Children (NMEC). (2024). *CyberTipline 2023 report*. https://www.missingkids.org/CyberTiplinedata

Ngo, V., Gajula, R., Thorpe, C. y Mckeever, S. (2023). Discovering child sexual abuse material creators. Behaviors and preferences on the dark web. https://arrow.tudublin.ie/scschcomart/201

Steel, C. M., Newman, E., O'Rourke, S. y Quayle, E. (2020). A systematic review of cognitive distortions in online child sexual exploitation material offenders. *Aggression and Violent Behavior*, *51*, 101375. https://doi.org/10.1016/j.avb.2020.101375

Tener, D., Wolak, J. y Finkelhor, D. (2015). A typology of offenders who use online communications to commit sex crimes against minors. *Journal of Aggression, Maltreatment and Trauma*, *24*(3), 319-337. https://doi.org/10.1080/10926771.2015.1009602

van der Bruggen, M. y Blokland, A. (2020). Child sexual exploitation communities on the Darkweb: How organized are they? En M. Weulen Kranenbarg y R. Leukfeldt (Eds.), *Cybercrime in context* (pp. 259-280). Springer. https://doi.org/10.1007/978-3-030-60527-8_15

Wolak, J., Finkelhor, D., Mitchell, K. J. e Ybarra, M. L. (2010). Online "predators" and their victims: Myths, realities, and implications for prevention and treatment. *APA PsycNet*. https://doi.org/10.1037/2152-0828.1.S.13

WeProtect (2023). *Global Threat Assessment 2023. Assessing the scale and scope of child sexual abuse online*. https://www.weprotect.org/global-threat-assessment-23/

III

LA INTELIGENCIA ARTIFICIAL Y EL ESTADO

SISTEMAS DE INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO DE AMÉRICA LATINA Y EL CARIBE

Juan David Gutiérrez, Sarah Muñoz Cadena

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.09

Introducción

Entidades públicas nacionales y subnacionales de casi toda América Latina y el Caribe están adquiriendo, desarrollando y utilizando sistemas de inteligencia artificial (IA) (Gómez Mont *et al*., 2020; Gutiérrez *et al*., 2023; Muñoz-Cadena *et al.* 2025; Organización para la Cooperación y el Desarrollo Económico [OECD] y Banco de Desarrollo de América Latina y el Caribe [CAF], 2022). Estos sistemas tienen dos características comunes pertinentes para el apoyo de las funciones estatales: (1) son sistemas computacionales que operan con cierto grado de autonomía y (2) pueden contribuir a automatizar parte de los procesos de toma de decisiones o apoyar dichos procesos al generar información para quienes hacen políticas públicas[1].

El uso adecuado de estos sistemas permitiría desarrollar los objetivos de la administración pública y facilitar las actividades de todo el ciclo de las políticas públicas. Sin embargo, como se explica más adelante, la aplicación de estos sistemas por parte de las entidades públicas también puede generar diversos riesgos, incluida la violación de los derechos humanos (Gutiérrez y Flórez, 2023; Gutiérrez y Muñoz-Cadena, 2023a).

Por lo tanto, no se puede dar por sentado que la adopción de estos sistemas vaya a favorecer a sus usuarios, a los beneficiarios finales previstos o a la población en general. Tampoco se puede asumir que los beneficios serán similares para

1 Para una discusión sobre qué significa "sistema de IA" y las implicaciones de dicha definición respecto de la gestión y las políticas públicas, véase Gutiérrez (2024e).

diferentes personas o grupos de personas; de hecho, algunas investigaciones han demostrado cómo ciertos sistemas pueden perpetuar o amplificar los prejuicios sobre una población específica. En los Países Bajos, por ejemplo, un sistema cuyo objetivo era la detección precoz del fraude en las prestaciones se convirtió en una herramienta para discriminar a las personas vulnerables (Peeters y Widlak, 2023).

Los potenciales impactos positivos derivados de estos sistemas dependen de distintas variables, como el tipo de tecnología que se utiliza, cómo se desarrolló (tanto el modelo como los datos), cómo se implementa el sistema, la capacidad institucional de la entidad que la adoptó, el contexto socioeconómico y cultural en el que se despliega, entre otras. Para una mejor comprensión de los impactos de los sistemas de IA, este capítulo ilustra cómo distintos sistemas de IA adoptados por diferentes entidades públicas en América Latina y el Caribe podrían contribuir con la consecución de diversos fines estatales.

La literatura sobre la aplicación de herramientas de IA en el sector público es incipiente. Por ejemplo, Tangi *et al*. (2024a) sostienen que una agenda de investigación sobre el uso de los sistemas de IA en las instituciones públicas debería explorar casos de uso de forma estructurada, para enriquecer nuestra comprensión del proceso de adopción de estas herramientas y la transformación de las entidades. Además, estudios recientes han tratado de identificar cómo la adopción de diversas innovaciones tecnológicas, como la IA, puede contribuir o no a la consecución de los objetivos gubernamentales, por ejemplo, en el marco del ciclo de las políticas públicas (Höchtl *et al*., 2016; Pencheva *et al*., 2020; Valle-Cruz *et al*., 2020). Sin embargo, la mayor parte de la literatura se centra en los sistemas de IA implementados en el sector público de los países del norte global.

Estudiar cómo se están adoptando y aplicando los sistemas de IA en los países del sur global es esencial, entre otras cosas, porque, como Arun (2020) ha señalado, los contextos particulares y diversos de estos países entrañan ciertos riesgos menos probables en sus homólogos del norte. Algunas de las preocupaciones mencionadas por esta autora son: (1) los sistemas diseñados para los países del norte global podrían importarse sin tener en cuenta los contextos particulares y diversos en los que se implementarán, por ejemplo, países en los que la cobertura de la red de internet es insuficiente o, incluso, hay cortes de electricidad; (2) el riesgo de que se introduzcan sistemas que utilicen "el reconocimiento facial, los drones y otras formas de vigilancia para oprimir a las poblaciones vulnerables" (p. 592); y (3) la inexistencia de leyes o reglamentos que protejan los datos y los derechos de los ciudadanos.

Este capítulo contribuye a la literatura al examinar cómo los sistemas de IA son adoptados por entidades públicas de América Latina y el Caribe. Con tal

fin, en el marco del proyecto Sistemas de Algoritmos Públicos de la Escuela de Gobierno, de la Universidad de los Andes, construimos una nueva base de datos que documenta 735 sistemas de IA piloteados o desplegados en el sector público de 25 países de la región, Puerto Rico y la Organización de los Estados Americanos[2].

Además, el capítulo estudia cómo las herramientas de IA se integran en actividades asociadas a las principales etapas del ciclo de las políticas públicas: agendamiento, formulación, implementación y evaluación. Para cada etapa del ciclo de las políticas públicas, mostramos diferentes tipos de tecnologías y técnicas y describimos cómo se utilizan los sistemas de IA para apoyar diferentes funciones y sectores gubernamentales. En relación con cada sistema de IA identificado, informamos su nombre, la entidad pública que lo adoptó, sus objetivos, cómo funciona (incluido el tipo de datos que utiliza) y quiénes son sus beneficiarios previstos. Es importante advertir que en nuestra clasificación de los sistemas como herramientas que contribuyen a las etapas del ciclo de las políticas públicas es posible que un sistema aporte a más de una de las etapas del ciclo.

La siguiente sección ofrece un breve repaso del marco teórico de los sistemas de IA, los retos de su adopción en el sector público y el ciclo de las políticas públicas; finaliza con una exposición de la metodología empleada en nuestra investigación, tanto para la construcción de la nueva base de datos de sistemas de IA del sector público de América Latina y el Caribe como para los casos de estudio que abordamos (dieciséis herramientas de IA). Luego, en la siguiente sección presentamos estadísticas descriptivas sobre los sistemas de IA mapeados en nuestra nueva base de datos. Posteriormente, se aborda cada una de las etapas del ciclo de las políticas públicas con los respectivos ejemplos de los sistemas de IA. Por último, el capítulo cierra con observaciones sobre las implicaciones políticas que pueden derivarse de este estudio y una reflexión sobre futuras vías de investigación.

2 Sistemas de Algoritmos Públicos es un proyecto académico interdisciplinario de la Escuela de Gobierno de la Universidad de los Andes. Este contribuye al conocimiento sobre los sistemas algorítmicos del sector público de América Latina y el Caribe, así como a la gobernanza de estas herramientas en nuestra región. Para más información sobre el proyecto, véase su portal web: https://algoritmos.uniandes.edu.co/

Marco conceptual, revisión de literatura y metodología

Sistemas de inteligencia artificial en el sector público

No hay consenso sobre una definición única de IA, por una parte, porque varían y se adaptan a medida que evolucionan los avances y las aplicaciones en este campo y, por otra, porque el significado del término *inteligencia* también es controvertido (Crawford, 2021; Mitchell, 2019; Russell y Norvig, 2004; Wirtz y Müller, 2019). En este capítulo, entendemos un sistema de IA como sistemas computacionales

> diseñados por humanos que, dado un objetivo complejo, actúan en la dimensión física o digital percibiendo su entorno mediante la adquisición de datos, interpretando los datos recogidos, estructurados o no, razonando sobre el conocimiento, o procesando la información, derivada de estos datos y decidiendo la(s) mejor(es) acción(es) a tomar para alcanzar el objetivo dado. (High-Level Expert Group on Artificial Intelligence [AI HLEG], 2020, p. 24)

Además, los sistemas de IA pueden diferenciarse por las técnicas y tecnologías utilizadas para desarrollarlos, como el aprendizaje automático, el aprendizaje profundo, las redes neuronales, el procesamiento de lenguaje natural, los sistemas expertos basados en reglas, la automatización robótica de procesos y los robots (Benbya *et al*., 2020)[3].

La adopción de sistemas de IA y otras tecnologías algorítmicas en el sector público se está extendiendo por todo el mundo (Ada Lovelace Institute *et al*., 2021; Global Partnership on Artificial Intelligence [GPAI], 2024; Muñoz-Cadena *et al*., 2025; Zuiderwijk *et al*., 2021). Aun así, la bibliografía sobre la relación entre la IA y las políticas públicas es escasa (Misuraca *et al*., 2020; Valle-Cruz *et al*., 2020). En general, aunque el entusiasmo generado por la introducción de nuevas tecnologías en el sector público podría ayudar a llevar a cabo procesos específicos de forma más eficiente y eficaz, mejorar la prestación de servicios e incluso aumentar la confianza en el Gobierno, también es cierto que la academia ha reconocido ciertos retos y problemas que han surgido de estas implementaciones (Chenou y Rodríguez Valenzuela, 2021; Gutiérrez, 2020; Misuraca *et al*., 2020).

3 Como Benbya *et al*. (2020) explican, la IA también pueden clasificarse en función del tipo de inteligencia que despliegan (estrecha, general y superinteligencia) o en función de la función que desempeñan (conversacional, biométrica, algorítmica y robótica).

Para que el despliegue de las herramientas contribuya con los objetivos perseguidos por las entidades públicas, es preciso que estas sean idóneas (que su funcionalidad sea compatible con el fin deseado), que funcionen suficientemente bien para todos los beneficiarios esperados (de lo contrario podría incrementar desigualdades sociales) y que los nuevos sistemas que han sido piloteados en ambientes controlados puedan escalarse con éxito en contextos reales (Parker y Davies, 2024).

Por otra parte, como mencionan Valle-Cruz *et al.* (2020), lo retos de la inclusión de herramientas de IA en los procesos de políticas públicas "están relacionados con cuestiones legales y morales, principios éticos, legitimidad, transparencia, sustitución de mano de obra a través de la automatización inteligente, predicciones y toma de decisiones" (p. 4). Por ejemplo, es posible que los algoritmos exacerben la desigualdad social, aumenten la discriminación contra determinados grupos de población y muestren información sesgada. Más concretamente, la toma de decisiones con apoyo de herramientas de IA entrenadas a partir de bases de datos desbalanceadas puede generar discriminación sistemática en contra de personas por razones de género, raza, origen étnico o condición socioeconómica (Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos [OHCHR], 2023; Slattery *et al.*, 2024).

El uso de sistemas de IA entrenados con datos personales y que operan con ellos, por ejemplo, genera el riesgo de un acceso no autorizado por terceros, si los Gobiernos no disponen de las salvaguardas necesarias para la seguridad de la información. De hecho, el despliegue de una herramienta con deficiencias de ciberseguridad puede dar lugar a una violación masiva de los derechos fundamentales, como el derecho a la privacidad y el derecho a la protección de los datos personales (Bundesamt für Sicherheit in der Informationstechnik [BSI], 2024).

Además, el uso de herramientas de IA que funcionan con un alto grado de opacidad en el contexto de investigaciones administrativas o judiciales con apoyo de herramientas podría vulnerar los derechos de defensa y debido proceso del investigado (Gutiérrez, 2020, 2024a; Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [Unesco], 2023). Por ejemplo, en un estudio sobre cómo es posible que la automatización erosione el derecho al debido proceso en los Estados Unidos, Citron (2008) argumenta que algunos sistemas de toma automatizada de decisiones en la administración pública "adjudican en secreto, mientras que otros carecen de registros de auditoría, lo que imposibilita la revisión de la ley y los hechos en los que se basan las decisiones del sistema" (p. 1253).

Por último, Calo y Citron (2021) también cuestionan la legitimidad de automatizar procesos de decisión por parte de agencias administrativas, en tanto ello

equivale a abdicar la aplicación de la experiencia de los tomadores de decisión y la posibilidad de que estos ejerzan sus funciones discrecionalmente, dos de los fundamentos que justifican la existencia de este tipo de organizaciones estatales; así mismo, que los litigios en los Estados Unidos "en diversos contextos administrativos ha revelado un patrón común: las agencias no comprenden ni pueden controlar las máquinas en las que han delegado su autoridad" (p. 818).

El ciclo de la política pública y los sistemas de inteligencia artificial

La idea del ciclo de las políticas públicas sirve para simplificar y desglosar la política pública, mediante un modelo que incluye una serie de etapas o pasos que facilitan su análisis y comprensión. Dicho análisis también se conoce como *enfoque de libro de texto* (Nakamura, 1987), porque el ciclo de las políticas públicas se utiliza como herramienta para la enseñanza de la materia.

Una de las ventajas de utilizar el ciclo es que "el modelo es lo suficientemente general como para permitir su utilización para cualquier política" (Roth Deubel, 2002, p. 49). En esta línea, Pencheva *et al*. (2020) sostienen que utilizar el ciclo de las políticas públicas como "modelo analítico" para explorar el papel de las tecnologías emergentes en el sector público es útil, porque ofrece "una plantilla básica" para orientar y enmarcar los debates sobre un tema que no se ha explorado lo suficiente.

Este modelo tiene sus críticas; entre las más recurrentes se encuentran: (1) el modelo puede generar la idea de que los procesos de las políticas públicas son rígidos y lineales; (2) no proporciona información sobre el tiempo o los requisitos que son necesarios para llevar a cabo las actividades asociadas a cada etapa; (3) no ofrece perspectivas sobre otros procesos que pueden afectarla, como el ciclo electoral; y (4) se acerca más a un enfoque político de *arriba hacia abajo* en la formulación de políticas (Gutiérrez *et al*., 2024; Hupe y Hill, 2015; Torres-Melo y Santander, 2013).

A pesar de las críticas al ciclo de las políticas públicas, este modelo es utilizado y enseñado, ya que permite "simplifica[r] las políticas públicas de una manera que resulta práctica para su estudio" (Torres-Melo y Santander, 2013, p. 67). Aunque existen diferentes modelos del ciclo de las políticas públicas, este capítulo se referirá a una versión sintetizada en cuatro etapas: agendamiento, formulación, implementación y evaluación.

La primera etapa del ciclo de las políticas públicas es la fijación de la agenda gubernamental o institucional, que consiste en el proceso de determinar qué problema o problemas públicos captan la atención del Gobierno (Gutiérrez

et al., 2024). Las herramientas tecnológicas pueden generar insumos o información para los Gobiernos sobre un tema y ayudarles a decidir si debe entrar en la agenda institucional. Uno de los usos de los sistemas de IA en el sector público es asistir a los funcionarios públicos para identificar de forma más oportuna y rápida qué tipo de cuestiones preocupan a los ciudadanos e incluso qué tipo de soluciones son mejor percibidas por la población (Pencheva *et al*., 2020; Valle-Cruz *et al*., 2020; van Noordt y Misuraca, 2022). Por otro lado, esta información podría contribuir a anticipar el surgimiento de problemas a medio y largo plazo, mediante la recopilación de datos y el seguimiento de diversas variables (socioeconómicas, medioambientales, entre otras).

En la etapa de formulación, los hacedores de política pública establecen objetivos y diseñan un curso o plan de acción basado en el diagnóstico de un problema. Esto requiere, idealmente, emprender actividades que apunten a (1) estructurar el problema, (2) diseñar alternativas de política pública y (3) priorizar las alternativas políticas y planificar (Torres-Melo y Santander, 2013). En esta fase del ciclo es posible que participen distintas partes interesadas, estatales y no estatales, que intentarán promover sus soluciones a los problemas públicos (Howlett y Cashore, 2014).

Los sistemas de IA pueden adoptarse en esta fase del ciclo para estructurar el problema público, diseñar alternativas y priorizar las intervenciones. La información obtenida a través de estos sistemas ayuda a legitimar una intervención determinada o decidir qué intervención, entre varias opciones, es la más popular entre los ciudadanos (Höchtl *et al*., 2016). Sin embargo, uno de los riesgos es que la información, obtenida a través de un sistema de IA que sirve de punto de referencia para formular un problema de política pública, pueda estar sesgada o ser inexacta y, por tanto, conduzca a una interpretación errónea del problema.

La etapa de implementación implica traducir el curso de acción, definido en la etapa de formulación, en pasos concretos y en la utilización de instrumentos de política pública como la ejecución de programas sociales o proyectos de inversión pública (Gutiérrez *et al*., 2024). En esta etapa, es posible utilizar los sistemas de IA para apoyar diferentes actividades, como la supervisión de la aplicación mediante la obtención y el procesamiento de información en tiempo real y la mejora de la capacidad de las entidades públicas para detectar patrones y anomalías. Además, los sistemas de IA también permiten a las entidades públicas realizar tareas o llegar a lugares a los que en otras circunstancias no podrían, así como reducir los costos de ejecución de estas labores (Valle-Cruz *et al*., 2020).

Idealmente, la última etapa del ciclo de la política pública es la evaluación de los efectos o impactos de las intervenciones. Para evaluar una política pública, es necesario analizar si se lograron los objetivos, dado un problema público

que se identificó en la primera etapa (agendamiento) y para el que se formuló un curso de acción (segunda etapa) que ya se implementó (tercera etapa) (Gutiérrez *et al*., 2024; Knill y Tosun, 2011). Además, la evaluación de las políticas públicas proporciona una retroalimentación para futuros cambios en la política pública y puede ser un mecanismo que promueva la rendición de cuentas del Gobierno (Arellano Gault y Blanco, 2020). A través de herramientas de IA es posible ayudar a identificar patrones y cambios, por ejemplo, en los espacios públicos, que permitan determinar el impacto potencial de una intervención política.

Sin embargo, la evaluación de políticas no se circunscribe a la identificación de sus efectos o impactos, sino que también abarca la valoración de las actividades, procesos, productos y resultados de la política implementada (Gutiérrez y Muñoz-Cadena, 2023b). Por lo tanto, para la evaluación de políticas se requiere recopilar información significativa, idealmente antes, durante y después de la intervención. En este sentido, una de las principales funciones de los sistemas de IA es ayudar a supervisar cómo progresa una política, lo que facilita las primeras evaluaciones de sus logros (Valle-Cruz *et al*., 2020; van Noordt y Misuraca, 2022).

La bibliografía sobre la relación entre la IA y el ciclo de las políticas públicas es incipiente. Sin embargo, algunos artículos recientes han explorado las razones por las cuales el enfoque del ciclo de las políticas públicas es útil como referencia para analizar la aplicación de las nuevas innovaciones tecnológicas en el sector público, sus limitaciones y los riesgos en cada una de las fases del ciclo, al igual que para documentar cómo pueden contribuir los sistemas específicos de IA en cada fase del ciclo[4]. Valle-Cruz *et al*. (2020) utilizaron como punto de referencia el ciclo de las políticas públicas para estudiar las implicaciones de cómo las entidades públicas emplean la IA para diversos medios y ofrecieron ejemplos de sistemas implementados en países del norte global para cada fase del ciclo.

En lo que respecta a América Latina y el Caribe, Gutiérrez *et al*. (2023) crearon una base de datos que en su primera versión del 2023 mapeó 111 sistemas de toma automatizada de decisiones (en particular sistemas de IA) piloteados o implementados en 51 entidades del sector público colombiano. Entre las cuarenta variables utilizadas para caracterizar los sistemas, los investigadores incluyeron las etapas del ciclo de las políticas públicas a las que podía apoyar eventualmente cada sistema. Gutiérrez y Muñoz-Cadena (2023a) informaron

4 El enfoque del ciclo de las políticas públicas no es la única forma de analizar los usos potenciales de la IA en el sector público. Por ejemplo, van Noordt y Misuraca (2022) utilizan una categorización según tres funciones gubernamentales: diseño de políticas, servicios públicos y gestión interna.

que el 12 % de los sistemas de esta base de datos podían contribuir al agendamiento, el 18 % a la formulación, el 98 % a la implementación y solo el 2 % a la evaluación de las políticas públicas.

