Referencias

Ahumada, H (2021). Inteligencia Artificial (IA) y Pronósticos Económicos, Publicación ANCE. https://anceargentina.org/

Chisari, O.O. (2021). Inteligencia Artificial e Infraestructura: evaluaciones en Equilibrio General Computado para seis países de América Latina, Publicación ANCE. https://anceargentina.org/

De Pablo, J.C. (2021). Inteligencias, Natural y Artificial, Publicación ANCE. https://anceargentina.org/ Elías, V.J. (2021). Punto de vista de un economista sobre los efectos posibles del arribo y adopción de la inteligencia artificial (IA) en la economía de un país, Publicación ANCE. https://anceargentina.org/ Fanelli, J.M. & R. Albrieu (2021). Crecimiento e inteligencia artificial: los desafíos de vivir entre Detroit y Bombay, Publicación ANCE. https://anceargentina.org/

Gasparini, L. (2021). Inteligencia Artificial, Empleo y Desigualdad, Publicación ANCE. https://anceargentina.org/ Heymann, D. & P. Mira (2021). Aspectos (Macro) Económicos de la Inteligencia Artificial, Publicación ANCE. https://anceargentina.org/

Kulesz, M & F. Navajas (2021). Inteligencia artificial, organización industrial y competencia, Publicación ANCE. https://anceargentina.org/

Montuschi, L. (2021). La Inteligencia Artificial, el Mercado de Trabajo y la Educación, Publicación ANCE. https://anceargentina.org/

30 de Septiembre de 2021

Desmitificando la Inteligencia Artificial

Laura Ación[1], Laura Alonso Alemany[2], Enzo Ferrante[3], Eric Lützow Holm[4], Vanina Martinez[5], Diego H. Milone[3], Ricardo Rodriguez[5], Guillermo Simari[6], Sebastian Uchitel[7]

La computación ha transformado al mundo en sucesivas olas. Las computadoras digitales, las computadoras personales, internet y los dispositivos móviles son ejemplos que la sociedad reconoció, en distintos momentos y con justa razón, como tecnologías disruptivas que estaban por cambiar nuestro mundo para siempre. La Inteligencia Artificial es, sin dudas, el área que hoy está por cambiar sustancialmente nuestro mundo, y hablamos en potencial porque aunque el impacto de la inteligencia artificial hoy es palpable e impactante, aún cuesta imaginar lo que se viene.

1 Instituto de Cálculo, CONICET/UBA.

2 Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba.

3 Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, Universidad Nacional del Litoral, CONICET.

4 Instituto de Cálculo, CONICET/UBA.

5 Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires e Instituto de Ciencias de la Computación, CONICET/UBA.

6 Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur e Instituto de Ciencias e Ingeniería de la Computación, CONICET/UNS.

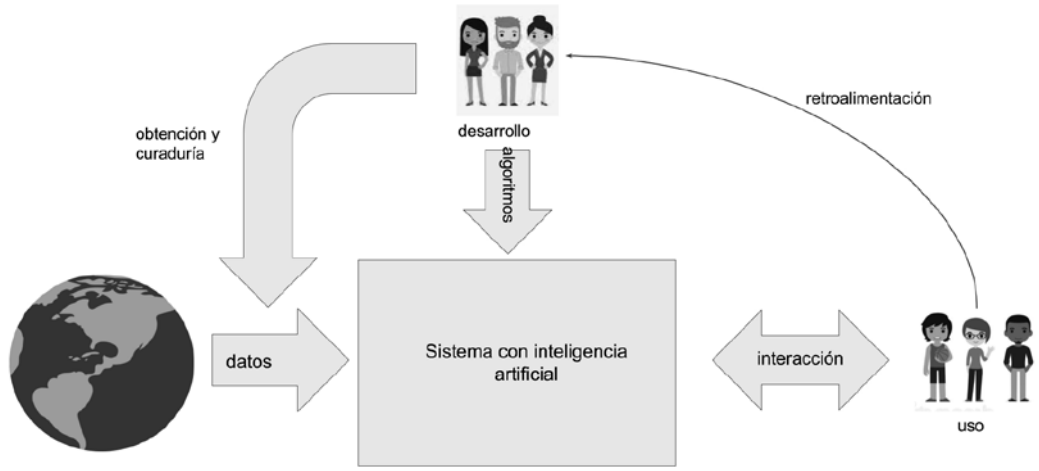
Una definición precisa del término *inteligencia artificial* admite muchos debates de interés, particularmente la referencia a inteligencia. Sin embargo, lo que es claro es que "artificial" remite al comportamiento exhibido por una máquina capaz de realizar cómputo automático. En este sentido, es fundamental, y lo hacemos en la **primera** sección, explicar cuál es la paleta de técnicas, fuertemente enraizadas en las ciencias de la computación, que hoy existen para dotar a una máquina de comportamiento que podría llamarse "inteligente", particularizando no sólo cómo funcionan estas técnicas sino sus fortalezas y debilidades.