Luego, Gutiérrez *et al.* (2025) ampliaron dicha base de datos, al mapear 400 sistemas de toma automatizada de decisiones en 171 entidades públicas colombianas. La base de datos documenta 306 sistemas en etapa de ejecución, 49 en pilotaje, 27 suspendidos y 18 descontinuados. De los 355 sistemas piloteados o implementados, 225 son herramientas de IA y el remanente corresponde a sistemas de automatización robótica de procesos. De estos sistemas de IA, el 20 % podrían contribuir en la fase de agendamiento, el 24 % a la formulación, el 99 % a la implementación y el 15 % a la evaluación de las políticas públicas.

De forma más general, algunos artículos académicos e informes de organizaciones de la sociedad civil han trazado el mapa de los sistemas de IA utilizados por el sector público en el norte global y han analizado sus implicaciones para sus ciudadanos. Por ejemplo, los informes publicados por AlgorithmWatch y Bertelsmann Stiftung (2019, 2020) y los artículos de van Noordt y Misuraca (2022), Medaglia y Tangi (2022), y Tangi *et al.* (2022) examinaron casos en la Unión Europea y Brauneis y Goodman (2018) exploraron ejemplos de la transparencia algorítmica de los Gobiernos en la aplicación de estos sistemas en los Estados Unidos. Es pertinente mencionar que el informe de GPAI (2024) sobre instrumentos de transparencia algorítmica en el sector público mapeó ochenta repositorios de algoritmos públicos creados por entidades supranacionales, nacionales, subnacionales, organizaciones de la sociedad civil y universidades; este documenta cuántos sistemas (que en gran parte usan IA) se han registrado en cada uno de estos repositorios.[5]

Nueva base de datos y casos de estudio sobre herramientas de inteligencia artificial implementadas por el sector público en América Latina y el Caribe

Construimos una nueva base de datos que documenta 735 sistemas de IA desarrollados o adoptados por entidades públicas en 25 países de la región, Puerto Rico y la Organización de los Estados Americanos (figura 9.1)[6]. En la base de

5 Para un mapeo más actualizado de 83 repositorios de algoritmos públicos en 15 países de América, Asia, Europa, y Oceanía, véase la base de datos de Gutiérrez y Muñoz-Cadena (2025).

6 También incluye un sistema (*chatbot*) implementado en varios países de Centroamérica por el Programa de las Naciones Unidas para el Desarrollo (PNUD) en América Latina y el

datos, cada sistema de IA es caracterizado a partir de ocho variables (en la siguiente sección presentamos las correspondientes estadísticas descriptivas)[7].

La caracterización de los sistemas de IA mapeados incluye información de sistemas registrados en los repositorios de Argentina, Brasil, Colombia, México, Chile y Uruguay; informes publicados por entidades multilaterales (Banco Interamericano de Desarrollo [BID] y Unesco, s. f.; CAF, 2021a; OECD y CAF, 2022; Unesco, 2021); artículos académicos (Chenou y Rodríguez Valenzuela, 2021; Gutiérrez y Castellanos-Sánchez, 2023; Gutiérrez y Muñoz-Cadena, 2023a, 2025; Hermosilla y Germán, 2024) e información disponible en las páginas web de las entidades públicas y organizaciones de la sociedad civil (AISur, s. f.; Dejusticia, 2022).

Además, en este capítulo ilustramos la adopción de herramientas de IA en el sector público de la región a partir de dieciséis casos en Argentina, Brasil, Chile, Colombia, Guatemala, Honduras, México y Perú. Para la selección de los casos que se describen, tuvimos en cuenta tres criterios: (1) el tipo de funciones y tareas que son realizados por las entidades públicas con los sistemas de IA, (2) los países que los adoptaron y (3) la disponibilidad de información sobre su implementación.

Los dos primeros criterios de selección apuntaban a lograr diversidad en los casos de estudio. Por tal motivo, escogimos una variedad de sistemas que ilustran diversas tecnologías (aprendizaje automático, procesamiento del lenguaje natural, visión por computador, etc.), las cuales han sido aplicadas en diferentes sectores gubernamentales (salud, educación, medio ambiente, movilidad, etc.) de ocho países de la región. El tercer criterio de selección buscaba que los casos escogidos consistieran en herramientas de IA efectivamente implementadas por los Estados (por oposición a meros anuncios de futura adopción), de las cuales pudiéramos contar con información básica acerca de cómo funcionan y en qué tipo de tareas son integradas.

Caribe, en alianza con la Agencia para el Desarrollo Internacional de los Estados Unidos (USAID). El *chatbot* Sara (Sistema de Atención y Respuesta Automatizada) se ha implementado en algunos países de Centroamérica, con el fin de brindar información y orientación a mujeres, niñas y adolescentes en riesgo de sufrir violencia de género e intrafamiliar.

7 Las ocho variables que caracterizan a cada sistema de IA son las siguientes: país, nombre del sistema, si usa datos personales, Clasificación de Funciones de Gobierno (COFOG), potencial aporte a objetivos de desarrollo sostenible (ODS), palabras clave que describen sus funcionalidades o técnicas con las cuales fueron desarrolladas, aporte a procesos de gobierno (clasificación nivel I JRC-UE) y tipo de interacción (entre Gobierno y ciudadanos [G2C], Gobierno y negocios [G2B] y Gobierno y Gobierno [G2G]).

Estadísticas sobre el uso de sistemas de inteligencia artificial en el sector público de América Latina y el Caribe

En Latinoamérica existen repositorios o registros en línea de algoritmos públicos en Argentina, Brasil, Colombia, Chile, México y Uruguay (GPAI, 2024). Dichos repositorios han sido creados por entidades públicas, académicos y organizaciones de la sociedad civil, y documentan los sistemas de IA adoptados por entidades públicas de dichos países. Además, en Perú se está realizando investigaciones para crear un registro similar. A partir de la información disponible en estos repositorios, documentos académicos, informes publicados por entidades multilaterales como el BID, CAF u OCDE, e información disponible en las páginas web de entidades públicas, organizaciones de la sociedad civil y empresas, hemos encontrado que en la última década las entidades públicas de los países latinoamericanos han piloteado o implementado al menos 735 sistemas de IA.

Los 735 sistemas identificados se encuentran en Colombia (254), Brasil (112), México (103), Chile (69), Argentina (46), Uruguay (36), Perú (26), Ecuador (19), Panamá, (12), Costa Rica (8), República Dominicana (7), El Salvador (6), Paraguay (5), Puerto Rico (5), Belice (4), Honduras (4), Guatemala (3), Jamaica (3), Bahamas (2), Venezuela (1), Guyana (2), Barbados (1), Bolivia (1), Granada (1), Nicaragua (1) y Trinidad y Tobago (1) (figura 9.1).

Tomando como punto de referencia la Clasificación de las Funciones del Gobierno (COFOG) de las Naciones Unidas, encontramos que, del total de 735 sistemas, 228 son implementados por entidades públicas que se clasifican en la categoría servicios públicos generales, 193 en asuntos económicos, 127 en orden público y seguridad, 66 en salud, 56 en educación, y los restantes sistemas están relacionados con cinco categorías: protección social; protección del medio ambiente; actividades recreativas, cultura, deportes y otros servicios sociales; vivienda y servicios conexos; y defensa (figura 9.2).

Ahora bien, al utilizar palabras clave para identificar y clasificar a los sistemas[8], encontramos que 148 corresponden a *chatbot*, 117 a detección de anomalías, 104 a reconocimiento facial, 92 a sistemas de predicción, 79 a reconocimiento de objetos, 64 a reconocimiento de patrones, 52 a sistemas de recomendación, 44 a clasificación, 18 a reconocimiento de sonidos, 12 a otros sistemas de soporte de interacción humano-máquina y 5 a un *voicebot* (figura 9.3).

8 Las palabras clave se seleccionaron a partir de aquellas que se utilizan en el repositorio Public Sector Tech Watch del Joint Research Centre (JRC), de la Comisión Europea de la Unión Europea, para clasificar los sistemas de IA.

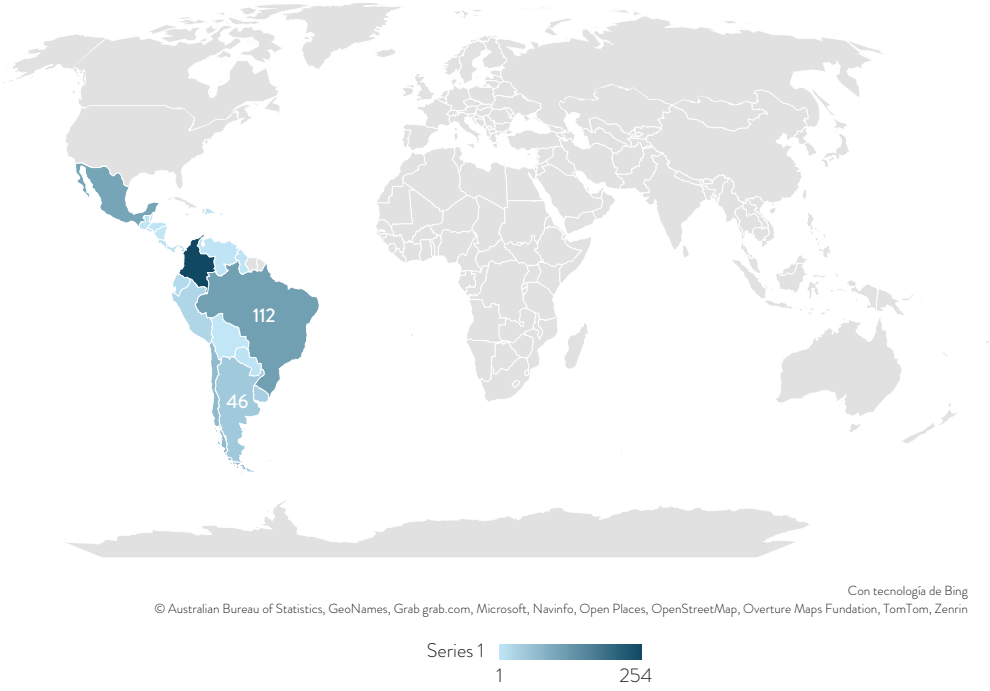


Figura 9.1. Cantidad de sistemas con IA que se utilizan en entidades del sector público de América Latina y el Caribe

Fuente: elaboración propia a partir de herramienta de Microsoft.

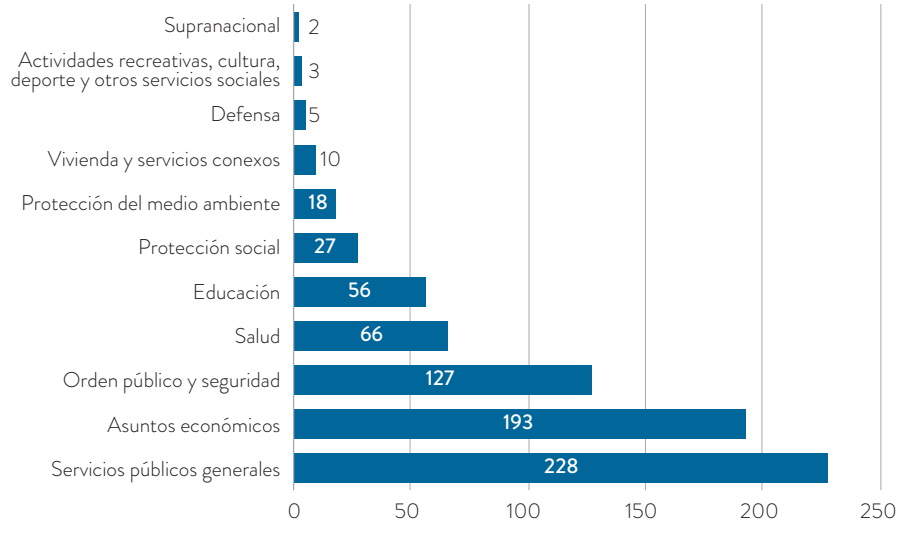


Figura 9.2. Tipos de entidades públicas que utilizan IA según la clasificación COFOG

Fuente: elaboración propia.

Por otra parte, identificamos que el 54 % de los sistemas utilizan datos personales de los usuarios y el 29 % no lo hace; entretanto, para el 17 % restante no es posible determinar si hace o no uso de datos personales con la información recolectada (figura 9.4). Como se comenta más adelante, este dato resalta que los Estados deben adoptar medidas adicionales de seguridad (administrativas, técnicas y humanas) para proteger la seguridad de los datos personales tratados para desarrollar, desplegar o usar los sistemas de IA.

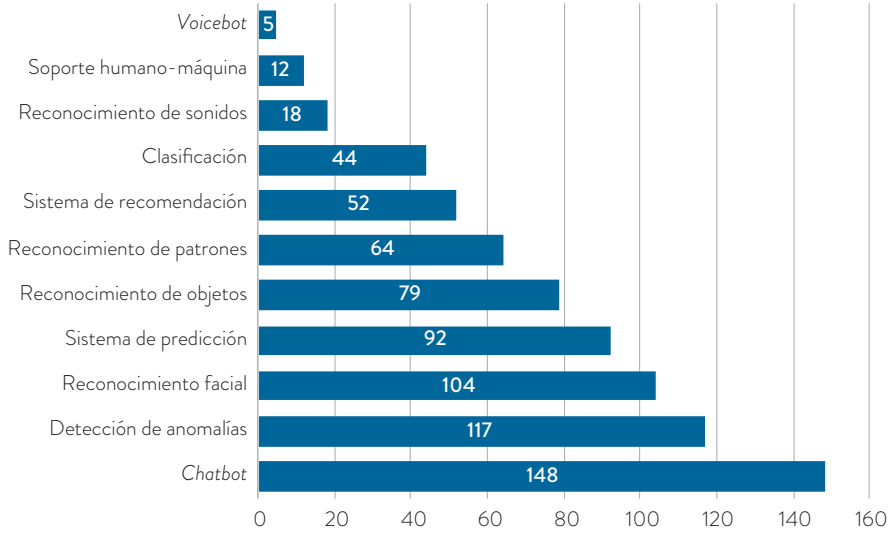


Figura 9.3. Tipos de sistemas según clasificación por palabras clave
Fuente: elaboración propia.

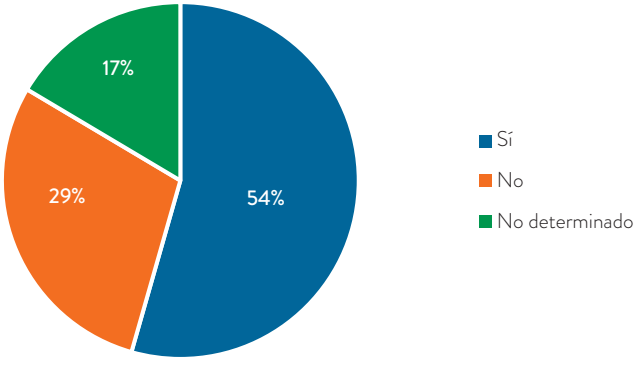


Figura 9.4. ¿Los sistemas utilizan datos personales?
Fuente: elaboración propia.

Los sistemas de IA pueden generar interacciones entre diferentes grupos de interés. Utilizando la clasificación propuesta en el documento "Methodology for the public sector Tech Watch use case collection: Taxonomy, data collection, and use case analysis procedures" (Tangi *et al.*, 2024a), se catalogaron cada uno de los sistemas si se da una interacción entre Gobierno y ciudadanos (G2C), Gobierno y negocios (G2B) y Gobierno y Gobierno (G2G), reconociendo que los sistemas pueden generar más de una interacción. Así, se observa que 547 sistemas generan interacciones entre G2C; 253 entre G2G, y 70 entre G2B (figura 9.5). Esto sugiere que el principal usuario de estos sistemas de IA son los propios funcionarios de las entidades públicas, quienes utilizan las herramientas para apoyar su gestión, mientras que en un porcentaje más bajo los usuarios y beneficiarios directos de los sistemas de IA son los ciudadanos y las empresas.

Con respecto a los posibles aportes de los sistemas de IA, los organizamos a partir de dos clasificaciones. Por un lado, tomamos como referencia la clasificación de aportes a los procesos de gobierno propuesta por Engstrom *et al.* (2020), utilizada también en el repositorio Public Sector Tech Watch, para clasificar los sistemas de IA (Tangi *et al.*, 2024a). Como se observa en la figura 9.6, 304 de los sistemas podrían aportarle a la categoría de cumplimiento de la ley; 180 a servicios públicos y participación; 143 a análisis, monitoreo e investigación de política pública; 94 a la gestión interna de procesos; y 14 al otorgamiento de beneficios.

Por otro lado, también se clasificaron los sistemas considerando su posible aporte a los objetivos de desarrollo sostenible (ODS); en esta clasificación es importante aclarar que un sistema puede aportar a más de un ODS. Como

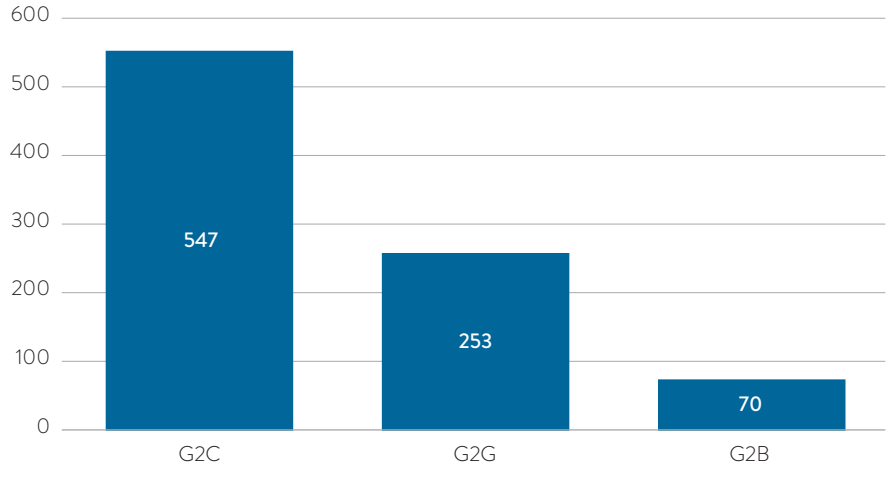


Figura 9.5. Tipo de interacción que ofrece el sistema
Fuente: elaboración propia.

se observa en la figura 9.7, los tres ODS respecto de los cuales más sistemas de IA podrían contribuir en su avance son el 16 (Paz, Justicia e Instituciones Sólidas), el 9 (Industria, Innovación e Infraestructura), y el 3 (Salud y Bienestar).

En las siguientes secciones de este capítulo presentaremos en detalle dieciséis de estos sistemas y explicaremos cómo podrían apoyar a las entidades públicas en las diferentes etapas del ciclo de las políticas públicas.

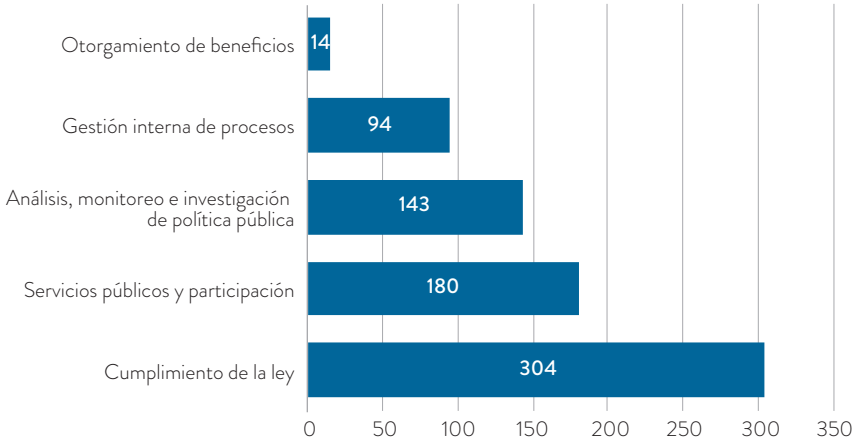


Figura 9.6. Posible aporte de los sistemas a los procesos de gobierno, según clasificación de la Unión Europea

Fuente: elaboración propia.

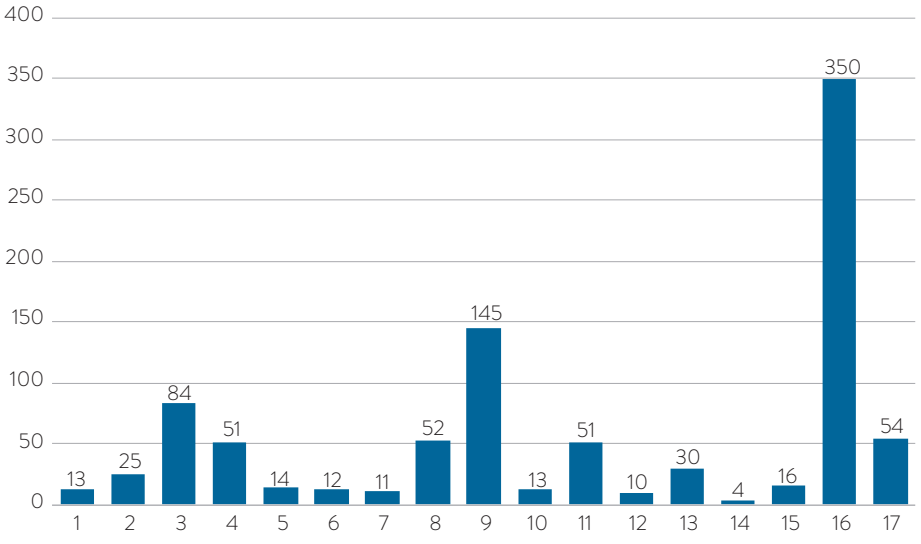


Figura 9.7. Posible aporte de los sistemas a los ODS

Fuente: elaboración propia.

Apoyo en procesos de agendamiento

Tres ejemplos de sistemas de IA en Colombia y Argentina contribuyen a procesos de agendamiento institucional mediante la producción y recolección de datos clave, los cuales las entidades públicas utilizan para prever problemas en los sectores de movilidad y salud (tabla 9.1).

Uno de los principales objetivos de Chatico, asistente virtual de la Alcaldía Mayor de Bogotá (Colombia), activo desde noviembre del 2021, es recoger información sobre los temas que más preocupan a los ciudadanos (Oficina Consejería Distrital TIC, 2022b). Este sistema fue diseñado para brindar atención permanente las veinticuatro horas del día, los siete días de la semana, a las solicitudes de información de los ciudadanos y visitantes sobre trámites y servicios digitales, la oferta de las entidades públicas distritales (por ejemplo, sobre programas sociales) e información de interés general (noticias de movilidad o sobre las fechas del racionamiento de agua que le corresponde a cada sector de la ciudad, en el marco de la crisis de abastecimiento en las fuentes hídricas, entre otras). Este sistema también permite a la administración local conocer cuáles son los asuntos que deberían priorizarse, dada la expresión de demandas de los ciudadanos a través de este canal digital.

Tabla 9.1. Ejemplos de sistemas de ia utilizados para el agendamiento institucional

País	Nombre de la herramienta o proyecto, y sector	Tipo de tecnología o técnica	Contribución técnica a las actividades de elaboración de programas
Colombia	Chatico (participación ciudadana)	*Chatbot*	Identificar las preocupaciones de los ciudadanos y promover su participación, además de brindar información.
Argentina	Gestión Epidemiológica basada en Inteligencia Artificial y Ciencia de Datos (salud)	Aprendizaje automático*	Anticiparse a las situaciones problemáticas.
Colombia	Identificación de Vías Terciarias con Imágenes de Satélite (movilidad)	Visión por ordenador y Aprendizaje automático.	Identificar la ubicación de los bienes estatales y los bienes públicos.

* Con base en la información disponible sobre el sistema, deducimos que utiliza este tipo de técnica.
Fuente: elaboración propia.

Además, Chatico ofrece la posibilidad a los ciudadanos de participar o votar en las iniciativas que se llevan a cabo en el marco de la estrategia de Gobierno Abierto de Bogotá, como los presupuestos participativos y las Causas Ciudadanas (Hernández, 2022; Oficina Consejería Distrital TIC, 2022a, 2022b; Secretaría Distrital de Planeación, 2022)[9]. Causas Ciudadanas es una iniciativa con la cual el Gobierno de Bogotá pretende promover "la participación directa de la población para conocer los temas que más aquejan al ciudadano y poder introducirlos en la agenda del Distrito" (Oficina Consejería Distrital TIC, 2022b), y para que los ciudadanos "participe[n] en campañas que buscan solucionar retos públicos" (Secretaría Distrital de Planeación, 2022, párr. 1)[10]. La Alcaldía informó que para la votación del presupuesto participativo del 2022 el 34 % de la participación fue a través de Chatico (Gobierno Abierto de Bogotá, 2023)[11].

Sin embargo, una de las trampas en las que pueden caer los funcionarios gubernamentales al utilizar este tipo de herramientas es asumir que lo que se discute en las redes sociales o se responde exclusivamente por medio de encuestas virtuales representa las preocupaciones de la población. Esto rara vez es así, ya que muchos ciudadanos no utilizan estos canales de comunicación o incluso no tienen acceso a un servicio de internet (en América Latina, se estima que alrededor del 32 % de la población no tiene acceso a este) (CAF, 2021b); además, terceros interesados en promover un problema (o incluso una solución) pueden generar ruido para desviar la atención, aprovechando la facilidad del acceso a las redes sociales (Valle-Cruz *et al*., 2020). Por lo tanto, confiar de forma exclusiva en este tipo de herramientas de IA para priorizar la acción gubernamental podría amplificar los efectos negativos de la brecha digital de la región. Otro riesgo es que actores externos consigan inflar artificialmente la demanda de servicios e

9 Bogotá es uno de los principales referentes de innovación pública de Colombia. La Alcaldía Mayor ha promovido el Gobierno Abierto y los proyectos tecnológicos a través del Laboratorio de Innovación Pública de Bogotá (iBO) (Gutiérrez y Dajer, 2023).

10 Este es un ejemplo de un sistema que puede contribuir a más de una etapa del ciclo de las políticas públicas, porque no solo permite encontrar qué problemas deben estar en la agenda del Gobierno, sino también apoyar el proceso de formulación de alternativas e incluso indirectamente el monitoreo o evaluación de "cómo avanzan las causas publicadas y selecciones, así como su implementación y ejecución" (Oficina Consejería Distrital TIC, 2022b).

11 Otra herramienta que está utilizando la Alcaldía de Bogotá para identificar las necesidades de sus ciudadanos es Keepcon. Este sistema, que utiliza técnicas de análisis del lenguaje natural con respecto a las preocupaciones, peticiones, quejas y reclamaciones recibidas a través de las cuentas de las redes sociales de la Alcaldía, pretende identificar y responder a los problemas planteados por los ciudadanos (Ágata, s. f.).

información mediante el uso de robots, que suplanten a los humanos e influyan estratégicamente en el Gobierno.

El segundo sistema estudiado ilustra cómo las herramientas de IA contribuyen a detectar con anticipación situaciones problemáticas a mediano y largo plazo. Este es el proyecto Gestión Epidemiológica basada en Inteligencia Artificial y Ciencia de Datos (ARPHAI), liderado en Argentina por el Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación (CIECTI), en asociación con el Ministerio de Ciencia, Tecnología e Innovación y el Ministerio de Salud (CIECTI, 2023; OECD y CAF, 2022). El objetivo de este sistema es la detección temprana de brotes epidémicos, a partir de la información de historias clínicas anonimizadas de establecimientos de salud del subsistema público de las provincias de La Rioja y San Juan (CIECTI, 2023; OECD y CAF, 2022). El sistema utiliza datos de contexto, como información sobre el clima, la geografía de la zona, las características socioeconómicas, las campañas de vacunación, entre otros (Engler y Pais, 2021). La anticipación de posibles brotes epidémicos permite a las autoridades de salud pública adoptar medidas preventivas y paliativas (CIECTI, 2023).

Por último, en la misma línea de proporcionar información para anticiparse a problemas futuros, los sistemas de IA también pueden ayudar a los Gobiernos a ubicar activos estatales y bienes públicos. En Colombia, por ejemplo, el Departamento Nacional de Planeación (DNP) desarrolló el sistema Identificación de Vías Terciarias con Imágenes Satelitales (Gutiérrez *et al*., 2023). Como su nombre indica, el sistema procesa grandes cantidades de imágenes de satélite para identificar la ubicación de la infraestructura vial; desde el 2019, ha generado una base de datos de carreteras terciarias (carreteras dentro de los municipios). El sistema de IA identificó y georreferenció las carreteras mediante imágenes de satélite de alta resolución de la constelación de PlanetScope. Hasta el 2022, se ha utilizado para procesar "cerca de 8000 imágenes de satélite, lo que equivale a 993 540 km² (87 %) del territorio nacional". (DNP, 2022, párr. 1).

Los principales beneficiarios de este sistema son los Gobiernos subnacionales colombianos, porque, según estimaciones del DNP (2022), "podrían ahorrar entre un 40 %, y un 60 % en el costo final de los inventarios viales, lo que se traduce en aproximadamente en COP 40 000 millones" (párr. 6). En resumen, el sistema permite a los Gobiernos identificar las deficiencias en la malla vial que comunica a los municipios y a las cabeceras municipales con las veredas con mayor precisión y menores costos. Es importante mencionar que esta herramienta también contribuye a otras fases del ciclo de las políticas públicas, ya que produce información que puede utilizarse para diseñar futuras intervenciones en la malla vial; así mismo, es una herramienta de gestión de la infraestructura existente.

Apoyo en procesos de formulación de política pública

A continuación, ilustramos cómo los sistemas de IA abordan algunos de los retos del diseño de políticas públicas, en particular los relacionados con el diagnóstico de un problema que ya ha entrado en la agenda, con miras a la formulación de alternativas de solución. Aquí describimos cuatro casos de Honduras, Guatemala, Chile y Colombia, que se refieren a sistemas diseñados para temas de educación, salud y libre competencia (tabla 9.2).

En Honduras y Guatemala, casi el 40 % de los alumnos de sexto grado abandonan sus estudios (Adelman *et al*., 2017; Escobar Gutiérrez *et al*., 2021; Haimovich *et al*., 2021). Con el fin de abordar el problema, ambos países implementaron sistemas de gestión de la información para predecir la deserción escolar y ayudar a las entidades gubernamentales, las escuelas y los profesores a decidir cuándo y dónde intervenir. En el 2021, la Secretaría de Educación de Honduras puso en marcha el Sistema de Alerta y Respuesta Temprana (SART), cuyo objetivo es identificar, atender y dar seguimiento a los estudiantes con mayor riesgo de abandonar la escuela (Secretaría de Educación, s. f.)[12]. Los maestros pueden entrar al sistema para verificar con el nombre o apellido de un alumno cuál es el riesgo de deserción y cuáles son los pasos por seguir para

Tabla 9.2. Ejemplos de sistemas de IA utilizados para la formulación de política pública

País	Nombre de la herramienta o proyecto, y sector	Tipo de tecnología o técnica	Contribución técnica a las actividades de formulación
Honduras y Guatemala	Sistema de Alerta y Respuesta Temprana (SART) (educación)	Modelo estadístico	Anticipar situaciones problemáticas y priorizar las intervenciones.
Chile	Sistema de Simulación para el Uso de Camas de Cuidados Intensivos (salud)	Aprendizaje automático	Anticipar situaciones problemáticas y planificar futuras intervenciones.
Colombia	Inspector (libre competencia)	Procesamiento del lenguaje natural	Identificar oportunidades para contribuir a los procesos de diseño de instrumentos de política pública.

Fuente: elaboración propia.

12 Este sistema también puede, eventualmente, generar insumos para monitorear la implementación de políticas públicas destinadas a atacar el abandono escolar.

reaccionar, dado el nivel de riesgo asignado (Asegurando la Educación, 2021).[13] Para predecir el abandono escolar,

> los datos sobre los estudiantes y a nivel de las escuelas [...] se digitalizan y se almacenan en redes de bases de datos administrativas interconectadas, que incluyen identificadores únicos de estudiantes para hacer seguimiento de los alumnos a lo largo del tiempo. (Adelman *et al*., 2019, párr. 3)[14]

Uno de los retos de estos sistemas radica en qué hacer en aquellas zonas del país donde la conexión a internet es deficiente y, por tanto, los profesores no pueden actualizar constantemente la información sobre sus alumnos.

En Guatemala, donde el Ministerio de Educación implementa el sistema, es posible vincular la información a las puntuaciones de los estudiantes en pruebas estandarizadas, que "se han realizado en el primer, tercer y sexto grado de educación primaria, en el último grado de educación secundaria inferior (noveno grado) y en el último grado de secundaria superior, con una frecuencia variable" (Adelman *et al*., 2017, p. 8)[15].

Estos sistemas de IA ayudan a los Gobiernos a organizar intervenciones para prevenir el abandono escolar (diagnóstico para la formulación), pero también se utilizan como herramientas de gestión (seguimiento de la implementación).

Los Gobiernos también emplean los sistemas de IA para prever situaciones problemáticas y planificar futuras intervenciones en el sector de la salud pública. Esto es especialmente importante en tiempos de crisis sanitaria, como la

13 El SART tiene una página web con estadísticas públicas, que permite al enlace departamental de migración ingresar al sistema para "llevar un control de su departamento para saber cuántos, cuáles y en qué centros educativos integrados [hay] niños identificados en riesgo leve, en riesgo moderado y en riesgo grave para dar seguimiento al proceso de atención de esta población de niños en riesgo de abandono" (Asegurando la Educación, 2021). Véase Secretaría de Educación (s. f.).

14 "Además del estado de matriculación individual, los datos incluyen algunos datos demográficos (sexo y edad), índices de asistencia, si el niño asiste a una institución pública o privada y el curso actual" (Adelman *et al*., 2017, p. 11).

15 En Perú, Chile y Argentina se han puesto en marcha sistemas similares destinados a abordar los problemas de abandono escolar en escuelas y universidades. El sistema Alerta Escuela, implementado en Perú en octubre del 2020, tiene como objetivo identificar a los estudiantes con mayor riesgo de abandono escolar, y orientar y gestionar acciones concretas para ayudarlos (Ministerio de Educación de Perú, 2023). En el 2019, la Universidad de Aysén de Chile implementó un sistema de alerta temprana para la deserción universitaria (GobLab, 2023); y en Argentina, en el 2022, la Dirección General de Escuelas de Mendoza presentó un sistema de alerta temprana (Prensa Gobierno de Mendoza, 2022).

provocada por la pandemia de covid-19. En este sentido, en marzo del 2020, el Hospital Clínico Regional Dr. Guillermo Grant Benavente (Concepción, Chile), con el apoyo de la Universidad de Concepción, la Universidad de Santiago de Chile y el Instituto Sistemas Complejos de Ingeniería (ISCI), puso en marcha un sistema de simulación de uso de camas de unidad de cuidados intensivos (UCI). El sistema permitió "estimar la necesidad de camas de UCI según las tendencias de contagio por covid-19 en la región del Biobío", "proyectar la oferta, y demanda de camas", para prever futuras saturaciones de capacidad y tomar medidas preventivas (GobLab, 2023)[16].

Uno de los posibles riesgos a los que se enfrentan este tipo de sistemas es que la información con la que fueron entrenados conduzca a reproducir sesgos. Por ejemplo, en un estudio para estimar el exceso de muertes por raza, etnia y residencia en una zona socialmente vulnerable, entre 498 adultos ingresados en una UCI de uno de los seis hospitales de Boston, mediante un sistema de puntuación de las normas de atención en crisis (CSOC), en medio de una oleada de covid-19 del 13 de abril al 22 de mayo del 2020, se descubrió que "casi el doble de la proporción de pacientes negros se puntuó en el grupo de prioridad más baja en comparación con todos los demás pacientes" (Riviello *et al*., 2022, p. 1).

El último ejemplo de un sistema de IA que contribuye a la formulación de políticas se da en otro sector gubernamental: la promoción de mercados competitivos. En Colombia, la Superintendencia de Industria y Comercio (SIC) es la autoridad nacional que, entre otras funciones, se encarga de proteger el derecho a la libre competencia económica. La función de protección de la competencia implica actividades cuyo objetivo es "promover un entorno competitivo en los mercados" (denominadas *abogacía de la competencia*) (Gutiérrez y Suárez, 2023, p. 146).

Una de estas actividades llevadas a cabo por la SIC consiste en el seguimiento de los proyectos regulatorios que puedan afectar negativamente las dinámicas competitivas de los mercados y la emisión de conceptos para que los reguladores consideren dichos potenciales impactos. Para ello, la SIC utiliza Inspector, una herramienta que

16 "La herramienta contiene un modelamiento de simulación mediante eventos discretos" ("Desarrollan herramienta que simula uso de camas UCI en Hospital Regional de Concepción", 2020), para lo cual se utiliza información del Ministerio de Salud y del hospital, "como el número real de pacientes que han requerido UCI" (GobLab, 2023).

> apoya el proceso de seguimiento por parte del Grupo de Defensa de la Competencia de los diferentes proyectos normativos publicados en las páginas web de los reguladores [...] [y] alerta a la autoridad de los proyectos normativos publicados detectados en sus páginas web oficiales que puedan afectar a la competencia. (Thibault y Groza, 2023, pp. 81-82)

Esta herramienta

> utiliza la tecnología web-scraping para detectar las regulaciones económicas que fueron expedidas por los reguladores colombianos, por ejemplo, aquellas en las que no se realizó una notificación previa a la autoridad de acuerdo con las normas de notificación de la defensa de la competencia. (Gutiérrez y Suárez, 2023, p. 169)

Una vez que la agencia de competencia detecta un proyecto de regulación que debería haber sido notificado, solicita información al regulador para preparar un dictamen que le ayude a decidir cómo diseñar su regulación. Por lo tanto, Inspector puede clasificarse como una herramienta que facilita la tarea de la SIC de participar en los procesos de elaboración de regulaciones, un tipo de instrumento de política pública.

Apoyo en procesos de implementación de política pública

En esta sección describimos seis sistemas adoptados en Brasil, Chile, Colombia, Perú y Argentina que contribuyen a la implementación de políticas públicas. Las herramientas son usadas para apoyar procesos asociados con la contratación pública y problemas medioambientales y sanitarios (tabla 9.3).

La Contraloría General de Brasil adoptó en el 2015 la Auditoria Preventiva em Licitações (Alice, por su acrónimo en portugués), una herramienta que examina las inconsistencias en las licitaciones públicas que se registran en el Portal de Compras del Gobierno Federal (Controladoria-Geral da União, 2021). Esta herramienta de supervisión se basa en la información del sistema de contratación pública de Brasil (Comprasnet, gestionado por el Ministerio de Economía). El sistema utiliza técnicas de minería de textos e IA (Cetina, 2020)[17].

17 El sistema "permite seleccionar automáticamente las ofertas para alertar a la administración sobre los riesgos en la contratación y evitar el gasto de fondos federales, anular o suspender las ofertas innecesarias o con indicios de fraude y, en otros casos, ajustar los valores estimados y los importes" (Controladoria-Geral da União, 2021).

Tabla 9.3. Ejemplos de sistemas de IA utilizados para la implementación de política pública

País	Nombre de la herramienta o proyecto, y sector	Tipo de tecnología o técnica	Contribución técnica a las actividades de aplicación
Brasil	Auditoria Preventiva em Licitações (Alice) (contratación pública)	Procesamiento del lenguaje natural y aprendizaje automático	Priorizar para la implementación de la ley.
Chile	Guardián del Bosque (medioambiental)	Aprendizaje automático*	Supervisar y priorizar la acción gubernamental.
Colombia	Guardianes de la Selva (medioambiental)	Aprendizaje automático	Detectar patrones y anomalías y priorizar la acción gubernamental.
Colombia	Guacamaya (medioambiental)	Visión por computador y aprendizaje automático	Detectar patrones y anomalías y priorizar la acción gubernamental.
Perú	AnemiaApp (salud)	Visión por computador y aprendizaje automático	Detectar anomalías y priorizar la acción gubernamental.
Argentina	Crecer con Salud (salud)	*Chatbot*	Mejorar el uso de los servicios públicos del Gobierno.

* Con base en la información disponible sobre el sistema, y en el sistema homólogo utilizado en Colombia, deducimos que utiliza este tipo de tecnología.
Fuente: elaboración propia.

Según el Gobierno brasileño, con base en las alertas emitidas por la herramienta y las subsecuentes acciones preventivas de la Contraloría, entre el 2015 y el 2022, "58 licitaciones, y concursos que estaban en curso, y presentaban debilidades, y riesgos en la contratación fueron ajustados, anulados o suspendidos, el monto involucrado es de R$8,61 mil millones" (Controladoria-Geral da União, 2021)[18]. Esta herramienta también entra en la categoría de sistemas de IA que contribuyen a la evaluación de políticas públicas.

18 En Colombia existen tres sistemas que también buscan detectar anomalías en los procesos de contratación: la Contraloría General de la República cuenta con Océano, que "establece

El segundo ejemplo es un sistema de IA utilizado para reducir los costos de implementación de una política en Chile. En concreto, se trata de una herramienta de supervisión que se desarrolló a través de una asociación público-privada, con el objetivo de contribuir a la implementación de políticas medioambientales. Desde agosto del 2021, el Ministerio de Medio Ambiente y la Superintendencia de Medio Ambiente, con el apoyo de Huawei, Rainforest Connection y Forest Ethics, implementaron Guardián del Bosque, "herramienta de monitoreo audio acústico basado en inteligencia artificial"; el propósito del sistema es la "preservación del zorro de Darwin a través de un sistema de monitoreo y análisis de audio asistido por inteligencia artificial (IA) a bajo costo" en la región del Biobío (GobLab, 2023)[19].

En otros países de la región se han aplicado estrategias similares. En Colombia, por ejemplo, el Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, con el apoyo de la Fundación Biodiversa Colombia, Rainforest Connection, Corantioquia y Huawei, desarrolló la herramienta Guardianes de la Selva (Instituto Humboldt, 2023). La herramienta ayuda a detectar los sonidos de actividades ajenas al ecosistema, como los producidos por la caza o las actividades de deforestación. Además, también en Colombia, el sistema Guacamaya, IA por la Amazonía, utiliza imágenes por satélite, grabaciones bioacústicas y cámaras trampa para proteger la fauna y la flora del Amazonas. Este proyecto se desarrolló a través de una alianza público-privada que incluyó a la Universidad de los Andes, Microsoft, el Instituto Humboldt y el Instituto Sinchi (Forero, 2023).

Existen otros casos en los que los sistemas de IA detectan anomalías y contribuyen a priorizar la acción gubernamental. En Perú, por ejemplo, el Gobierno y la Universidad Peruana Cayetano Heredia desarrollaron un sistema

---

relaciones entre los contratos celebrados a nivel nacional y los analiza para detectar posibles casos de corrupción; a través de un análisis matricial de redes y construcción de vectores" (Santiso y Cetina, 2022, p. 117); la Procuraduría General de la República utiliza el Análisis de Redes Criminales en el Contexto de la Procuraduría General de la República (ARCPGN), para "identificar estructuras ilícitas de cooptación institucional" (Santiso y Cetina, 2022, p. 108); por su parte, "Sherlock es el proyecto de herramienta de análisis de datos que busca apoyar a los investigadores de la SIC en la identificación de indicios o patrones que sugieran posibles conductas anticompetitivas en los procesos de contratación pública" (*bid-rigging*) (Schrepel y Groza, 2022, p. 87).