Una familia de técnicas que impulsa esta revolución de la Inteligencia Artificial está orientada a que un sistema aprenda a través de ejemplos, también llamados *datos de entrenamiento*, previo a su puesta en funcionamiento y/o incorporando la experiencia adquirida mientras está en uso. La forma en que se diseña este proceso de aprendizaje y los datos con que se alimenta son claves en el futuro del comportamiento del sistema. Claramente, quienes son responsables de hacer el diseño son, somos, seres humanos. Consecuentemente, y aunque pueda resultar evidente para algunos, aún no está incorporado al sentido común que el comportamiento de los sistemas que incorporan elementos de inteligencia artificial pueda tener fuertes sesgos que repitan, o incluso profundicen, los errores, prejuicios e injusticias que cometen los mismos seres humanos. Abordamos esta temática en la **segunda** sección.

Desde el reconocimiento que finalmente en el corazón de la Inteligencia Artificial están los seres humanos (ver el esquema más abajo), las consideraciones éticas son centrales. Abordamos en la **tercera** sección cómo la ética impacta en el diseño de sistemas inteligentes y cómo puede pensarse el diseño de estos sistemas para que tengan consideraciones éticas desde su construcción misma.

Finalmente, abordamos uno de los desafíos técnicos más acuciantes que la Inteligencia Artificial tiene por delante. La necesidad de que los resultados computados por un sistema de inteligencia artificial puedan ser explicados automáticamente de manera que personas, tanto expertas en el dominio de aplicación como aquellas que no lo son, puedan entender cómo el sistema llegó a una decisión particular.

Cerramos el capítulo con un llamado a fortalecer la investigación en Ciencias de la Computación en general y en Inteligencia Artificial en particular en la Argentina.



Esquema básico de la interacción entre las personas y los sistemas con inteligencia artificial. Las personas están presentes en todas las fases de la producción del sistema: desde el desarrollo hasta su uso final. Las personas encargadas del desarrollo discuten el uso que se le dará al sistema, introducen los algoritmos con los que este operará y obtienen y preparan los datos para que el sistema funcione correctamente. Los datos pueden provenir de diversas fuentes, pero son personas las que deben determinar su correcta utilización. El sistema toma los datos, los procesa en base a los algoritmos y genera salidas que son interpretadas por personas en el contexto de uso. Estas personas, a su vez, pueden incidir en el sistema y repercutir en su desarrollo o en el de otros sistemas.