19 El sistema funciona con "minicomputadoras solares que se instalan en las copas de los árboles, […] estos dispositivos graban audio, y envían toda la información grabada a Huawei Cloud […] [para] identificar las vocalizaciones grabadas con IA para reconocer los patrones de las diferentes especies" (Superintendencia del Medio Ambiente de Chile, 2021, párr. 4).

que detecta la anemia e informa a las autoridades sobre los casos graves[20]. La universidad diseñó la AnemiaApp, una aplicación móvil desde la que la gente puede acceder al sistema que detecta la anemia con base en una fotografía del párpado inferior. Esta iniciativa, que se utiliza sobre todo en zonas remotas, cuenta con el apoyo del Ministerio de Desarrollo e Inclusión Social del Perú (Midis) (Oré Arroyo, 2019; Zimic, 2019). Con la fotografía del párpado y un formulario con los datos del paciente (nombre, número de identificación, fecha de nacimiento y sexo), el algoritmo procesa la imagen, analiza las características de la membrana que recubre la superficie externa y determina el nivel de hemoglobina (OECD y CAF, 2022; Oré Arroyo, 2019). Cuando se detecta un caso de anemia grave, se envía automáticamente un mensaje de alerta al Midis para que se tomen las medidas correspondientes (Oré Arroyo, 2019; Zimic, 2019). Un riesgo crítico que pueden crear estos sistemas es el acceso y uso no autorizados de datos personales; por ello, el uso de sistemas de IA desarrollados con datos personales y que los utilicen debe estar sujetos a protocolos de protección de dicha información, que incluye datos sensibles (relacionados con la salud) y datos de menores.

Por último, también relacionado con la implementación de políticas sanitarias, el Gobierno nacional argentino desarrolló un *chatbot* que pretende recordar a las mujeres los controles prenatales y posnatales. Esto fue reconocido como un problema público, ya que el 30 % de las mujeres no completó el cronograma de controles durante el embarazo, y una de cada diez nunca asistió al médico antes del momento del parto (Observatory of Public Sector Innovation [OPSI], 2018). El asistente virtual, llamado Crecer con Salud, funciona a través de Facebook Messenger[21] y fue diseñado para "dirigirse a mujeres embarazadas y madres con bebés menores de un año en Argentina"; entre otras funciones, "envía recordatorios de asistencia" a los controles (OPSI, 2018). Una vez más, los sistemas de IA que utilizan datos personales sensibles requieren los máximos esfuerzos de los Gobiernos para proteger la privacidad.

20 Se mencionó que, en principio, este sistema contaba con el apoyo del Ministerio de Desarrollo e Inclusión Social (Midis), según informes del Observatorio fAIr ALC (BID, s. f.b); sin embargo, no fue posible confirmar si esta alianza con la Universidad Peruana Cayetano Heredia sigue vigente.

21 "El Gobierno eligió Facebook Messenger porque la plataforma es utilizada por más de 30 millones de argentinos, incluido el 90 % de las embarazadas en las maternidades, según una investigación interna del Gobierno" (OECD y CAF, 2022).

Apoyo en procesos de evaluación de política pública

En esta sección caracterizamos tres sistemas adoptados en Colombia, Chile y México que contribuyen a evaluar las políticas públicas relacionadas con problemas de infraestructura y medio ambiente (tabla 9.4).

La Unidad Administrativa Especial de Rehabilitación y Mantenimiento Vial (UAERMV), de la Alcaldía Mayor de Bogotá (Colombia), implementó en el 2020 el Sistema de Información Geográfica Misional y de Apoyo (SIGMA) (Morales, 2020). Su propósito es "analizar, y desplegar información geográfica de manera centralizada, e integrada" (UAERMV, 2022), para optimizar la reparación y el mantenimiento de la red vial de la ciudad (Morales, 2020). Para ello, utiliza información relacionada con el número de personas que transitan por la vía que se va a intervenir, cuáles son las horas de mayor volumen de tráfico, el número de vehículos, y la cercanía a colegios y hospitales, la cual se obtiene de otras entidades o sistemas de información de la Alcaldía, como ArcGIS, el Instituto de Desarrollo Urbano o la Infraestructura de Datos Espaciales de Bogotá (IDECA) (Morales, 2020; UAERMV, 2021, 2022). Esta herramienta se utiliza para la "formulación, planeación, ejecución y seguimiento de la conservación de la red vial local para la toma oportuna de decisiones" (UAERMV, 2022)[22]; por lo tanto, se clasifica al menos en las tres últimas etapas del ciclo de las políticas públicas.

Tabla 9.4. Ejemplos de sistemas de IA utilizados para la evaluación de política pública

País	Nombre de la herramienta o proyecto, y sector	Tipo de tecnología o técnica	Contribución técnica a las actividades de evaluación
Colombia	Sistema de Información Geográfica Misional y de Apoyo (SIGMA) (movilidad)	Aprendizaje automático*	Supervisión
Chile	Sistema de Vigilancia con Infometría por Satélite (medio ambiente)	Visión por ordenador	Supervisión
México	Plataforma del Centro Intercultural de Estudios de Desiertos y Océanos (CEDO) (medio ambiente)	Aprendizaje automático	Análisis de datos

* Con base en la información disponible sobre el sistema, deducimos que utiliza este tipo de tecnología.
Fuente: elaboración propia.

22 El sistema proporciona: "(1) Información centralizada de la Red Local de Carreteras. (2) Disponibilidad de diagnósticos de carreteras sobre el terreno en tiempo real. (3) Reducción

Un segundo caso de un sistema que facilita la recopilación de información tiene lugar en Chile, donde la Superintendencia de Medio Ambiente implementó en el 2017 el Sistema de Monitoreo con Infometría Satelital. La Superintendencia utiliza el sistema para "fiscalizar que la ubicación de los Centros de Engorda de Salmones (CES) cumpla con las concesiones otorgadas" (GobLab, 2023). Este sistema se nutre de "la información del Radar de Apertura Sintética (SAR) captada por los satélites Sentinel-1 […] gracias a la cooperación con el Programa Copérnico de la Unión Europea y la Agencia Espacial Europea" (GobLab, 2023). La información obtenida con el SAR es procesada por un "algoritmo de Clasificación Polarimétrica Dual", que permite identificar, incluso en condiciones de nubosidad, si las CES se encuentran dentro de las concesiones, aprovechando que las jaulas salmoneras tienen una dispersión de oleaje diferente a la del agua circundante[23].

El último ejemplo que examinaremos se trata de un sistema de IA que contribuye a la fase de evaluación de políticas mediante el apoyo en el análisis de datos[24]. La implementación eficaz de las políticas públicas depende, entre otros factores, de la capacidad de comprender las realidades, los temores y los contextos de los ciudadanos, así como de adaptar el lenguaje para que la información llegue a los directamente implicados. En las comunidades costeras de México, por ejemplo, "la pesca artesanal desempeña un papel social y económico fundamental", de ahí que la información proporcionada a los pescadores "sobre el cambio climático y las percepciones del cambio ambiental […] influye en la aplicación de las políticas, [mientras que] las experiencias de los pescadores [a su vez] pueden ayudar a informar la adopción de nuevas políticas" (BID, s. f.a).

La plataforma del Centro Intercultural de Estudios de Desiertos y Océanos (CEDO) analiza la información disponible sobre el cambio climático y "evalúa

---

de los tiempos de operación. (4) Agilidad en la respuesta a las solicitudes de información. (5) Fortalecimiento de la coordinación intersectorial entre: Instituto de Desarrollo Urbano-IDU, Unidad Administrativa Especial de Rehabilitación y Mantenimiento Vial-UAERMV, Fondos de Desarrollo Local-FDL" (UAERMV, 2022).

23 "Entre 2018 y 2019 se utilizó el sistema en las 4 regiones más australes de Chile. De las imágenes procesadas, hubo 78 hallazgos, de los cuales 27 correspondieron a jaulas de CES fuera de concesión" (GobLab, 2023).

24 Este se trata de una herramienta desarrollada por una coalición de organizaciones sin ánimo de lucro de México y Estados Unidos. A pesar de que su origen no es gubernamental, decidimos incluir el caso en la medida en que en su desarrollo estuvieron involucrados un centro de investigación público de México y universidades públicas de México y Estados Unidos, además de su potencial contribución a la generación de valor público.

la forma en que los medios de comunicación nacionales, regionales y locales interpretan y comunican los mensajes sobre el cambio climático en los Estados costeros de México" (BID, s. f.a). Con esta información, los responsables públicos de la toma de decisiones frente a la gestión de los recursos podrán conocer cómo entienden los pescadores locales la información disponible sobre el cambio climático (BID, s. f.a). El CEDO aplica el análisis de sentimientos para deducir qué temas preocupan a los pescadores locales y cómo las entidades gubernamentales pueden ser más eficaces con su estrategia de comunicación (BID, s. f.a). Esta herramienta contribuye a evaluar la política y a identificar y estructurar nuevas cuestiones problemáticas, por lo que también podría apoyar los procesos de agendamiento y formulación de política pública.

Conclusiones, implicaciones de política pública y futuras vías de investigación

El objetivo principal de este capítulo fue examinar cómo los sistemas de IA adoptados por las entidades públicas de América Latina y el Caribe apoyan el desempeño de las actividades asociadas a las principales etapas del ciclo de las políticas públicas: agendamiento, formulación, implementación y evaluación. Para ello, el capítulo presentó las principales estadísticas descriptivas de 735 sistemas de IA de la región y exploró dieciséis sistemas de IA implementados por entidades públicas nacionales o subnacionales en Argentina, Brasil, Chile, Colombia, Guatemala, Honduras, México y Perú.

El capítulo describe los sistemas de IA adoptados en diversos sectores, como salud, medio ambiente, movilidad y educación. Los sistemas también apoyan actividades transversales relacionadas con los procesos de contratación, la participación ciudadana y la libre competencia; además, se desarrollaron mediante diversas técnicas, como el aprendizaje automático, el procesamiento de lenguaje natural, la visión por ordenador, entre otras.

También argumentamos que un sistema de IA puede contribuir a una o varias de las fases del ciclo de las políticas públicas. Por ejemplo, es posible que un sistema ayude a identificar una situación problemática que afecte a los ciudadanos y genere, a su vez, aportes y alternativas para contribuir a una solución.

Otro hallazgo clave, a partir de la construcción de la nueva base datos, es que la información disponible públicamente sobre los sistemas de IA adoptados por los Gobiernos latinoamericanos no siempre es de fácil acceso o está incompleta. Encontramos poca información sobre cómo se desarrollaron los sistemas, quién los desarrolló, los costos y las evaluaciones de los resultados e impactos asociados a su uso. Solo seis países de América Latina y el Caribe

(Argentina, Brasil, Colombia, Chile, México y Uruguay) cuentan con repositorios de algoritmos públicos, y en algunos de estos repositorios parecen presentar un subregistro de sistemas de IA (GPAI, 2024; Gutiérrez y Muñoz-Cadena, 2023c; Muñoz-Cadena y Gutiérrez, 2025). Esto contrasta con las promesas realizadas por las iniciativas de gobierno abierto en América Latina y el Caribe, que deberían conllevar ciertos niveles de transparencia respecto a los sistemas de IA que despliegan los Gobiernos (Gutiérrez y Castellanos-Sánchez, 2023).

Por lo tanto, una implicación política derivada de nuestra investigación es que los Estados latinoamericanos deberían informar de forma más proactiva sobre el uso de sistemas de IA para (semi)automatizar o apoyar procesos de toma de decisiones. Esto es particularmente crítico dado que algunos de estos pueden tener un gran impacto en la vida y los derechos fundamentales de millones de personas o contribuir a amplificar sesgos o estereotipos sobre ciertos segmentos de la población. Muy pocos países de la región cuentan con repositorios en línea de algoritmos públicos, y algunos de los existentes, como los publicados por el Gobierno colombiano, registran pocos algoritmos y no incluyen información actualizada (GPAI, 2024; Gutiérrez, 2024d; Gutiérrez y Muñoz-Cadena, 2023a; Muñoz-Cadena y Gutiérrez, 2025).

De igual manera, se identificó cómo los sistemas de IA podrían contribuir positivamente a los objetivos estatales. Así mismo, respecto de algunos de los sistemas, se estableció cómo su implementación podría crear nuevos riesgos que deben ser gestionados por las entidades públicas que despliegan y usan las herramientas. Este punto, que podría ser profundizado en futuras investigaciones, es esencial para la adquisición, el desarrollo, el despliegue y la utilización de herramientas de IA de forma ética, responsable y compatible con la protección de los derechos humanos por parte del Estado.

En este sentido, los sistemas de IA utilizados por los Gobiernos deben diseñarse teniendo en cuenta a los usuarios finales y a los beneficiarios (Flórez Rojas, 2023). Si no se tienen en cuenta múltiples perspectivas, los sistemas de IA pueden reproducir información sin un valor añadido o, lo que es peor, amplificar los sesgos sociales. Además, si un Estado utiliza sistemas de IA para apoyar funciones en las que los derechos humanos pueden estar en juego (por ejemplo, la educación o la salud), deberían establecerse mecanismos para prevenir y gestionar los riesgos (como contar con evaluaciones de impacto antes de desplegar la herramienta).

Otra implicación que se deriva de las limitaciones identificadas de los sistemas de IA estudiados es que estas herramientas no resuelven la necesidad de abordar los retos políticos asociados a los procesos del ciclo de las políticas públicas. Además, los sistemas no pueden sustituir los juicios de valor que los

responsables de la toma de decisiones deben realizar en sociedades democráticas y pluralistas, en las que la elaboración de políticas requiere equilibrar derechos e intereses. De ahí que las entidades públicas interesadas en adquirir o desarrollar sistemas de IA deban ser conscientes de su limitada capacidad para sustituir la toma de decisiones humana.

De nuestra investigación se desprenden tres grandes vías futuras de investigación. En primer lugar, es necesario comprender cómo y por qué las entidades públicas deciden adquirir, desarrollar y adoptar sistemas de IA para apoyar o automatizar sus procesos de toma de decisiones. Del mismo modo, sería pertinente explorar qué situaciones o contextos pueden o no favorecer la adopción e implementación de sistemas de IA por parte de las entidades públicas en los diferentes países latinoamericanos.

En segundo lugar, futuros estudios podrían explorar los factores determinantes del éxito del diseño y la aplicación de sistemas de IA que apoyen las decisiones en el ciclo de las políticas públicas. Como sostienen Tangi *et al.* (2024b), la identificación de los casos de mejores prácticas

> puede poner de relieve elementos concretos de los procesos de innovación y la forma en que las organizaciones del sector público están adoptando y utilizando la IA, lo que permitirá a los responsables políticos identificar los objetivos y el camino a seguir a la hora de desarrollar agendas de innovación y políticas relacionadas. (p. 222)

Por último, con corte de mayo del 2025, al menos dieciocho países de América Latina y el Caribe están elaborando leyes o reglamentos para regular la IA: Argentina, Bahamas, Barbados, Bolivia, Brasil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Guatemala, El Salvador, Honduras, México, Panamá, Perú, República Dominicana y Uruguay (Gutiérrez, 2024c, 2024b; Gutiérrez y Hurtado, 2025). Países como Chile han dado pasos para crear marcos normativos que obligan a las entidades públicas a informar sobre los sistemas de IA que utilizan, en especial cuando estos pueden afectar a los derechos de los ciudadanos; sin embargo, en otros países no se están estudiando normativas similares ni cuentan con repositorios o registros de los sistemas de IA que utilizan las entidades públicas (GPAI, 2024). Esto es preocupante porque, en la medida en que los ciudadanos no disponen de información actualizada sobre los sistemas que utilizan los Gobiernos, resulta difícil supervisarlos y, por tanto, alertar sobre posibles riesgos. De ahí que el estudio de los marcos regulatorios del desarrollo y uso de los sistemas de IA por parte de los Estados pueda ser también una prometedora vía de investigación.

Referencias

Ada Lovelace Institute, AI Now Institute y Open Government Partnership. (2021). *Algorithmic accountability for the public sector. Learning from the first wave of policy implementation*. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Adelman, M., Haimovich, F. y Alasino, E. (2019, 29 de octubre). Who will drop out of school? Leveraging management information systems to predict dropouts in Guatemala and Honduras. *World Bank Blogs*. https://blogs.worldbank.org/education/who-will-drop-out-school-leveraging-management-information-systems-predict-dropouts

Adelman, M., Haimovich, F., Ham, A. y Vázquez, E. (2017). *Predicting school dropout with administrative data: New evidence from Guatemala and Honduras.* World Bank.

Ágata. (s. f.). *¿Qué hacemos?* https://agatadata.com/queHacemos

AISur. (s. f.). *Reconocimiento facial en América Latina*. https://estudio.reconocimientofacial.info

AlgorithmWatch y Bertelsmann Stiftung. (2019). *Automating society: Taking stock of automated decision making in the EU*. AlgorithmWatch. https://www.algorithmwatch.org/automating-society

AlgorithmWatch y Bertelsmann Stiftung. (2020). *Automating society report 2020*. AlgorithmWatch. https://automatingsociety.algorithmwatch.org

Arellano Gault, D. y Blanco, F. (2020). *Políticas públicas y democracia*. (2.ª ed.). Instituto Federal Electoral.

Arun, C. (2020). AI and the global south: Designing for other worlds. En M. D. Dubber, F. Pasquale y S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 588-606). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.38

Asegurando la Educación. (2021, 7 de julio). *Fases y etapas: Sistema de Alerta y Respuesta Temprana (SART)* [video]. https://www.youtube.com/watch?v=0_q0oyu2VD8

Banco de Desarrollo de América Latina y el Caribe (CAF). (2021a). *Experiencia: Datos e inteligencia artificial en el sector público*. http://scioteca.caf.com/handle/123456789/1793

Banco de Desarrollo de América Latina y el Caribe (CAF). (2021b). *Desigualdad 4.0: A cerrar la brecha digital*. https://www.caf.com/es/actualidad/noticias/desigualdad-40-a-cerrar-la-brecha-digital/

Banco Interamericano de Desarrollo (BID). (s. f.a). CEDO intercultural. https://fairlac.iadb.org/cedo-intercultural

Banco Interamericano de Desarrollo (BID). (s. f.b). MIDIS de AYNI LAB. https://fairlac.iadb.org/midis

Banco Interamericano de Desarrollo (BID) y Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco). (s. f.). *Observatorio fAIr LAC*. https://fairlac.iadb.org/observatorio

Benbya, H., Davenport, T. H. y Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, *19*(4). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3741983

Brauneis, R. y Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale Journal of Law and Technology*, *20*, 103-176.

Bundesamt für Sicherheit in der Informationstechnik (BSI). (2024). *Generative AI models. Opportunities and risks for industry and authorities*. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Generative_AI_Models.pdf?__blob=publicationFile&v=4

Calo, R. y Citron, D. K. (2021). The automated administrative state: A crisis of legitimacy. *Emory Law Journal*, *70*(4), 797-845.

Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación (CIECTI). (2023). *ARPHAI: Gestión epidemiológica basada en inteligencia artificial y ciencia de datos*. http://www.ciecti.org.ar/arphai/

Cetina, C. (2020). *Tres preguntas sobre el uso de los datos para luchar contra la corrupción*. Banco de Desarrollo de América Latina y el Caribe (CAF). https://scioteca.caf.com/bitstream/handle/123456789/1544/Tres_preguntas_sobre_el_uso_de_los_datos_para_luchar_contra_la_corrupcion.pdf?sequence=1&isAllowed=y

Chenou, J.-M. y Rodríguez Valenzuela, L. E. (2021). Habeas data, habemus algorithms: Algorithmic intervention in public interest decision-making in Colombia. *Law, State and Telecommunications Review*, *13*(2), 56-77. https://doi.org/10.26512/lstr.v13i2.34113

Citron, D. (2008). Technological due process. *Washington University Law Review*, *85*(6), 1249-1313.

Controladoria-Geral da União. (2021, 29 de marzo). *Auditoria Preventiva em Licitações-Alice*. Presidência da República. https://www.gov.br/cgu/pt-br/centrais-de-conteudo/campanhas/cgu-contra-corrupcao/temas/alice

Crawford, K. (2021). *The atlas of AI*. Yale University Press.