No es automágico: una breve introducción

En octubre de 1950, Alan M. Turing publicó un ensayo titulado "Computing Machinery and Intelligence", Mind, LIX (236): 433–460, en el que discutía la posibilidad de que una máquina pudiera pensar en el sentido humano que usualmente se asocia con ese término. Al hablar de maquinaria capaz de computar, Turing se refería a los sistemas computacionales que en ese entonces se comenzaban a desarrollar; luego de analizar las posibles objeciones a lograr ese objetivo, concluía con una respuesta positiva[1]. Es interesante notar que la maquinaria física (hardware) nunca fue el centro del debate salvo por su capacidad de ejecutar la programa-

1 Una buena introducción a la historia y desarrollo de la IA puede encontrarse en: *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going.* Michael Wooldridge Flatiron Books, 2021.

ción que produciría el comportamiento inteligente. Más tarde, en 1955, se realizó una propuesta liderada por John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon para realizar al año siguiente un taller de trabajo en Dartmouth College, New Hampshire, EE.UU.; en esa propuesta apareció por primera vez el término *Inteligencia Artificial* que se debía interpretar como inteligencia realizada en un sistema computacional. Años después, John McCarthy expresó que hubiera sido mejor emplear el término *Inteligencia Computacional* porque hubiera sido más preciso, pero el término original perduró y es el que usamos actualmente.

Resulta imposible dar una definición precisa de la disciplina Inteligencia Artificial (IA), dado que tampoco existe una definición clara de *inteligencia*. El término es usado para describir un área de trabajo amplísima a la que han contribuido todas las ramas del saber humano. Por eso, es interesante explorar el concepto a través de algunas preguntas cuyas respuestas pueden aclarar su denotación. Estas preguntas podrían ser: ¿qué es la IA?, ¿qué se espera de la IA? y ¿cómo sería posible concretarla?

La primera pregunta es por supuesto *ontológica* y ha tenido respuestas variadas a lo largo de las décadas que pasaron desde la introducción del término. Como se ha dicho, la mayor dificultad es que no existe una comprensión precisa de lo que comúnmente se considera inteligencia. Reconociendo esto, Turing propuso en el trabajo mencionado un test que podría decidir si un sistema computacional exhibe un comportamiento que fuera aceptado como inteligente. El test representa una forma de decidir por comparación con un ejemplo, que en este caso es el ser humano. Esencialmente, el test involucra por un lado dos participantes: un sistema y un ser humano; además, se agrega un interrogador que realiza preguntas a los dos participantes sin saber cuál es el ser humano, y su tarea es decidir a través del análisis de las respuestas a sus preguntas cual es el ser humano. Si no es posible discriminar entre los dos interrogados, se reconocerá que el sistema es inteligente. Desde 2006 se realiza una competición por el Premio Loebner en el escenario descrito por Turing con resultados no definitivos pero que van aproximando la solución.

A lo largo de los años, se han ofrecido diversas descripciones de la IA; por ejemplo, si un sistema exhibe un comportamiento que se aceptaría como inteligente en un ser humano, entonces el sistema será considerado inteligente. Es claro que estas descripciones siguen el modelo del test de Turing que compara el sistema con el ser humano para decidir. Una parte importante del problema es que tenemos un solo ejemplo de inteligencia para estudiar el concepto haciendo que sea complicado separar lo superficial de lo esencial. Una distinción interesante y útil es la separación en dos tipos de inteligencia: la Inteligencia Artificial General (IAG) que sería

similar a la humana, y la Inteligencia Artificial Enfocada que sería especializada en una tarea. Si bien existen otras descripciones ofrecidas en la literatura, la gran mayoría se divide en dos posibles perspectivas. Una incluye la ofrecida por Turing: se podría decir que un sistema computacional es inteligente si es percibido como *pensando* en forma similar a la humana. La otra se limita al comportamiento sin poner limitaciones internas; aquí tendríamos inteligencia computacional si esta se *comporta* como un ser humano y ese comportamiento es considerado inteligente.

Al considerar la segunda pregunta acerca de lo que puede producir la IA, estamos apuntando a una *descripción funcional*. Crear una IAG requiere producir un sistema que pueda comportarse de manera inteligente en escenarios diversos, como lo hace un ser humano; i.e., un ser humano puede conducir un auto, jugar al ajedrez y conversar inteligentemente, probablemente alternando sucesivamente entre estas actividades y en algunos casos de manera simultánea. Por otra parte, la Inteligencia Artificial Enfocada intenta producir sistemas capaces de desarrollar una actividad inteligente particular; i.e., jugar al ajedrez, mantener una conversación o conducir un auto; estos intentos han tenido éxito variado en los últimos años y son el centro del entusiasmo actual por la IA.