Dejusticia. (2022, 2 de diciembre). Reconocimiento facial y DDHH: 13 historias para entender sus implicaciones. https://www.dejusticia.org/reconocimiento-facial-y-dd-hh-13-historias-para-entender-sus-implicaciones/

Departamento Nacional de Planeación (DNP). (2022, 25 de febrero). DNP en articulación con el Ministerio de Transporte desarrolló el proyecto de identificación de vías terciarias con inteligencia artificial. https://www.dnp.gov.co/Paginas/DNP-y-MiniTransporte-desarrollo-el-proyecto-de-identificacion-de-vias-terciarias-con-inteligencia-artificial.aspx

Desarrollan herramienta que simula uso de camas UCI en Hospital Regional de Concepción. (2020, 14 de abril). *Noticias UdeC*. https://noticias.udec.cl/desarrollan-herramienta-que-simula-uso-de-camas-uci-en-hospital-regional-de-concepcion/

Engler, C. y Pais, C. (2021, 22 de diciembre). *Modelo basado en agentes para la epidemiología de covid-19 con movilidad y actividades sociales representadas por un conjunto de modelos ocultos de Markov* [video]. https://www.youtube.com/watch?v=Cae4chHX_Kk

Engstrom, D. F., Ho, D. E., Sharkey, C. M. y Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *SSRN Electronic Journal*, (20-54). https://doi.org/10.2139/ssrn.3551505

Escobar Gutiérrez, E., Ramírez Roa, D. P., Quevedo Hernández, M., Insuasti Ceballos, H. D., Jiménez Ospina, A., Montenegro Helfer, P., Numpaque Cano, J. S., Rocha Ruiz, C. A., Ruiz Saenz, J. A., Berniell, M. L. y Zapata, E. (2021). *Aprovechamiento de datos para la toma de decisiones en el sector público*. Banco de Desarrollo de América Latina y el Caribe (CAF) y Departamento Nacional de Planeación (DNP). https://cafscioteca.azurewebsites.net/handle/123456789/1776

Flórez Rojas, M. L. (2023). Pensamiento de diseño y marcos éticos para la inteligencia artificial: Una mirada a la participación de las múltiples partes interesadas. *Desafíos*, *35*. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-40352023000100002&nrm=iso

Forero, N. (2023, 31 de agosto). "Guacamaya", para salvar la Amazonía con inteligencia artificial. *Noticias Universidad de los Andes*. https://uniandes.edu.co/es/noticias/ambiente-y-sostenibilidad/guacamaya-para-salvar-la-amazonia-con-inteligencia-artificial

Global Partnership on Artificial Intelligence (GPAI). (2024). *Algorithmic transparency in the public sector: A state-of-the-art report of algorithmic transparency instruments*. https://gpai.ai/projects/responsible-ai/algorithmic-transparency-in-the-public-sector/algorithmic-transparency-in-the-public-sector.pdf

Gobierno Abierto de Bogotá. (2023). *¿Cómo se han implementado presupuestos participativos en Bogotá?* https://www.sdp.gov.co/sites/default/files/eventos/infografia_dialogos_ciudadanos.pdf

GobLab. (2023). *Algoritmos públicos*. Universidad Adolfo Ibáñez. https://algoritmospublicos.cl

Gómez Mont, C., Del Pozo, C. M., Martínez Pinto, C. y Martín del Campo Alcocer, A. V. (2020). *Artificial intelligence for social good in Latin America and the Caribbean: The regional landscape and 12 country snapshots*. Inter-American Development Bank. https://publications.iadb.org/publications/english/document/Artificial-Intelligence-for-Social-Good-in-Latin-America-and-the-Caribbean-The-Regional-Landscape-and-12-Country-Snapshots.pdf

Gutiérrez, J. D. (2020). Retos éticos de la inteligencia artificial en el proceso judicial. En *Derecho procesal: Nuevas tendencias. XLI Congreso Colombiano de Derecho Procesal* (pp. 499-516). Instituto Colombiano de Derecho Procesal (ICDP) y Universidad Libre. https://doi.org/10.2139/ssrn.4011179

Gutiérrez, J. D. (2024a). Chapter 24: Critical appraisal of large language models in judicial decision-making. En *Handbook on public policy and artificial intelligence* (pp. 323-338). Edward Elgar Publishing. https://doi.org/10.4337/9781803922171.00033

Gutiérrez, J. D. (2024b). *Consultation paper on AI regulation: Emerging approaches across the world*. United Nations Educational, Scientific and Cultural Organization (Unesco). https://unesdoc.unesco.org/ark:/48223/pf0000390979

Gutiérrez, J. D. (2024c). Regulación sobre IA. *Foro Administración, Gestión y Política Pública*. https://forogpp.com/inteligencia-artificial/regulacion-sobre-ia/

Gutiérrez, J. D. (2024d). Repositorios y registros públicos de algoritmos. *Foro Administración, Gestión y Política Pública*. https://forogpp.com/inteligencia-artificial/repositorios-y-registros-de-algoritmos/

Gutiérrez, J. D. (2024e). De qué hablamos cuando hablamos de IA. *Foro Administración, Gestión y Política Pública*. https://forogpp.com/2024/11/15/de-que-hablamos-cuando-hablamos-de-ia/

Gutiérrez, J. D. y Castellanos-Sánchez, M. (2023). Transparencia algorítmica y Estado abierto en Colombia. *Revista Reflexión Política*, *25*(52). https://revistas.unab.edu.co/index.php/reflexion/issue/archive

Gutiérrez, J. D., Castellanos-Sánchez, M. y Muñoz-Cadena, S. (2025). *Sistemas automatizados de toma de decisiones en el sector público*

colombiano (versión 2.1) [Dataset]. https://sistemaspublicos.tech/sistemas-automatizados-de-toma-de-decisiones-en-el-sector-publico-de-colombia/

Gutiérrez, J. D. y Dajer, D. M. (2023). Pensamiento de diseño y procesos de política pública. *Desafíos*, *35*(1), 1-27.

Gutiérrez, J. D. y Flórez, M. L. (2023). Presentación de número: Retos de la gobernanza de datos y de inteligencia artificial en el sector público Iberoamericano. *GIGAPP Estudios Working Papers*, *10*(270), 329-334.

Gutiérrez, J. D. y Hurtado, S. (2025). *Regulación sobre IA en América Latina y el Caribe* (versión 1) [Dataset]. Sistemas de Algoritmos Públicos, Universidad de los Andes. https://sistemaspublicos.tech/regulacion-sobre-ia-en-america-latina/

Gutiérrez, J. D. y Muñoz-Cadena, S. (2023a). Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia. *GIGAPP Estudios Working Papers*, *10*(270), 365-395.

Gutiérrez, J. D. y Muñoz-Cadena, S. (2023b). Assessing government design practices from a human-centered perspective: Case study of an improved cookstoves program in Colombia. *Desafíos*, *35*(1), 1-38.

Gutiérrez, J. D. y Muñoz-Cadena, S. (2023c). Building a repository of public algorithms: Case study of the dataset on automated decision-making systems in the Colombian public sector. En L. Belli y W. B. Gaspar (Eds.), *The quest for AI sovereignty, transparency and accountability: Official outcome of the UN IGF data and artificial intelligence governance coalition* (pp. 325-340). Getulio Vargas Foundation. https://www.intgovforum.org/en/filedepot_download/288/26421

Gutiérrez, J. D. y Muñoz-Cadena, S. (2025). Proactive algorithmic transparency in government: The case of the Colombian repositories of public algorithms. En *Handbook of Governance and Data Science*. Edward Elgar.

Gutiérrez, J. D., Muñoz-Cadena, S. y Castellanos-Sánchez, M. (2023). *Sistemas de decisión automatizada en el sector público colombiano* [Dataset]. Universidad del Rosario. https://doi.org/10.34848/YN1CRT

Gutiérrez, J. D., Muñoz-Cadena, S. y Corcione, M. C. (2024). Participación e incidencia de las agencias reguladoras en el ciclo de las políticas públicas: Caso de estudio comparado en Colombia. *Estudios de Derecho*, *81*(178), 131-166.

Gutiérrez, J. D. y Suárez, A. F. (2023). Using competition law to link regulation and development. *Law and Development Review*, *16*(1), 145-184. https://doi.org/10.1515/ldr-2022-0045

Haimovich, F., Vazquez, E. y Adelman, M. (2021). *Scalable early warning systems for school dropout prevention: Evidence from a 4000-school randomized controlled trial*. World Bank. https://openknowledge.worldbank.org/entities/publication/65141903-10be-5c76-8d74-a52f2b7afbc7

Hermosilla, M. P. y Germán, M. (2024). Implementación responsable de algoritmos e inteligencia artificial en el sector público de chile. *Revista Chilena de la Administración del Estado*, (11). https://doi.org/10.57211/revista.v11i11.185

Hernández, L. (2022, 5 de diciembre). "Chatico" es la nueva herramienta de WhatsApp para los bogotanos: Así funciona. *El Tiempo*. https://www.eltiempo.com/tecnosfera/apps/que-es-chatico-la-nueva-herramienta-de-whatsapp-para-los-bogotanos-723526

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2020). *The Assessment List for Trustworthy AI (ALTAI) for self-assessment*. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

Höchtl, J., Parycek, P. y Schoellhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*. https://doi.org/10.1080/10919392.2015.1125187

Howlett, M. y Cashore, B. (2014). Conceptualizing public policy. En I. Engelli y C. Rothmayr Allison (Eds.), *Comparative policy studies. Conceptual and methodological challenges* (pp. 17-33). Palgrave Macmillan. https://doi.org/10.1057/9781137314154_2

Hupe, P. L. y Hill, M. J. (2015). "And the rest is implementation": Comparing approaches to what happens in policy processes beyond great expectations. *Public Policy and Administration*, *31*(2), 103-121. https://doi.org/10.1177/0952076715598828

Instituto Humboldt. (2023, 23 de febrero). Escuchar para resguardar los bosques. http://www.humboldt.org.co/es/boletines-y-comunicados/item/1813-escuchar-para-resguardar-los-bosques

Knill, C. y Tosun, J. (2011). Policy-making. En D. Caramani (Ed.), *Comparative politics* (2.ª ed.) (pp. 373-388). Oxford University Press.

Medaglia, R. y Tangi, L. (2022). The adoption of artificial intelligence in the public sector in Europe: Drivers, features, and impacts. En *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance* (pp. 10-18). https://doi.org/10.1145/3560107.3560110

Ministerio de Educación de Perú. (2023). *Preguntas frecuentes alerta escuela para directores de instituciones educativas.* https://www.ugel02.gob.pe/file/25566/

Misuraca, G., Van Noordt, C. y Boukli, A. (2020, 23 de septiembre). The use of AI in public services: Results from a preliminary mapping across the EU. En *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance* (pp. 90-99). https://doi.org/10.1145/3428502.3428513

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Pelican.

Morales, J. D. (2020, 31 de enero). Tecnología para mejorar y reparar las vías de Bogotá. *El Tiempo*. https://www.eltiempo.com/tecnosfera/novedades-tecnologia/tecnologia-para-mejorar-y-reparar-las-vias-de-bogota-457436

Muñoz-Cadena, S., Gutiérrez, J. D., Castellanos-Sánchez, M. y Peralta, D. S. (2025). *Sistemas de IA en el sector público de América Latina y el Caribe* (versión 2.3) [Dataset]. https://sistemaspublicos.tech/sistemas-de-ia-en-america-latina/

Muñoz-Cadena, S. y Gutiérrez, J. D. (2025). *Repositorios de algoritmos públicos en el Mundo* (versión 2) [Dataset]. Sistemas de Algoritmos Públicos, Universidad de los Andes. https://sistemaspublicos.tech/otras-bases-de-datos-relacionadas-con-sistemas-de-algoritmos-e-ia/

Nakamura, R. T. (1987). The textbook policy process and implementation research. *Review of Policy Research*, *7*(1), 142-154.

Observatory of Public Sector Innovation (OPSI). (2018). Crecer con Salud: Virtual assistant for pregnancy and early childhood. https://oecd-opsi.org/innovations/crecer-con-salud-virtual-assistant-for-pregnancy-and-early-childhood/

Organización para la Cooperación y el Desarrollo Económico (OECD) y Banco de Desarrollo de América Latina y el Caribe (CAF). (2022). *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe*. OECD Publishing. https://doi.org/10.1787/5b189cb4-es

Oficina Consejería Distrital TIC. (2022a, 5 de febrero). Avanzan las pruebas del nuevo agente virtual de Bogotá, Chatico. https://tic.bogota.gov.co/node/322

Oficina Consejería Distrital TIC. (2022b, 26 de agosto). Ciudadanía podrá votar por sus "Causas Ciudadanas" a través de WhatsApp. https://tic.

bogota.gov.co/node/472#:~:text=La%20votación%20se%20realiza%20a,virtual%20de%20Gobierno%20Abierto%20Bogotá.

Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (OHCHR). (2023). *Taxonomy of human rights risks connected to generative AI: Supplement to B-Tech's foundational paper on the responsible development and deployment of generative AI*. https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf

Oré Arroyo, C. (2019, 23 de junio). Crean *app* para detectar anemia al tomar una foto del párpado inferior. *El Comercio*. https://elcomercio.pe/juntos-contra-anemia/crean-app-detectar-anemia-foto-parpado-inferior-lima-inteligencia-artificial-noticia-647883-noticia/?ref=ecr

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco). (2021). *Los sistemas de alerta temprana para prevenir el abandono escolar en América Latina y el Caribe*. https://unesdoc.unesco.org/ark:/48223/pf0000380354

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco). (2023). *Kit de herramientas global sobre IA y el estado de derecho para el poder judicial*. https://unesdoc.unesco.org/ark:/48223/pf0000387331_spa

Parker, I. y Davies, M. (2024, 18 de julio). AI in the public sector: White heat or hot air? *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/blog/ai-public-sector-white-heat-hot-air/

Peeters, R. y Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review*, *83*(4), 863-877. https://doi.org/10.1111/puar.13615

Pencheva, I., Esteve, M. y Mikhaylov, S. J. (2020). Big data and AI: A transformational shift for government: So, what next for research? *Public Policy and Administration*, *35*(1), 24-44. https://doi.org/10.1177/0952076718780537

Prensa Gobierno de Mendoza. (2022, 5 de octubre). Escuelas presentó el Sistema de Alerta Temprana (SAT). https://www.mendoza.gov.ar/prensa/la-dge-presento-el-sistema-de-alerta-temprana-sat/

Riviello, E. D., Dechen, T., O'Donoghue, A. L., Cocchi, M. N., Hayes, M. M., Molina, R. L., Moraco, N. H., Mosenthal, A., Rosenblatt, M., Talmor, N., Walsh, D. P., Sontag, D. N. y Stevens, J. P. (2022). Assessment of a crisis standards of care scoring system for resource prioritization and estimated excess mortality by race, ethnicity, and socially vulnerable

area during a regional surge in covid-19. *JAMA Network Open*, *5*(3), e221744-e221744. https://doi.org/10.1001/jamanetworkopen.2022.1744

Roth Deubel, A. N. (2002). *Políticas públicas: Formulación, implementación y evaluación* (1.ª ed.). Aurora.

Russell, S. J. y Norvig, N. (2004). *Inteligencia artificial: Un enfoque moderno* (2.ª ed.). Pearson Educación.

Santiso, C. y Cetina, C. (2022). *DIGIntegridad: La transformación digital de la lucha contra la corrupción*. Banco de Desarrollo de América Latina y el Caribe (CAF). https://scioteca.caf.com/handle/123456789/1901

Schrepel, T. y Groza, T. (Eds.). (2022). *The adoption of Computational Antitrust by agencies: 2021 Report*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4142225

Secretaría de Educación. (s. f.). *SART: Sistema de Alerta y Respuesta Temprana*. https://sart.se.gob.hn

Secretaría Distrital de Planeación. (2022, 13 de diciembre). Alcaldía de Bogotá presenta agente virtual que facilitará el acceso a servicios distritales y campañas de participación ciudadana. https://www.sdp.gov.co/node/30609

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. y Thompson, N. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv*. https://arxiv.org/abs/2408.12622

Superintendencia del Medio Ambiente de Chile. (2021, 17 de agosto). Alianza público-privada protegerá a zorro de Darwin a través de monitoreo basado en inteligencia artificial. https://portal.sma.gob.cl/index.php/2021/08/17/alianza-publico-privada-protegera-a-zorro-de-darwin-a-traves-de-monitoreo-basado-en-inteligencia-artificial/

Tangi, L., Combetto, M. y Martin Bosch, J. (Eds.). (2024a). *Methodology for the public sector tech watch use case collection: Taxonomy, data collection, and use case analysis procedures*. European Commission. https://doi.org/10.2760/078522

Tangi, L., Ulrich, P., Schade, S. y Manzoni, M. (2024b). Chapter 12: Taking stock and looking ahead. Developing a science for policy research agenda on the use and uptake of AI in public sector organizations in the EU. En *Research handbook on public management and artificial intelligence* (pp. 208-225). Edward Elgar Publishing. https://doi.org/10.4337/9781802207347.00023

Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D. y Pignatelli, F. (2022). *AI watch: European landscape on the use of artificial intelligence by the public sector*. Publications Office of the European Union.

Thibault, S. y Groza, T. (2023). *The adoption of Computational Antitrust by agencies: 2nd annual report*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4476321

Torres-Melo, J. y Santander, J. (2013). *Introducción a las políticas públicas: Conceptos y herramientas desde la relación entre Estado y ciudadanía*. Instituto de Estudios del Ministerio Público, Procuraduría General de la Nación. https://www.funcionpublica.gov.co/eva/admon/files/empresas/ZW1wcmVzYV83Ng==/imgproductos/1450056996_ce38e6d218235ac89d6c8a14907a5a9c.pdf

Unidad Administrativa Especial de Rehabilitación y Mantenimiento Vial (UAERMV). (2021, septiembre). *Mi Calle*, *79*.

Unidad Administrativa Especial de Rehabilitación y Mantenimiento Vial (UAERMV). (2022, 17 de mayo). *Transformación digital. SIGMA*. UAERMV.

Valle-Cruz, D., Criado, J. I., Sandoval Almazan, R. y Ruvalcaba Gómez, E. (2020). Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Government Information Quarterly*, *37*(4). https://doi.org/10.1016/j.giq.2020.101509

van Noordt, C. y Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3), 101714. https://doi.org/10.1016/j.giq.2022.101714

Wirtz, B. W. y Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review*, *21*(7), 1076-1100. https://doi.org/10.1080/14719037.2018.1549268

Zimic, M. (2019). Sistema portátil para el diagnóstico de anemia basado en el análisis de la conjuntiva ocular usando un smartphone e inteligencia artificial. *Evidencia MIDIS*. https://evidencia.midis.gob.pe/wp-content/uploads/2019/11/MEM_Zimic.pdf

Zuiderwijk, A., Chen, Y.-C. y Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577

IMAGINARIOS SOCIOTÉCNICOS Y PRÁCTICAS ANTICIPATORIAS EN EL CUBRIMIENTO MEDIÁTICO DE LA INTELIGENCIA ARTIFICIAL Y SU RELACIÓN CON EL ESTADO EN COLOMBIA*

Miller Díaz-Valderrama, Natalia Niño-Machado, Javier Guerrero-C., Catalina González-Uribe

* Este trabajo es resultado de las reflexiones de los siguientes proyectos: Enhancing Tools for Response, Analytics and Control of Epidemics in Latin America and the Caribbean (TRACE-LAC), financiado por el International Research Centre (IDRC) de Canadá [Grant n.° 109848-001]; y El Futuro de los Derechos Humanos, financiado por la Fundación Botnar [Grant n.° REG-22-002]. El IDRC y la Fundación Botnar no tuvieron ningún papel en la recolección o el análisis de datos, ni en la escritura de este manuscrito.

Para citar este capítulo:
http://dx.doi.org/10.51573/Andes.9789587988444.9789587988451.9789587988468.10

Introducción

La inteligencia artificial (IA) es considerada una tecnología disruptiva, la cual, empleada por el Estado, promete una transformación sin precedentes de este y de la sociedad. La promoción de la IA por parte del Estado es de nuestro interés precisamente por la capacidad de esta tecnología —y la recolección de datos asociada a su desarrollo— de reorganizar la vida social, los mercados y las funciones provistas por los Estados. La IA promete servicios que son a la vez altamente individualizados y flexiblemente diferenciados, según las demandas de los ciudadanos. Las promesas de la IA en relación con el Estado es el punto de partida de este capítulo.

Así, rastreamos en el discurso mediático imaginarios sobre la IA y su relación con el Estado e identificamos aquellas acciones estatales que surgen como una respuesta anticipatoria a la creciente incursión de la IA en la sociedad colombiana. Como los medios de comunicación no solo reportan o reflejan el sentido de la vida social, sino que contribuyen activamente en su producción, nuestro análisis se orienta a comprender las visiones de futuro que allí se construyen sobre el Estado y sus potenciales transformaciones, teniendo en mente dos intereses: por un lado, ofrecer una descripción general de cómo se está construyendo la imaginación sobre este tema; y, por otro, analizar cómo los imaginarios de esta relación hacen eco de visiones localizadas sobre el Estado en Colombia, su modernización y otros proyectos tecnológicos.

El concepto que guía nuestra exploración es *imaginarios sociotécnicos*. Este se define como las "formas de vida social y de orden social imaginadas colectivamente que se reflejan en el diseño y la realización de proyectos científicos

y/o tecnológicos" (Jasanoff y Kim, 2009, p. 120), y que son sostenidas de forma colectiva, institucionalmente estabilizadas y desplegadas en público. Estos imaginarios ayudan a organizar grupos y crear alianzas, son múltiples y coexisten en tensión dentro de una sociedad, siendo algunos foros públicos (como las legislaturas, las cortes o los medios de comunicación) los lugares donde se priorizan o superponen unas visiones sobre otras, construyendo posiciones dominantes (Jasanoff, 2015).

Los estudios sociales de la ciencia y la tecnología han empleado este concepto para examinar las formas en las que tecnologías emergentes como la IA desempeñan un papel fundamental al prever y determinar sus trayectorias de desarrollo, agendas de investigación y usos (Richter *et al*., 2023). El concepto de imaginarios sociotécnicos va más allá de las representaciones, actitudes y percepciones sociales y populares acerca de una tecnología. Su escala no es individual, ni su atención está puesta en la agregación de visiones individuales para comprender tendencias en la percepción.