La última pregunta nos lleva al análisis de *cómo* construir un sistema inteligente. Para alcanzar el objetivo de crear un sistema computacional que exhiba un comportamiento inteligente, se han desarrollado herramientas variadas siguiendo las líneas que se han trazado desde el estudio de la inteligencia humana en las diversas áreas disciplinares asociadas.

Mucho antes de que el psicólogo y economista Daniel Kahneman (premio Nobel de Economía 2002) publicara el resultado de sus investigaciones en el libro titulado *Thinking Fast and Slow*, (2011) Penguin Books, los trabajos de investigación en IA se dividieron en dos formas generales de pensar el problema de emular la inteligencia. Estos caminos coinciden de alguna manera con los resultados de Kahneman, quien postula la existencia de dos componentes en la estructura mental humana: el *Sistema* 1, que está siempre activo, realiza un procesamiento rápido, reflejo, automático, inconsciente (no observable) y en general estereotípico; y el *Sistema* 2 cuyo mecanismo, que solo se activa cuando se lo requiere, es lento, lógico y consciente (observable). Es cierto que existen formas intermedias en las que el pensamiento actúa utilizando ambos sistemas en forma concurrente y colaborativa, y resulta dificultoso clasificarlas en una categoría precisa. El Sistema 2 usaría como insumos las contribuciones del Sistema 1, e.g., el reconocimiento de patrones del Sistema 1 produce un símbolo como subrogante, o substituto del patrón observado y el Sistema 2 usa este símbolo para razonar sobre la situación, como sucede al

reconocer un rostro en una multitud: inmediatamente nuestros procesos mentales asociados con el Sistema 1 recurren al nombre de la persona que fue reconocida para continuar con su trabajo.

Al elaborar pensamientos se recurre a conexiones conocidas entre los símbolos considerando posibles conclusiones. Los posibles vínculos (o reglas) de encadenamiento representan relaciones: algunas son generales, e.g., conociendo la regla general que todos los mamíferos tienen pelo puedo concluir que los seres humanos tienen pelo; otras son personales, e.g., conociendo que los gatos son mamíferos puedo obtener que un gato particular, Garfield, tiene pelo por ser una característica común a todos los mamíferos. El Sistema 1 luego de muchas interacciones en situaciones parecidas reconoce la conexión y la conserva para ser usada cuando sea necesario, aunque posiblemente deba modificar el resultado cuando disponga de más experiencias. El Sistema 1 "aprende" patrones y los usa de forma refleja al reconocerlos configurando una herramienta indispensable para actuar efectivamente. Pero el acto deliberativo consciente de analizar una situación y obtener conclusiones corresponde a lo que hemos descrito como Sistema 2, y si bien es cierto que algunas de sus "habilidades" son aprendidas a través del Sistema 1, las más elaboradas han desarrollado progresivamente a partir de otras destrezas que son conocidas desde hace mucho tiempo como parte de la mente racional utilizando distintas formas de la Lógica. También es cierto que algunos procesos comienzan siendo manejados por el Sistema 2 y luego se transforman en parte del Sistema 1, como sucede por ejemplo con el aprendizaje de las operaciones aritméticas elementales a medida que adquirimos experiencias.

Como decíamos, las investigaciones más importantes en IA se han dividido en dos grupos. Por un lado, se ha buscado comprender la forma como nuestra mente "delibera" a partir de lo que "conoce"; por otro lado, se ha trabajado en producir sistemas orientados al "reconocimiento de patrones" a partir de "observaciones".