En el presente capítulo adoptamos esta perspectiva conceptual para preguntarnos: ¿qué tipo de situaciones están siendo definidas como reales o posibles por determinados grupos alrededor de discusiones sobre la IA y su relación con el Estado? ¿Qué imaginarios son posicionados como hegemónicos o dominantes? ¿Qué respuestas anticipatorias reclaman del Estado colombiano y cómo estas prometen transformarlo? Y, por último ¿qué imaginarios son opacados/negados en estas versiones particulares de la imaginación sociotécnica?

Los imaginarios sociotécnicos y la anticipación

Quienes producen métodos, proyecciones numéricas, infraestructuras tecnológicas y políticas para tomar decisiones imaginan un porvenir en el que ciencia, tecnología y sociedad se dan forma mutuamente, para enfrentar desafíos presentes y problemas que incluso aún no tienen lugar. La creación de marcos regulatorios exige previsión, y la gobernanza de escenarios anticipados se ha constituido en la acción por excelencia de los Estados contemporáneos, así como en la razón principal de su justificación (Wenger *et al*., 2020).

La puesta en marcha de la imaginación a futuro es, a la vez, técnica y social, pues involucra el despliegue de "nuevos conjuntos específicos de infraestructuras tecnológicas materiales, significados sociales y órdenes morales, todo ello en torno a nuevas formas de políticas de la información" (Felt, 2015, p. 177). Implícita o explícitamente, proyectos tecnocientíficos y burocráticos se proponen ajustar órdenes tecnológicos y de conocimiento, como la creación de infraestructuras

materiales, al igual que órdenes sociales, que conciernen valores, identidades y la asignación de responsabilidades sobre individuos y organizaciones.

¿De qué manera estos ordenamientos se llevan a cabo? Esta es una pregunta tácita del diseño tecnológico (Latour, 1990), y muchas veces se encuentra explícita en las políticas públicas orientadas a la creación de infraestructuras tecnológicas. Ambas acciones, diseños tecnológicos y formulación de políticas, elaboran distintos imaginarios sociotécnicos.

Los imaginarios sociotécnicos involucran la definición de lo que puede lograrse a través de la ciencia y la tecnología, y visiones "sobre cómo la vida debe o no debe ser vivida; así, expresan visiones compartidas del bien y del mal en una sociedad" (Jasanoff, 2015, p. 4). Los imaginarios sociotécnicos prescriben relaciones sociales y desarrollos tecnológicos, produciendo acciones que anticipan y responden a escenarios futuros.

La idea de anticipación es central para la operacionalización de este concepto, pues pretende hacer visible la forma en la que tales imaginarios prefiguran relaciones sociales en su ejecución, y también las maneras en las que promueven "prácticas anticipatorias", es decir, acciones especulativas presentes orientadas a responder a un futuro esperado. Los imaginarios sociotécnicos hablan de visiones de futuros deseables —o *aspiraciones*— y de las formas en las que tales visiones se traducen en prácticas y actitudes presentes, esto es, en *performances* (Schmid *et al*., 2022).

De la experticia y la política al cubrimiento mediático

Existen distintos puntos de partida para asir los imaginarios sociotécnicos sobre la IA. En la literatura pueden encontrarse tres tipos de enfoques: aquellos centrados en los expertos (Hautala y Ahlqvist, 2024; Law, 2023; Natale y Ballatore, 2017), los que prestan atención a los imaginarios explícitos e implícitos en documentos regulatorios y de política pública (Bakiner, 2023; Bareis y Katzenbach, 2021; Chan, 2021; Paltieli, 2022; Pham y Davies, 2024) y los que se centran en medios de comunicación. Este último cuerpo de literatura ha venido creciendo en los últimos años, presentando distintos debates sobre cómo la prensa cubre el tema de la IA y qué imaginarios futuros construye alrededor de su entrada en escena.

Por un lado, estos estudios prestan atención a los imaginarios sociotécnicos que se construyen alrededor de la promoción (*hype*) de la IA en los medios como tecnología emergente. Este cuerpo de literatura se dedica a entender las expectativas ubicuas que emergen alrededor del uso vago del concepto de IA en el debate público y cómo las noticias *median, influencian* y *amplifican* tales

expectativas (Brennen *et al*., 2020). Una forma de abordar esto es explorando la disparidad entre el cubrimiento mediático de la IA, en términos generales, y las noticias que se enfocan en sus aplicaciones (Züger *et al*., 2023). Mientras que el primer escenario tiende a conformar visiones de futuro optimistas y de *solucionismo tecnológico*, donde limitaciones y riesgos actuales (definidos en términos vagos) serán subsanados en un futuro cercano (Obozintsev, 2018), el segundo hace posible que surjan visiones críticas sobre los alcances, posibilidades y límites de la IA, al entender que tal nombre agrupa un grupo heterogéneo de tecnologías y que las promesas amplificadas pueden transformarse en desilusiones igualmente magnificadas (Hansen, 2022; Züger *et al*., 2023). Esta coexistencia y transformación de expectativas contradictorias ha sido muy bien estudiada en otras tecnologías emergentes, categorizadas bajo el concepto de ciclos de promesa y decepción o *hype cycle* (Sovacool y Hess, 2017; Van der Maarel *et al*., 2023).

En términos generales, el uso vago del concepto de la IA en el discurso público permite una visión futura promisoria, en la que el número de roles y potenciales que puede tener es tan abierto como difuso (Vrabič Dežman, 2024), siendo la principal función de este tipo de cobertura mediática, no ofrecer un contenido robusto sobre el tema, sino posicionar un asunto como noticioso, relevante para un contexto social amplio (Roberge *et al*., 2020).

En este sentido, las noticias periodísticas desempeñan un papel importante en definir agendas del debate público, en delimitar aquello que merece atención, y es usual que lo haga al promover los beneficios de nuevas tecnologías, enmarcándolas en una visión sobre el progreso científico o sobre los beneficios económicos que pueden implicar. Al respecto, los imaginarios sociotécnicos sobre la IA en la prensa también presentan similitudes con el tratamiento de otras tecnologías emergentes (Brennen *et al*., 2020), y no es de extrañar que las visiones de futuro optimistas predominen su cubrimiento mediático (Fast y Horvitz, 2016; Roberge *et al*., 2020). Estos imaginarios optimistas están principalmente enmarcados en el abordaje de temas como los beneficios económicos y de negocios, y la investigación en ciencia y tecnología (Chuan *et al*., 2019).

Por este motivo, otro cuerpo importante de literatura es aquel dedicado a explorar las formas en las que grupos de interés como la industria dominan el debate público sobre la IA y sus discursos controlan los imaginarios sociotécnicos que la prensa construye alrededor de su futuro (Brennen, 2018). Estos discursos e imaginarios también permean las agendas que actores gubernamentales (Maldonado y Arroyave, 2024) despliegan en sus opiniones y promociones de la IA en los medios, lo que suscita imaginarios sobre el desarrollo económico y el posicionamiento nacional favorable en un contexto global (Köstler y Ossewaarde, 2022).

Así, es usual que la IA sea presentada como prometedora, en lo que Pham y Davies (2024) describen como la imaginación de una *carrera global de suma cero*, construyendo un escenario en el que, independientemente del contexto social o económico actual, las naciones se enfrentan a una oportunidad para solucionar cualquier problema y ponerse al frente de una revolución tecnológica y económica sin precedentes (Brennen, 2018; Brennen *et al*., 2020; Köstler y Ossewaarde, 2022). Sin embargo, esto tampoco es un aspecto *sui generis* de los debates públicos sobre la IA, es algo común en el cubrimiento de otras tecnologías (Donk *et al*., 2011), donde los imaginarios de promesas económicas, productividad y cambio de las reglas de juego en la "competencia internacional" han influido tradicionalmente en la producción de política y la regulación tecnológica (Ulnicane *et al*., 2021).

Muchos de estos imaginarios se construyen por medio de metáforas, como el uso de la palabra *revolución* para hablar de las transformaciones imaginadas por el uso extendido de tecnologías de IA en diferentes campos de aplicación. Por este motivo, algunos estudios se ocupan de las metáforas con las que se nombran los beneficios y riesgos de la IA —o la IA misma—, y que ayudan a construir visiones de futuro. Tales estudios toman las metáforas no como simples descripciones del mundo, sino como mecanismos discursivos que llevan consigo connotaciones normativas (Wyatt, 2004).

El trabajo de Sally Wyatt (2021a, 2021b), que explora las formas en las que metáforas construyen imaginarios futuros para tecnologías como la internet y el *big data*, ha sido particularmente influyente en este tipo de estudios. La importancia de esta exploración radica en que las metáforas que se dan por sentado constituyen sustitutos del pensamiento (McCloskey, 1983; Wyatt, 2021a), es decir, encierran asuntos de hecho difíciles de cuestionar —si no son examinados— con una potente fuerza retórica. Las metáforas funcionan como una suerte de cajas negras. La importancia de estos estudios es poner en cuestión el carácter oscuro y de autojustificación que estas formas de hablar sobre la IA suponen, las cuales dificultan la deliberación pública y democrática sobre las aplicaciones deseables de la IA, creando visiones de inevitabilidad (Bones *et al*., 2021; Kajava y Sawhney, 2023). A la vez, permiten prestar atención a las versiones alternativas imposibilitadas, negadas o socavadas por estos recursos narrativos (Suchman, 2008).

Esta indagación sobre las imaginaciones metafóricas ha tenido más espacio en el abordaje de imaginarios generalizados sobre la IA, donde la equivalencia entre IA y mente, así como entre mente y máquina, son centrales en la comprensión pública de este grupo de tecnologías y constituyen narrativas engañosas que opacan la discusión sobre sus limitaciones actuales; a su vez, son la

base de la atención noticiosa —*hype*— que recibe (Campolo y Crawford, 2020; Natale y Ballatore, 2017).

Este tipo de metáforas también han sido exploradas en documentos de política pública, como la ley de inteligencia artificial de la Comisión Europea, en la que el uso de metáforas como el "nuevo petróleo" o la "mina de oro" circunscriben el futuro del uso de la IA en términos de riqueza económica (Marčetić y Nolin, 2023, p. 107). Estas metáforas también han sido analizadas en el debate público en un sentido amplio (Roberge *et al*., 2020); no obstante, su aproximación desde el cubrimiento mediático ha sido tangencial, llamando en especial la atención sobre la necesidad de generar nuevas formas de nombrar e imaginar futuros para la IA, más allá de la hegemonía discursiva establecida por la industria (Mager y Katzenbach, 2021).

Así mismo, este llamado es una muestra de la necesidad de estudios que enfoquen su atención en las metáforas que forman imaginarios sociotécnicos sobre la IA por parte de actores heterogéneos, incluidas las organizaciones estatales y la sociedad civil. Así, el cubrimiento mediático de la IA en relación con servicios estatales (o cuya responsabilidad ha sido tradicionalmente asignada al Estado) puede representar un terreno apropiado para permitir esta reflexión, al escapar de la tendencia que enmarca el debate sobre la IA en términos de inevitabilidad. Como señalan Mager y Katzenbach (2021):

> En términos más generales, el entusiasmo actual parece sugerir que la IA es inevitable y que cambiará fundamentalmente nuestra forma de vivir, comunicarnos, trabajar y viajar. Si bien estas afirmaciones son claramente producto de una exageración contingente, tienen poderosos efectos en la forma en que estructuran a los actores y los recursos. (p. 232)

Por ello, se hace necesario pensar cómo este entusiasmo está siendo enmarcado cuando se discute el papel del Estado, pues muchas de las promesas actuales de la IA se expanden al mejoramiento de la administración pública (Esko y Koulu, 2023).

En particular para países del sur global, los imaginarios sobre la transformación digital de los Estados requieren ser analizados. Por un lado, las promesas de mejorar la eficiencia del Estado y resolver problemas estructurales de sociedades afectadas por una fuerte desigualdad necesitan una comprensión sopesada; y, por otro, es pertinente pensar en cómo imaginarios generalizados necesitan de una mayor contextualización/localización, sensible a las historias y especificidades de estos países (Barreneche *et al*., 2021). Al hablar de imaginarios se deja abierta la posibilidad de divisar futuros digitales alternativos, que tengan en cuenta a grupos tradicionalmente marginalizados (Suárez-Estrada y

Lehuede, 2022) y que entren en discusion con las lógicas coloniales de los imaginarios sociotécnicos del norte global (Ricaurte *et al*., 2024).

Metodología

En este capítulo presentamos un análisis de los imaginarios sobre la IA, su relación con el Estado y las prácticas anticipatorias que esto supone, a través de una revisión de notas de prensa publicadas en dos periódicos tradicionales de Colombia, *El Tiempo* y *El Espectador*, así como en los comunicados de prensa del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), entre mayo del 2023 y junio del 2024. De igual manera, en las notas periodísticas de MinTIC encontramos evidencias de acciones anticipatorias o de gobernanza anticipada (Wenger *et al*., 2020) para responder a estos imaginarios.

Nos concentramos en el periodo comprendido entre mayo del 2023 y julio del 2024. Entendemos que la construcción de un imaginario no tiene momentos puntuales de inicio y fin, pero nos interesa explorarlos de forma condensada en un lugar como la prensa, que puede movilizar percepciones colectivas e imaginarios públicos sobre la IA y han tenido un papel decisivo en hacer de este tema un asunto de interés público (Hansen, 2022).

Seleccionamos el mes de mayo del 2023 por ser la fecha oficial de finalización de la pandemia por covid-19, un tema que ocupó gran parte del interés mediático y que fue empleado por algunos actores para movilizar la idea de la necesidad de la transformación digital del Estado. Por otro lado, aunque la discusión pública sobre la IA contempla un periodo más extenso, tras el lanzamiento de ChatGPT a finales del 2022, el 2023 se convierte en un año de discusión pública sobre los alcances, las posibilidades y los futuros de la IA generativa; y tras las distintas instancias de discusión y acuerdos sobre la ley de IA en la Unión Europea, el 2024 se configuró como el año de discusiones sobre las regulaciones de este grupo de tecnologías.

Dentro del periodo señalado encontramos 220 notas —*El Espectador* (85), *El Tiempo* (80) y MinTIC (55)— en Colombia. Se seleccionaron *El Espectador* y *El Tiempo* por ser dos periódicos tradicionales en Colombia que han ocupado un lugar predominante en la producción de opinión pública en el país por décadas; además, son dos de los medios con mayor circulación nacional, influencia y de cercanía al poder[1].

1 De acuerdo con el informe Reuters (Newman *et al*., 2024), *El Tiempo* y *El Espectador* son los periódicos de mayor confianza para los encuestados, en su estudio de consumo de noticias digitales en el mundo.

En la selección de notas nos centramos en noticias que discutieran impactos sociales, económicos y gubernamentales, y que significaran de alguna manera un llamado sobre acciones estatales como el *uso*, la *regulación*, la *capacitación* y la *gestión* gubernamental de la IA.

Clasificamos cada una de las noticias por fecha, tema principal de discusión, tipo de acción estatal sobre el que llama la atención y percepción general de la IA (optimista, neutral o pesimista). De igual manera, ordenamos el contenido de las noticias por fragmentos, considerando aspectos como fuente, fecha y categorías de codificación emergentes, y se prestó especial atención a las metáforas empleadas para definir la IA, sus posibilidades, alcances y limitaciones.

Para apoyar el análisis de narrativas e identificación de imaginarios sociotécnicos, utilizamos el marco de análisis explorado por Guenduez y Mettler (2023) en la revisión de narrativas de políticas públicas, adaptado al análisis de imaginaros sobre Estado e IA en prensa. Esta adaptación (tabla 10.1) consistió en tomar tres de los elementos base del análisis de narrativas (contexto, moral y trama), manifestarlos en términos de imaginarios colectivos de futuros y dar cuenta de los elementos que configuran la posibilidad de la imaginación, la forma de expresarla y las relaciones que establece. De igual manera, como este tipo de análisis se enfoca en la construcción de personajes, sus relaciones y la atribución de responsabilidades, en términos de la imaginación sociotécnica fue provechoso identificar cómo las distintas formulaciones de la imaginación atribuyen responsabilidad al Estado, qué acciones demanda de este y qué versiones implícitas socava.

Este marco nos permite, en primer lugar, a través de la descripción de aspectos generales de las noticias y la percepción de la IA, mostrar cómo se configura un debate predominantemente optimista. En segundo lugar, revelar cómo la innovación es un tema central y transversal a las noticias que hablan de los distintos roles del Estado en respuesta a la IA, y cómo este tema conforma los diferentes campos de la imaginación que identificamos. En tercer lugar, presentar los campos identificados, sus imaginarios asociados, los argumentos que los constituyen, las acciones anticipatorias que promueven y las versiones alternativas que socaban.

Tabla 10.1. Elementos narrativos para la identificación de imaginarios sociotécnicos

Marco de narrativas	Elementos narrativos para la identificación de imaginarios sociotécnicos sobre Estado e IA
Contexto Provee el contexto en el que se cuenta una historia, en la que se incluyen teorías y suposiciones, que son usualmente dadas por sentado, aunque sean controversiales en algún punto para algunos.	Marco de imaginación Provee el campo sobre el cual se construyen imaginarios sociotécnicos. Dispone los límites de la imaginación, habilitando un lenguaje común —metáforas y otros dispositivos narrativos— para la discusión sobre desarrollos e innovaciones tecnológicas.
Moral Se presenta como una solución política o una llamada a la acción.	Imaginarios Se expresan en formulaciones, declaraciones o descripciones sobre una tecnología, su desarrollo e impacto social, las cuales prescriben su futuro.
Trama Vincula personajes y contextos, asignando responsabilidades.	Argumento Consiste en la expresión de los imaginarios en la descripción de ordenamientos técnicos, sociales y morales, que asignan responsabilidades a actores particulares y proscriben, imposibilitan o minimizan versiones alternativas del orden social.
Personajes y roles gubernamentales Para la identificación de narrativas, personajes descritos como víctimas, villanos y héroes son fundamentales, sean estos individuos, grupos u organizaciones.	Acción del Estado y acción anticipatoria Cómo los imaginarios asignan responsabilidades a actores como el Estado, organizaciones y sujetos determinados. Esta clasificación ayuda a identificar la forma en la que el papel del Estado es enmarcado por los imaginarios y las acciones anticipatorias que los argumentos promueven.
	Versiones alternativas Son las versiones de los imaginarios y argumentos alternativos proscritos por los imaginarios posicionados como dominantes. La lógica misma del campo de la imaginación.

Fuente: elaboración propia con base en Guenduez y Mettler (2023, p. 5).

Un panorama especulativo optimista y centrado en la innovación

La idea de la innovación es un tema transversal al cubrimiento mediático sobre la IA. El análisis de las notas de prensa nos permitió identificar cuatro campos de la imaginación en relación con la IA: (1) innovación vs. regulación; (2) cuarta revolución industrial; (3) primicia, liderazgo y modernización regional; y (4) desarrollo y presencia nacional amplificada.

De las 165 entradas de periódicos, el 52 % fue publicado entre mayo y diciembre del 2023 (ocho meses) y el 48 %, en los primeros seis meses del 2024. Para las notas de prensa de MinTIC, el 40 % se publicó en el 2023 y el 60 %, en el primer semestre del 2024. Aunque evidentemente casual, la diferencia entre el primero —"Inteligencia artificial: ¿perderá su empleo?"— y el último titular —"Salud digital: poder transformador"— encontrados en nuestra indagación muestra un cambio del discurso en relación con la IA.

Hay una diferencia importante en el cubrimiento de este asunto de interés en las noticias de periódicos nacionales y las notas de MinTIC: estas últimas tienen la función de demostrar o documentar las acciones del Gobierno sobre la transformación digital del Estado, la promoción de la IA y la gestión estatal

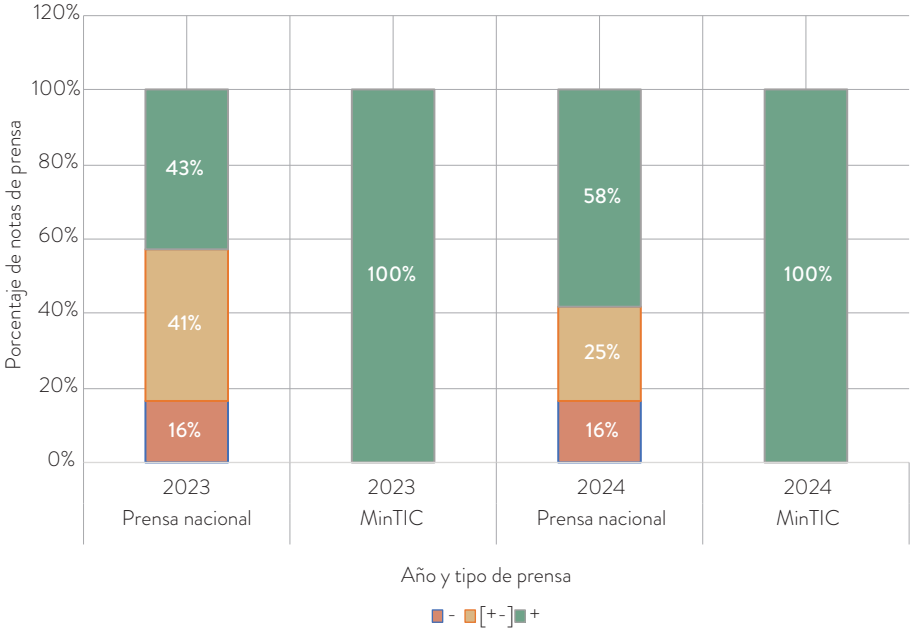


Figura 10.1. Percepción de la IA por tipo de fuente y año

Fuente: elaboración propia a partir de la recolección y clasificación de notas de prensa.