En la primera, conocida como IA simbólica, el término "simbólico" se refiere a un sistema de representación en el que los constituyentes atómicos de las representaciones son, a su vez, representaciones. Tal sistema de representación tiene asociadas una sintaxis y una semántica. Un ejemplo de sistema simbólico es una teoría lógica interpretada. En este área se caracteriza la deliberación como un proceso de razonamiento que puede ser descrito de distintas formas representando el conocimiento en diferentes maneras utilizando algún tipo de formalismo lógico que posiblemente considere diferentes aspectos de la incertidumbre asociada con el conocimiento.

En la segunda, caracterizada como IA subsimbólica o presimbólica, una representación "subsimbólica" está compuesta por entidades que no son a su vez representaciones. Ejemplos de este tipo de representación son: píxeles, imágenes, sonidos, señales. Asimismo, las unidades subsimbólicas en las redes neuronales pueden considerarse casos particulares de esta categoría. Aquí se analizan conjuntos de observaciones (datos) buscando encontrar los patrones que permitan obtener conclusiones; los patrones generales permiten crear estructuras que reaccionan a una situación particular de manera análoga a la forma en que se reaccionó en las situaciones en que se obtuvieron las observaciones.

En el área referida como IA simbólica, conocida como Representación de Conocimiento y Razonamiento (en inglés *Knowledge Representation and Reasoning*), se busca encontrar formas de representar información acerca del entorno en el que un agente inteligente autónomo debe desarrollar su actividad resolviendo tareas complejas por sí mismo; se espera que el comportamiento del agente responda a la información con la que cuenta, pudiendo este comportamiento ser explicado en función de esa información. Las experiencias obtenidas en psicología experimental, economía, lógica, matemática y filosofía contribuyen a muchos de los intentos de modelar inteligencia por esta vía, donde la norma es el uso de formalismos lógicos. En inglés se usa la expresión *Knowledge-driven* AI haciendo énfasis en el uso del conocimiento del dominio que se utiliza para implementar los sistemas basados en conocimiento. La complejidad computacional, esto es, la forma en la que crece el tiempo necesario para computar un algoritmo y la manera en la que se expanden los requerimientos de espacio para almacenar datos, representa una gran dificultad en la creación de sistemas basados en conocimiento. Estos costos son independientes de la capacidad de cómputo de un equipo particular en un momento específico y las funciones que describen estos costos tienen factores de crecimiento que no pueden ser ignorados y limitan el "tamaño" de los problemas a resolver. También resulta problemática la elicitación del conocimiento necesario para la construcción de las bases de conocimiento que son requeridas en estas implementaciones, así como también su organización y mantenimiento a lo largo de su vida útil. Por otro lado, dado que la infraestructura cognitiva del sistema es explícita y está disponible en todo momento, las respuestas que ofrece son explicables y analizables. También se facilita el aprovechamiento de una base de conocimiento en diferentes aplicaciones y la construcción de sistemas que usan múltiples bases de conocimiento provenientes de fuentes variadas.

En el área que se mencionó como IA subsimbólica, cuyo principal representante es el aprendizaje de máquina o aprendizaje automático (*Machine Learning*), se uti-

lizan métodos de análisis de conjuntos de datos que automatizan la construcción de modelos basados en esos datos que son conocidos como "datos de entrenamiento". Estos modelos representan patrones inferidos a partir de los datos, y así ayudan al sistema a actuar de manera inteligente por medio de diversos mecanismos y beneficiándose de las decisiones tomadas a partir de los patrones usados para crear los modelos. Es importante destacar que no existe una programación explícita para la obtención de las respuestas. En inglés se utiliza el término *Data-driven AI*, que remarca el uso de los datos del dominio a partir de los que se implementa el sistema. Los algoritmos de aprendizaje crean sistemas cuya complejidad computacional es mucho menor y su respuesta es en general inmediata, lo que representa una ventaja enorme. Esto se logra realizando el aprendizaje de forma previa (*offline*) al uso del sistema. Sin embargo, existen varias dificultades importantes con estos sistemas. La más importante, y que se propaga de varias formas en el análisis, es su fuerte dependencia del conjunto de datos usados en el entrenamiento. En algunos casos, pequeños cambios en el conjunto llevan a cambios significativos en las respuestas. También es importante reconocer la existencia de "inclinaciones" o "sesgos" (*biases*) escondidos en esos conjuntos, lo que lleva a respuestas que son disímiles a preguntas donde las respuestas deberían ser "cercanas" por razones que no pueden analizarse dada la opacidad de los sistemas que luego se despliegan. La incapacidad de explicar las razones sobre las que se fundamentan las respuestas son complicaciones adicionales que son un foco importante de investigación actualmente. Otro detalle esencial es que, si bien muchas de estas aproximaciones son *bio-inspiradas,* generalmente no se conciben con la finalidad de ser realmente un modelo biológico.