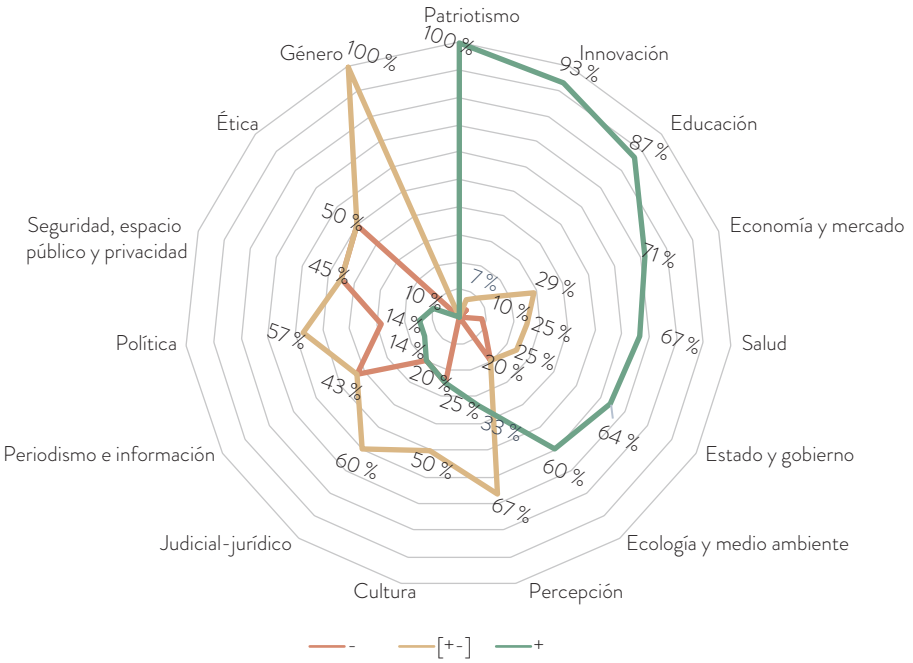


Figura 10.2. Percepción de la IA por categoría temática
Fuente: elaboración propia a partir de la recolección y clasificación de notas de prensa.

que se ha llevado a cabo para transformar a Colombia en lo que el Ministerio ha denominado como una "potencIA mundial", un centro de desarrollo económico y tecnológico importante gracias a las herramientas de IA. En ese sentido, no sorprende que las notas reportadas por MinTIC son absolutamente optimistas, mientras que la proporción de noticias optimistas revisadas en la prensa nacional fue del 50 %, las neutrales de un 33 % y las negativas de un 16 %.

Campo de la imaginación 1: innovación vs. regulación

Este campo delimita la discusión sobre el papel regulador del Estado, donde la necesidad de regular se deriva de imaginar una carrera global por definir los riesgos, límites y posibilidades de la IA; sin embargo, en este campo la regulación es subsidiaria de la innovación. La acción anticipatoria que promueve esta discusión es precisamente la adecuación de marcos éticos y normativos sobre la producción, la explotación de datos y el desarrollo de IA. Recordemos que gran parte de la cultura de las empresas digitales ha estado vinculada a la idea de la *innovación disruptiva*, un concepto asociado al *creative destruction* de Schumpeter

y reinterpretado en Sillicon Valley para promover ideas como "moverse rápido y romper cosas", lema de Zuckenberg en los primeros años de Facebook.

Imaginario: la regulación de la IA es una carrera global, y es necesario posicionarse

Dado que el escenario regulatorio se presenta igualmente incierto para todos los Estados, actuar en esta materia se percibe como un paso necesario para que el país pueda convertirse en "productor" y no solo en "consumidor" de IA, colocando un "sello propio" en el universo de las regulaciones actuales (García Rico, 2024). En este marco, no actuar en la regulación se percibe como un contexto propicio para el rezago y el aumento de las desigualdades entre norte-sur: "quedarnos atrás en este contexto podría generar una nueva brecha creativa y económica entre el norte y el sur global que profundice la precarización y la relación desigual" (Rangel, 2024).

Imaginario: la regulación no reactiva promueve la innovación

El imaginario de una carrera global por la regulación tiene como consecuencia la definición de la IA como una tecnología cuyo avance es irreversible y cuyos impactos son tales que no anticiparse en la definición de normas y marcos éticos supone un riesgo considerable. Pero, a la vez, supone que es necesario centrarse más en los beneficios que en los riesgos, para no "quedarse atrás". Por esto, la regulación se pone en términos de "proactividad" en oposición a la "reactividad" (García, 2023): "regulación no es prohibición. Regular es colocar unos acuerdos mínimos, unos pilares fundamentales" (Chacón Orduz, 2024). El imaginario del rol regulador del Estado es aquí, entonces, promover la innovación a través de la legislación, en lugar de responder a riesgos presentes o previsibles, pues "exagerar en la regulación de la IA también puede sofocar la innovación" (Rueda, 2023).

El campo de la imaginación conformado por estos dos imaginarios deja de lado la discusión sobre la posibilidad de una innovación controlada, es decir, "abordar la inteligencia artificial como herramienta y no como fin por sí misma. Mirarla en su contexto, dentro del tema concreto" (Ojeda, 2024). De igual manera, presume que la regulación debe enfocarse en aspectos mínimos, siempre que su fin sea la innovación. El afán regulatorio, por ejemplo, en lugar de estar centrado en la producción e innovación, podría estarlo en definir cómo las sociedades desearían hacer uso de la IA: "Lo claro es que el mundo está corriendo a regular, con la influencia de las grandes empresas tecnológicas

buscando intervenir. El problema es que las implicaciones en juego son difíciles de medir" ("Llegamos tarde y sin claridad a ver qué hacer con la inteligencia artificial", 2023).

Por último, pensar en una regulación de riesgos futuros pierde de vista la necesidad de regular riesgos y problemas actuales, como el uso de la IA en organizaciones estatales. Aunque este ha sido un aspecto señalado por algunos actores, está lejos de formar parte de la imaginación hegemónica replicada por MinTIC, para quien la prioridad es la definición de límites para la regulación misma, más que la definición de límites para la IA:

> hay que saber hasta qué punto esta regulación, que saldría de los Gobiernos, iría en contra de la innovación, por lo que la discusión que se abre es dónde debe estar ese límite. Es un debate que hay que dar de inmediato, para buscar un balance entre desarrollo tecnológico y uso responsable. (MinTIC, 2024i)

Campo de la imaginación 2: cuarta revolución industrial

Este campo enmarca las discusiones sobre el uso de la IA desde su enunciación como *revolución*, como evento histórico de transformación social, tecnológica y económica.

Imaginario: la IA es una revolución tecnoeconómica sin precedentes y con impacto innovador en innumerables campos de acción

Aquí se destaca la definición de la IA como una revolución tecnoeconómica sin precedentes, aunque similar a la primera revolución industrial, con la capacidad de transformar cualquier área de la vida social: "Hoy día, con el avance acelerado de la tecnología, estamos observando y siendo partícipes de una transformación de la vida y condición humana sin precedentes" (Montoya Castaño, 2024). Como a ninguna otra tecnología, este imaginario le atribuye a la IA ser "la primera tecnología que va a cambiar la historia, la cultura y las relaciones sociales de la humanidad" (MinTIC, 2023b). Sin embargo, este poder transformador, cuando tiene que ser delimitado, es puesto en términos económicos y de producción: "la IA puede mejorar la eficiencia y la productividad en una gran variedad de campos, desde la salud hasta la industria manufacturera"; y de innovación, como lenguaje general para hablar de los beneficios de este grupo de tecnologías:

> En resumen, la IA tiene muchas bondades y oportunidades que pueden ser beneficiosas para la humanidad. Podemos utilizarla para mejorar la eficiencia y la productividad, abordar los desafíos sociales y ambientales, fomentar la creatividad y la innovación, y mucho más. (Maussa, 2023)

Imaginario: subirse al bus de la IA o quedarse atrás

Al atribuirle a la IA un potencial sin igual, el riesgo de no participar en su producción se concibe también de una magnitud excepcional. El llamado al Estado es el de "subirse al bus", posicionarse cuanto antes en la "carrera global" por recoger sus beneficios económicos, que son siempre mencionados en términos globales (como un botín para quien llegue primero en la competencia):

> hay una "fuerte carrera" por el liderazgo global, encabezada principalmente por Estados Unidos, China y la Unión Europea. En esta última, el Parlamento estima que para 2030 la IA contribuirá con más de 11 000 millones de euros a la economía mundial, con un crecimiento en el PIB europeo del 20 %. (Gutiérrez, 2023)

En este imaginario, la IA es inevitable, domina cualquier forma de futuro y no hay manera de responderle sino adaptándose:

> El mundo avanza a pasos agigantados y aquellos que no utilizan la tecnología están destinados a quedar obsoletos [...] [hay una] imperiosa necesidad de adaptarnos a los rápidos avances tecnológicos, centrándonos especialmente en el potencial transformador de la inteligencia artificial. (Méndez, 2023)

Desde esta forma específica de imaginación, el Estado colombiano debería promover la innovación para todo tipo de necesidades, "y así consolidar un papel relevante en la escena TIC mundial" (Redacción Economía y Negocios, 2024a).

Este campo de la imaginación limita la discusión sobre los beneficios de la IA al bienestar económico, el aumento de la productividad y la eficiencia, imposibilitando el diálogo sobre los límites actuales y propios de la IA, los contextos de desigualdad económica (pues la idea de una carrera global elimina cualquier desigualdad previa) o los beneficios reales y estimables para un país:

> Para 2030, según una investigación de PwC, la IA podría aportar hasta USD$15,7 billones a la economía mundial. Pero incluso si su protagonismo sigue siendo fuerte en los próximos años, es difícil predecir exactamente qué formas y aplicaciones de la IA proporcionarán un valor comercial atractivo. (Redacción Especiales, 2024)

De igual modo, al imaginar la IA como un agente transformador de todos los aspectos sociales, no se concibe la posibilidad legítima de que algunos grupos y sectores no utilicen la IA.

Campo de la imaginación 3: primicia, liderazgo y modernización regional

Aquí el campo de la imaginación se configura alrededor de las ideas de *primicia* y *liderazgo* en la transformación digital de la sociedad colombiana, a través de la capacitación en habilidades de analítica de datos e IA. Estas habilidades se imaginan como una oportunidad para la modernización regional, así como para lograr una profesionalización y una capacitación laboral de grandes magnitudes, elementos que prometen la generación de nuevos empleos.

Imaginario: la IA es una solución tecnoeconómica promisoria para superar las desigualdades regionales

Al igual que los campos de la imaginación que describimos, este imaginario se nutre de la idea de una carrera por el dominio de la IA, pero lo hace en escalas de municipios, departamentos y regiones. Tal como se ve una oportunidad para posicionarse en la escena económica global gracias a la IA, las distintas regiones de Colombia se imaginan en igualdad de condiciones para sumarse a la competencia nacional y global, mediante la creación de centros de formación de IA:

> hoy nos estamos ubicando en el mapa internacional como un referente que cierra la brecha de tecnología en la ciudad. Nos preparamos para consolidar esa Barranquilla global, con los desafíos y retos del mundo actual, y de esta forma seguir con el firme compromiso de brindar oportunidades equitativas para todos nuestros ciudadanos", expresó el alcalde Alejandro Char. (Díaz Ospino, 2024)

De igual manera, los beneficios de la IA se perciben tales que pueden impactar la cotidianidad de los colombianos, superando escenarios de empobrecimiento y desigualdad: "a través del fortalecimiento de la IA, el país logrará grandes avances en la lucha contra la pobreza y la desigualdad" (MinTIC, 2023a).

Imaginario: la IA es una oportunidad de profesionalización y generación de empleo

Aunque el trabajo es uno de los aspectos de la IA sobre los que más se presume sentir temor, es usualmente minimizado en este campo de la imaginación, al contrarrestar la posibilidad de la pérdida de empleo con la de su generación. Así, este imaginario utiliza la idea de la IA como revolución tecnoeconómica para asegurar que temores sobre el reemplazo por las máquinas son tan viejos como la primera revolución industrial y que, como en aquella, esta significará un avance en el desarrollo tecnológico y social: "los reemplazos de labores liderados por maquinarias siempre tienden a ser oportunidades para mejorar la capacidad productiva de las economías (lo que se traduce en mayores oportunidades de desarrollo), así como impulsores de nuevas disciplinas y profesiones" (Redacción Economía y Negocios, 2024c).

En este imaginario, la idea de modernización ocupa un lugar central, en especial en el cuerpo de notas periodísticas de MinTIC, pues en la imaginación a futuro de poblaciones capacitadas, el involucramiento de la juventud y la niñez en la alfabetización digital y la adecuación de espacios para el aprendizaje de habilidades útiles para la producción de IA se ve como una forma de ponerse al día con el desarrollo:

> El propósito es que los jóvenes se formen con nuestros programas. (MinTIC, 2024h)
>
> Necesitamos que nuestros niños y niñas aprendan el lenguaje de las máquinas. (MinTIC, 2024e)
>
> Para masificar los conocimientos sobre inteligencia artificial, generando competencias digitales y formación académica en áreas técnicas, para la producción de contenidos digitales, aplicaciones y desarrollo de software. (MinTIC, 2024c)

Este campo de la imaginación da por sentado la necesidad de capacitación en habilidades de analítica de datos e IA para toda la población, proscribiendo así la discusión sobre la pertinencia o no de generar transformaciones en áreas de desempeño profesional específicas, las dificultades que el desarrollo de este tipo de habilidades supone para la población mayor o la posibilidad de que las desigualdades regionales se acentúen.

Campo de la imaginación 4: desarrollo y presencia nacional amplificada

Por último, este campo de la imaginación define el rol de gestión del Estado en términos de una amplificación de su presencia. Aquí, el principal cuerpo de noticias es el de MinTIC, que se enfoca en la gestión realizada por el Gobierno nacional para llevar la IA a distintas regiones o para posicionar distintos espacios locales como "pontencIAs" mundiales, tras la creación de centros de IA o simplemente poniendo al día a algunas regiones con la conectividad y acceso a computadores.

Imaginario: la IA es una promesa de completitud para una modernidad inconclusa

Este imaginario trata la adecuación tecnológica de espacios para la producción de IA igual que otros proyectos de infraestructura —por ejemplo, la creación de carreteras en Colombia— como vía para el desarrollo. Así, la IA es definida en su potencial de superar la brecha digital en diferentes regiones históricamente excluidas como periferias —sean zonas de frontera, como el Amazonas, o localidades empobrecidas dentro de grandes ciudades como Bogotá—. La creación de centros de IA se acompaña con la gestión de otros proyectos, como Computadores y Tabletas para Educar, que prometen realizar el imaginario anterior en cuanto a infraestructura física. Esta movilización de recursos, que se expresa en número de computadores o metros cuadrados de lotes adquiridos para la transformación tecnológica de las regiones, es percibida como un momento histórico de completitud de una promesa de modernidad:

> "Es la primera vez en la historia que un ministro llega a la región con los representantes de los diferentes operadores de telefonía móvil del país. Nunca antes los responsables le habían dado la cara a la comunidad para generar soluciones en conectividad", destacó el gobernador de Amazonas, Óscar Enrique Sánchez. (MinTIC, 2024d)

> Así cambiaremos la dinámica de la ciudad, pues vamos a producir la ciencia, la tecnología y la Inteligencia Artificial para el mundo desde los barrios populares. (MinTIC, 2024b)

Imaginario: la IA como prótesis del Estado nación

La democratización de la IA (que sea diseñada para todos, que esté al alcance de todos, y que pueda abarcar un sinfín de campos) está en la base de la imaginación de un Estado amplificado. Un Estado que sea capaz de llegar a sus ciudadanos de una manera más eficiente, siempre que las adecuaciones de infraestructura para el aprendizaje y desarrollo de IA "prometen acercar la tecnología a todos los ciudadanos" (Redacción Economía y Negocios, 2024b), esperando que "más personas tengan acceso a herramientas tecnológicas y puedan beneficiarse de los avances que la inteligencia artificial puede brindar en áreas como la educación, la salud y la seguridad" (MinTIC, 2024f), es decir, servicios por excelencia de la imaginación moderna de los Estados, que prometen llegar de forma eficaz a "zonas históricamente desatendidas" (MinTIC, 2024a).

Poner la mirada en algún territorio nacional —"El Gobierno nacional ha puesto su mirada en el Pacífico colombiano" (Ortiz Landecho, 2024)— o la idea misma de que con transformación digital el Estado hace presencia en algunas zonas del país son expresiones que alimentan este imaginario:

> ¿Cuántos años tuvieron que pasar para que el Bolívar profundo tuviera acceso a la conectividad digital? Nosotros consideramos que la transformación de los territorios inicia con conectividad digital y vial, porque con ellas llega la presencia del Estado, llega la presencia institucional, que es la que nos ayuda a tener control territorial, para empezar a hablar de paz y reconciliación. Hoy el Estado hace presencia a través del programa de conectividad digital en más de 800 instituciones educativas que van a tener internet, que van a tener conexión al mundo a través del MinTIC. (MinTIC, 2024g)

Este campo limita la discusión sobre la relación entre IA y Estado en términos del aumento o potencia de la presencia estatal en regiones consideradas olvidadas o periféricas, al valorarlo positivo. Además, lo presenta como una novedad. La concentración de proyectos de infraestructura e innovación tecnocientífica en este tipo de territorios nacionales no es novedad y valdría la pena considerar qué tipo de continuidades y discontinuidades se promueven en la acción anticipatoria de adecuar infraestructura para la enseñanza, la producción y el consumo de IA en las distintas regiones.

Reflexiones finales

Hemos presentado cuatro campos de la imaginación sociotécnica que construyen distintas noticias acerca de la relación entre la IA y el Estado: (1) innovación

vs. regulación; (2) cuarta revolución industrial; (3) primicia, liderazgo y modernización regional; y (4) desarrollo y presencia nacional amplificada. Esta producción confina el debate al uso de metáforas y problemas que son similares y están vinculadas, como aquellos que tiene que ver con la revolución, la novedad, el liderazgo, la innovación, el posicionamiento, el desarrollo, la eficiencia y la productividad, que se conectan con la promesa de que estas herramientas son objetivas y eficientes (Fourcade y Healey, 2024).

Estos distintos campos no se contraponen entre sí; antes bien, cada uno comparte con otros algunos recursos transversales, como la imaginación de una carrera global de suma cero en la que tanto Colombia como sus regiones tienen la oportunidad de posicionarse como líder en la innovación tecnológica y la producción. Esta afinidad entre campos es posible gracias al lenguaje de la innovación, que es central a todos los imaginarios que hemos presentado. Estos equiparan la transformación tecnológica de determinados lugares, la adecuación de centros de IA y las infraestructuras para la explotación de datos con el hacer real y concreto el ideal de un Estado "presente" en todo su territorio. En este sentido, la IA es también imaginada con un potencial de modernidad y de amplificación del Estado, en un país donde la narrativa del Estado ausente ha sido recurrente para describir las relaciones de desigualdad entre centros y periferias en el territorio nacional. La promesa aquí es similar a la de una prótesis para el Estado, una extensión del brazo del leviatán, capaz de llegar de mejor manera y de la más eficiente a sus ciudadanos.

En el pasado, proyectos como los de infraestructura y Computadores y Tabletas para Educar comparten con la IA la promesa de completitud de una modernidad incompleta, un discurso ampliamente difundido y presente en el sur global (De Greiff *et al*., 2020). Estos proyectos han sido sobre todo impulsados por el Estado, pero, a diferencia de estos, los imaginarios en torno a la IA hacen pocas referencias a los materiales necesarios para crear y mantener la infraestructura digital. Para terminar, argumentamos que si bien hay algo de ciclos de *hype* (Sovacool y Hess, 2017; Van der Maarel *et al*., 2023) frente al miedo a perder la nueva ola de desarrollo, comunes a otras tecnologías, estos imaginarios se inscriben en imaginarios nacionales propios de proyectos de modernización tecnológica y de infraestructura.

Referencias

Bakiner, O. (2023). Pluralistic sociotechnical imaginaries in artificial intelligence (AI) law: The case of the European Union's AI Act. *Law, Innovation and Technology*, *15*(2), 558-582. https://doi.org/10.1080/17579961.2023.2245675

Bareis, J. y Katzenbach, C. (2021). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values*, *47*(5), 855-881. https://doi.org/10.1177/01622439211030007

Barreneche, C., Lombana-Bermúdez, A. y Ramos-Martín, J. (2021). Datificación en contextos de corrupción: imaginarios sociotécnicos y prácticas de resistencia frente a sistemas antipobreza en Colombia. *Palabra Clave*, *24*(3). https://doi.org/10.5294/pacla.2021.24.3.4

Bones, H., Ford, S., Hendery, R., Richards, K. y Swist, T. (2021). In the frame: The language of AI. *Philosophy & Technology*, *34*, 23-44.

Brennen, J. (2018). *An industry-led debate: How UK media cover artificial intelligence*. Reuters Institute for the Study of Journalism.

Brennen, J. S., Howard, P. N. y Nielsen, R. K. (2020). What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news. *Journalism*, *23*(1), 22-38. https://doi.org/10.1177/1464884920947535

Campolo, A. y Crawford, K. (2020). Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*, *6*, 1-19. https://doi.org/10.17351/ests2020.277

Chan, J. (2021). The future of AI in policing: Exploring the sociotechnical imaginaries. En *Predictive policing and artificial intelligence* (pp. 41-57). Routledge.

Chacón Orduz, M. (2024, 12 de febrero). Minciencias lanza hoja de ruta de inteligencia artificial en Colombia. *El Tiempo*. https://www.eltiempo.com/vida/ciencia/minciencias-lanza-hoja-de-ruta-de-inteligencia-artificial-en-colombia-854166

Chuan, C.-H., Tsai, W.-H. S. y Cho, S. Y. (2019). Framing artificial intelligence in American newspapers. En *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 339-344). https://doi.org/10.1145/3306618.3314285

de Greiff A., A., Herazo, E. L. y Soto Triana, J. S. (2020). Local, global and fragmented narratives about road construction: An invitation to look beyond our disciplinary space. *The Journal of Transport History*, *41*(1), 6-26. https://doi.org/10.1177/0022526620903018

Díaz Ospino, F. (2024, 20 de junio). Barranquilla tiene el primer Centro de Excelencia de Inteligencia Artificial de Colombia y Latinoamérica. *El Tiempo*. https://www.eltiempo.com/colombia/barranquilla/barranquilla-tiene-el-primer-centro-de-excelencia-de-inteligencia-artificial-de-colombia-y-latinoamerica-3354721

Donk, A., Metag, J., Kohring, M. y Marcinkowski, F. (2011). Framing emerging technologies: Risk perceptions of nanotechnology in the German press. *Science Communication*, *34*(1), 5-29. https://doi.org/10.1177/1075547011417892

Esko, T. y Koulu, R. (2023). Imaginaries of better administration: Renegotiating the relationship between citizens and digital public power. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231164113

Suárez-Estrada, M. y Lehuedé, S. (2022). Towards a Terrestrial Internet: Re-imagining digital networks from the ground up. *Tapuya: Latin American Science, Technology and Society*, *5*(1). https://doi.org/10.1080/25729861.2022.2139913

Fast, E. y Horvitz, E. (2016). Long-term trends in the public perception of artificial intelligence. En *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:6965911

Felt, U. (2015). Sociotechnical imaginaries of "the internet" digital health information and the making of citizen-patients. En *Science and Democracy* (pp. 176-197). Routledge.

Fourcade, M. y Healy, K. (2024). *The ordinal society*. Harvard University Press

García, H. (2023, 16 de mayo). Inteligencia artificial: ¿Perderá su empleo? *El Espectador*. https://www.elespectador.com/opinion/columnistas/hectorgarcia/inteligencia-artificial-perdera-su-empleo/

García Rico, J. C. (2024, 22 de junio). Las oportunidades de Colombia PotencIA Digital: Así puede formarse gratis, emprender en TIC y aprovechar la revolución tecnológica. *El Tiempo*. https://www.eltiempo.com/tecnosfera/novedades-tecnologia/colombia-sera-la-capital-de-la-inteligencia-artificial-en-latinoamerica-mintic-3355302

Guenduez, A. A. y Mettler, T. (2023). Strategically constructed narratives on artificial intelligence: What stories are told in governmental artificial intelligence policies? *Government Information Quarterly*, *40*(1). https://doi.org/10.1016/j.giq.2022.101719

Gutiérrez, C. (2023, 18 de noviembre). Ensayo: ¿Quién le pone el cascabel a la inteligencia artificial? *El Espectador*. https://www.elespectador.com/tecnologia/ensayo-quien-le-pone-el-cascabel-a-la-inteligencia-artificial/

Hansen, S. S. (2022). Public AI imaginaries: How the debate on artificial intelligence was covered in Danish newspapers and magazines 1956-2021. *Nordicom Review*, *43*(1), 56-78.

Hautala, J. y Ahlqvist, T. (2024). Integrating futures imaginaries, expectations and anticipatory practices: Practitioners of artificial intelligence between now and future. *Technology Analysis & Strategic Management*, *36*(9), 2100-2112. https://doi.org/10.1080/09537325.2022.2130041

Jasanoff, S. (2015). Future imperfect: Science, technology, and the imaginations of modernity. En S. Jasanoff y K. Sang-Hyun Kim (Eds.), *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power* (pp. 1-33). University of Chicago Press.

Jasanoff, S. y Kim, S.-H. (2009). Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva*, *47*(2), 119-146. https://doi.org/10.1007/s11024-009-9124-4

Kajava, K. y Sawhney, N. (2023). Language of algorithms: Agency, metaphors, and deliberations in AI discourses. En S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 224-236). Edward Elgar Publishing.

Köstler, L. y Ossewaarde, R. (2022). The making of AI society: AI futures frames in German political and media discourses. *AI & Society*, *37*(1), 249-263. https://doi.org/10.1007/s00146-021-01161-9

Latour, B. (1990). Technology is society made durable. *The Sociological Review*, *38*(S1), 103-131.

Law, H. (2023). Computer vision: AI imaginaries and the Massachusetts Institute of Technology. *AI and Ethics*, *4*, 657-663. https://doi.org/10.1007/s43681-023-00389-z

Llegamos tarde y sin claridad a ver qué hacer con la inteligencia artificial. (2023, 11 de noviembre). *El Espectador*. https://www.elespectador.com/opinion/editorial/llegamos-tarde-y-sin-claridad-a-ver-que-hacer-con-la-ia/

Mager, A. y Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified. *New Media & Society*, *23*(2), 223-236. https://doi.org/10.1177/1461444820929321

Maldonado-Castañeda, O. J. y Arroyave, C. A. (2024). Promesas algorítmicas: Una exploración a los imaginarios sociotécnicos de la inteligencia artificial en el cuidado de la salud. En N. Ángel-Cabo y R. Ureña-Hernández (Eds.), *Derecho, poder y datos. Aproximaciones críticas al derecho y las nuevas tecnologías* (pp. 41-65). Ediciones Uniandes.

Marčetić, H. y Nolin, J. (2023). Utopian and dystopian sociotechnical imaginaries of big data. *Journal of Digital Social Research*, *5*(4), 93-125.

Maussa, A. (2023, 11 de junio). Bondades y oportunidades de la inteligencia artificial. *El Espectador*. https://www.elespectador.com/opinion/lectores/antieditorial/bondades-y-oportunidades-de-la-inteligencia-artificial/

McCloskey, D. N. (1983). The rhetoric of economics. *Journal of Economic Literature*, *21*(2), 481-517.

Méndez, A. L. (2023, 10 de diciembre). Comandante de FF. MM. incorporará la inteligencia artificial en operaciones. *El Tiempo*. https://www.eltiempo.com/justicia/conflicto-y-narcotrafico/inteligencia-artificial-comandante-de-las-ff-mm-la-incorporara-en-operaciones-834128

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, *8*(2), 193-210.

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2023a, 15 de junio). Colombia tendrá laboratorio de inteligencia artificial. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-276507.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2023b, 12 de julio). MinTIC lidera la Mesa Internacional de Inteligencia Artificial para Colombia. *Sala de Prensa*. https://mintic.gov.co/portal/inicio/Sala-de-prensa/Noticias/276849:MinTIC-lidera-la-Mesa-Internacional-de-Inteligencia-Artificial-para-Colombia

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024a, 22 de mayo). Beneficios y retos de la Inteligencia Artificial en la medicina fueron destacados por el Ministro TIC. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382391.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024b, 6 de mayo). "Con las Juntas de Internet estamos potenciando la conectividad en los barrios populares de Cundinamarca": Ministro Lizcano. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382079.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024c, 19 de junio). Conectividad e inteligencia artificial transformarán a Buenaventura gracias al Ministerio TIC. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382911.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024d, 6 de marzo). "La conectividad para Leticia mejora o mejora": Ministro Mauricio Lizcano. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-334437.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024e, 9 de mayo). La creación de facultad de inteligencia artificial y los estudios para un proyecto de fabricación de microprocesadores potencian digitalmente a Manizales. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382147.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024f, 18 de junio). Ministro TIC explicó los avances de Colombia hacia una política nacional de inteligencia artificial con enfoque en gobernanza y ética. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382883.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024g, 14 de marzo). MinTIC prepara proyecto de ley para que Colombia se convierta en un país productor de datos. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-334561.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024h, 4 de abril). Palmira tendrá un Centro PotencIA Digital para la formación en inteligencia artificial. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-337785.html

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2024i, 29 de mayo). "Tenemos que trabajar para democratizar la inteligencia artificial y que no esté en manos de unos pocos países": Ministro Lizcano en Suiza. *Sala de Prensa*. https://www.mintic.gov.co/portal/715/w3-article-382551.html

Montoya Castaño, D. (2024, 8 de septiembre). Formación integral y la inteligencia artificial. *El Espectador*. https://www.elespectador.com/opinion/columnistas/dolly-montoya-castano/formacion-integral-y-la-inteligencia-artificial/

Natale, S. y Ballatore, A. (2017). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence*, *26*(1), 3-18. https://doi.org/10.1177/1354856517715164

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A. y Kleis Nielsen, R. (2024). *Reuters Institute Digital News Report 2024*. https://doi.org/10.60625/risj-vy6n-4v57

Obozintsev, L. (2018). *From Skynet to Siri: An exploration of the nature and effects of media coverage of artificial intelligence*. University of Delaware.

Ojeda, D. (2024, 13 de marzo). Los detalles de la ley que regula el uso de la IA en la Unión Europea. *El Espectador*. https://www.elespectador.com/tecnologia/los-detalles-de-la-ley-que-regula-el-uso-de-la-ia-en-la-union-europea/

Ortiz Landecho, N. T. (2024, 20 de junio). Ministerio TIC abrirá "Centros PotencIA" para el aprendizaje de inteligencia artificial. *El Tiempo*. https://www.eltiempo.com/mas-contenido/ministerio-tic-abrira-centros-potencia-para-el-aprendizaje-de-inteligencia-artificial-3354528

Paltieli, G. (2022). The political imaginary of national AI strategies. *AI & Society*, *37*(4), 1613-1624. https://doi.org/10.1007/s00146-021-01258-1

Pham, B.-C. y Davies, S. R. (2024). What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy. *Critical Policy Studies*, 1-19. https://doi.org/10.1080/19460171.2024.2373786

Rangel, V. (2024, 30 de mayo). Algoritmos y arte: políticas culturales en la era de la inteligencia artificial. *El Espectador*. https://www.elespectador.com/opinion/columnistas/columnista-invitada/algoritmos-y-arte-politicas-culturales-en-la-era-de-la-inteligencia-artificial/

Redacción Economía y Negocios. (2024a, 16 de febrero). El acuerdo con el que Colombia busca pasar de consumir a producir tecnología. *El Espectador*. https://www.elespectador.com/economia/el-acuerdo-con-el-que-colombia-busca-pasar-de-consumir-a-producir-tecnologia/

Redacción Economía y Negocios. (2024b, 19 de junio). MinTIC invirtió en programa para ampliar el acceso a internet en Valle del Cauca. *El Espectador*. https://www.elespectador.com/economia/mintic-invirtio-en-programa-para-ampliar-el-acceso-a-internet/

Redacción Economía y Negocios. (2024c, 4 de marzo). Seis de cada diez empleos en Colombia podrían ser automatizados: Fedesarrollo. *El Espectador*. https://www.elespectador.com/economia/finanzas-personales/seis-de-cada-diez-empleos-en-colombia-podrian-ser-automatizados-fedesarrollo/

Redacción Especiales. (2024, 22 de febrero). Inteligencia artificial, automatización y seguridad: ¿Cómo será la próxima década? *El Espectador*. https://www.elespectador.com/contenido-patrocinado/inteligencia-artificial-automatizacion-y-seguridad-como-sera-la-proxima-decada/

Ricaurte, P., Gómez-Cruz, E. y Siles, I. (2024). Algorithmic governmentality in Latin America: Sociotechnical imaginaries, neocolonial soft power, and authoritarianism. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241229697

Richter, V., Katzenbach, C. y Schäfer, M. S. (2023). Imaginaries of artificial intelligence. En *Handbook of critical studies of artificial intelligence* (pp. 209-223). Edward Elgar Publishing.

Roberge, J., Senneville, M. y Morin, K. (2020). How to translate artificial intelligence? Myths and justifications in public discourse. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720919968

Rueda, M. I. (2023, 10 de diciembre). ¿Qué ventajas y qué peligros trae Bard, la nueva forma de IA de Google? *El Tiempo*. https://www.eltiempo.com/tecnosfera/novedades-tecnologia/que-ventajas-y-que-peligros-trae-bard-la-nueva-forma-de-ia-de-google-834199

Schmid, B., Serlavós, M. y Hirt, L. F. (2022). Community energy initiatives as a space for emerging imaginaries? Experiences from Switzerland. En S. Löbbe, F. Sioshansi y D. Robinson (Eds.), *Energy communities* (pp. 167-181). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-91135-1.00006-7

Sovacool, B. K. y Hess, D. J. (2017). Ordering theories: Typologies and conceptual frameworks for sociotechnical change. *Social Studies of Science*, *47*(5), 703-750.

Suchman, L. (2008). Feminist STS and the Sciences of the Artificial. En *The handbook of science and technology studies* (pp. 139-164). Massachusetts Institute of Technology Press.

Ulnicane, I., Eke, D. O., Knight, W., Ogoh, G. y Stahl, B. C. (2021). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, *46*(1-2), 71-93. https://doi.org/10.1080/03080188.2020.1840220

van der Maarel, S., Verweij, D., Kramer, E.-H. y Molendijk, T. (2023). "This is not what I signed up for": Sociotechnical imaginaries, expectations, and disillusionment in a Dutch military innovation hub. *Science, Technology, & Human Values*, 1-22. https://doi.org/10.1177/01622439231211032

Vrabič Dežman, D. (2024). Promising the future, encoding the past: AI hype and public media imagery. *AI and Ethics*, *4*, 743-756. https://doi.org/10.1007/s43681-024-00474-x

Wenger, A., Jasper, U. y Dunn Cavelty, M. (2020). Governing and probing the future: The politics and science of prevision. En A. Wenger, U. Jasper y M. Dunn Cavelty (Eds.), *The politics and science of prevision: Governing and probing the future* (pp. 3-23). Routledge.

Wyatt, S. (2004). Danger! Metaphors at work in economics, geophysiology, and the internet. *Science, Technology & Human Values*, *29*(2), 242-261. https://doi.org/10.1177/0162243903261947

Wyatt, S. (2021a). Metaphors in critical Internet and digital media studies. *New Media & Society*, *23*(2), 406-416. https://doi.org/10.1177/1461444820929324

Wyatt, S. (2021b). Past and present metaphors of interaction and virtuality. En C. Ernst y J. Schröter (Eds.), *(Re-)imagining new media: Techno-imaginaries around 2000 and the case of "Piazza virtuale" (1992)* (pp. 7-14). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32899-3_2

Züger, T., Kuper, F., Fassbender, J., Katzy-Reinshagen, A. y Kühnlein, I. (2023). Handling the hype: Implications of AI hype for public interest tech projects. *TATuP-Journal for Technology Assessment in Theory and Practice*, *32*(3), 34-40.

SOBRE LOS AUTORES

Diana Agudelo
Profesora asociada del Departamento de Psicología, Universidad de los Andes. Doctora en Psicología Clínica y de la Salud por la Universidad de Granada.

Pablo Andrés Arbeláez
Doctor con honores en Matemáticas Aplicadas por la Universidad París-Dauphine, Francia. Profesor asociado del Departamento de Ingeniería Biomédica, Universidad de los Andes. Director del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA).

Jaime Humberto Borja
Profesor titular del Departamento de Historia, Universidad de los Andes. Doctor en Historia por la Universidad Iberoamericana de Ciudad de México.

Ana María Bustamante Duarte
Profesora asistente de la Facultad de Arquitectura y Diseño, Universidad de los Andes. Doctora en Geoinformática por la Westfälische Wilhelms-Universität Münster, Alemania.

Andrés Calderón
Investigador en ciencia de datos con especialización en Big Data, Minería de Datos y Visualización, enfocado en datos espaciales. Doctor en Ciencias de la Computación por la Universidad de California, Riverside.

Yuly Calderón
Psicóloga por la Universidad de los Andes. Coordinadora de Monitoreo y Evaluación de la ONG Aulas en Paz.

Nicolás Cardozo
Profesor asociado del Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes. Doctor en Ingeniería por la Université Catholique de Louvain, Bélgica, y doctor en Ciencias por la Vrije Universiteit Brussel, Bélgica.

Alejandro Castañeda Molano
Jefe del Centro de Internet Seguro Viguías, Red PaPaz. Magíster en Estudios Interdisciplinarios sobre Desarrollo por la Universidad de los Andes.

Ángela Castillo Aguirre
Doctora en Filosofía por la Universidad de los Andes. Miembro del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA).

Nicolás Díaz
Científico de datos. Magíster en Ingeniería de Sistemas y Computación por la Universidad de los Andes.

Miller Díaz Valderrama
Investigador del Centro de Objetivos de Desarrollo Sostenible para América Latina y el Caribe, Universidad de los Andes.

Ivana Dusparic
Profesora asociada, School of Computer Science and Statistics en Trinity College Dublin. Doctor en Ciencias de la Computación por la Trinity College Dublin.

Luis Felipe Giraldo
Profesor asociado del Departamento de Ingeniería Biomédica, Universidad de los Andes. PhD, Electrical and Computer Engineering, The Ohio State University.

Catalina González Uribe
Investigadora del Centro de Objetivos de Desarrollo Sostenible para América Latina y el Caribe, Universidad de los Andes. Doctora en Epidemiología y Salud Pública por la University College London.

Javier Enrique Guerrero
Investigador del Centro de Objetivos de Desarrollo Sostenible para América Latina y el Caribe, Universidad de los Andes. Doctor en Estudios de Ciencia y Tecnología de la Universidad de Edimburgo.

Juan David Gutiérrez
Profesor asociado de la Escuela de Gobierno, Universidad de los Andes. Director del proyecto Sistemas de Algoritmos Públicos. Doctor en Política Pública por la Universidad de Oxford.

Laura Camila Hernández Gutiérrez
Investigadora y profesional en Psicología por la Universidad Católica de Colombia. Analista de Viguías, Red PaPaz.

Laura Viviana Manrique
Doctora en Historia con énfasis de investigación en Inteligencia Artificial.

Rubén Manrique
Profesor asistente del Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes. Doctor en Ingeniería por la Universidad de los Andes.

Olga Mariño
Profesora asociada del Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes. PhD en Informática, Universidad Joseph Fourier (actual Universidad Grenoble Alpes), Francia.

Sarah Muñoz Cadena
Magíster en Economía de las Políticas Públicas por la Universidad del Rosario. Politóloga bilingüe con estudios complementarios en periodismo y profesional en Gobierno y Asuntos Públicos por la Universidad de los Andes.

Natalia Niño Machado
Investigadora del Centro de Objetivos de Desarrollo Sostenible para América Latina y el Caribe, Universidad de los Andes. Doctora en Estudios de Ciencia y Tecnología por la Universidad de Edimburgo.

Haydemar Núñez
Profesora del Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes. Doctora en Inteligencia Artificial por la Universidad Politécnica de Cataluña, España.

Wilman Osejo
Abogado, especialista en Derecho Penal y Ciencias Forenses. Consultor del proyecto Prevención del Abuso en Línea de Niñas, Niños y Adolescentes en Latinoamérica, Universidad de los Andes.

Andrés Páez
Profesor titular del Departamento de Filosofía y miembro del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA), Universidad de los Andes. Doctor en Filosofía por The City University of New York.

Diego Pajarito Grajales
Investigador asociado del Urban Big Data Centre, University of Glasgow. Doctor en Geoinformática por la Universidad Jaume I, Castellón, España.

Gilbert Paquette
Profesor emérito del Departamento de Ciencia y Tecnología, Université Teluq. PhD en Inteligencia Artificial y Educación por la Université du Maine, Francia.

Leonardo Parra Agudelo
Profesor asociado de la Facultad de Arquitectura y Diseño, Universidad de los Andes. Doctor en Transformación Urbana y Social por el Laboratorio de Informática Urbana de Queensland, University of Technology, Australia.

Carolina Paz
Psicóloga e ingeniera de *software*, con amplia experiencia en investigación del comportamiento humano y la tecnología, con énfasis en violencia sexual. Analista de Viguías, Red PaPaz.

Manuel Portela
Investigador posdoctoral y catedrático del Grupo de Ciencia Web y Computación Social de la Universidad Pompeu Fabra de Barcelona. Doctor en Geoinformática por la Universidad Jaume I, Castellón, España.

Juanita Puentes Mozo
Magíster en Ingeniería Biomédica por la Universidad de los Andes. Investigadora del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA).

Nicanor Quijano
Profesor titular del Departamento de Ingeniería Eléctrica y Electrónica, Universidad de los Andes. PhD en Electrical and Computer Engineering, The Ohio State University.

Viviana Quintero Salgado
Psicóloga especializada en protección de niñas, niños y adolescentes en línea. Becaria Fulbright y Humphrey Fellow en el Washington College of Law de la American University, centrada en derechos humanos y prevención de la trata de seres humanos. Asesora técnica de Aulas en Paz.

Lina María Saldarriaga
Psicóloga, magíster en Psicología por la Universidad de los Andes y Doctora en Psicología del Desarrollo por la Universidad de Concordia, Canadá. Directora de operaciones de la ONG Aulas en Paz.

Rocío Sierra
Profesora asociada del Departamento de Ingeniería Química y de Alimentos, Universidad de los Andes. PhD en Ingeniería Química por la Universidad de Texas A&M.

David Vásquez
Geocientífico, Universidad de los Andes. Magíster en Ingeniería de la Información.

Inteligencia artificial
Teorías, aplicaciones, futuro
fue compuesto en caracteres
Brandon Grotesque y Minion Pro.

Bogotá, octubre del 2025