Esta división en áreas de la IA es imprecisa y existen sistemas que operan con arquitecturas mixtas variadas. Dadas las limitaciones de espacio, no sería posible en este texto hacer una descripción más detallada de las dos áreas que hemos descrito. La tendencia a combinar ambos abordajes en sistemas que aprovechan las ventajas de cada uno y disminuyen los problemas está tomando impulso en años recientes. Estas decisiones de diseño reflejan de manera bastante directa la forma en que se comprenden hoy las estructuras de la inteligencia humana. La investigación en arquitecturas que combinan diversos elementos provenientes tanto del área simbólica como de la subsimbólica ha dado lugar a sistemas que demuestran que la complejidad de la inteligencia puede ser comprendida y realizada mejor a través de componentes diversos que se integran para mejorar el desempeño de estos sistemas.

Finalmente, es interesante mencionar que existen diversos problemas éticos y sociales asociados con la puesta en servicio de los sistemas de inteligencia artificial, muchos de ellos imposibles de resolver. Comenzando por la pregunta acerca de la existencia de IAG similar a la humana y cómo sería nuestro comportamiento en relación a ella. Si es similar a la inteligencia humana, ¿el sistema tendría consciencia de sí mismo? Si fuera así deberíamos reconocer esta característica y tratarlos como si fueran humanos rechazando su uso como una especie esclavizada. Estas preguntas de índole filosófica también llevan a explorar nuestra condición humana tratando de entender qué nos hace especiales, objetivo que está claramente fuera del alcance de este capítulo. Más adelante en este texto se analizarán otros aspectos relacionados con la implementación de sistemas inteligentes que afectan a nuestra sociedad de maneras algunas veces sorprendentes[2].

¿Sumar no discrimina, pero dividir si? Sesgos en inteligencia artificial

En el año 2015, el periódico estadounidense The New York Times publicó un artículo titulado "Google Photos Mistakenly Labels Black People 'Gorillas'"[3]. El artículo hacía referencia a un desarrollador de software afrodescendiente que había denunciado en las redes sociales cómo la conocida aplicación de Google para la gestión de imágenes había asignado a él y sus amigos la etiqueta de 'Gorilla'. Pero no es necesario remontarnos hasta 2015 para encontrar un ejemplo como este. Recientemente, en Septiembre de 2021, el mismo diario se hizo eco[4] de la denuncia de un grupo de personas en Facebook, a quienes luego de mirar un video donde aparecían hombres de piel oscura, la plataforma les preguntó si les gustaría "seguir viendo videos sobre Primates", lo que hizo que la compañía investigara y deshabilitara la función del sistema de IA que generaba el mensaje. Estos casos, en donde un sistema basado en IA comete errores sistemáticamente en detrimento de una subpoblación en particular, ilustran claramente el concepto de *sesgo algorítmico*. En este apartado intentaremos indagar sobre dicho término, sus implicancias en distintos contextos de aplicación y algunas de las razones detrás de la existencia de estos sesgos, considerando a las personas que desarrollan los sistemas, los datos que usualmente se utilizan para entrenarlos y los modelos de IA que se usan actualmente.

2 Como referencia general podemos mencionar el siguiente texto utilizado como referencia principal en la mayoría de los cursos de IA: Artificial Intelligence: A Modern Approach (4th Ed.), Stuart Russell, Peter Norvig.Pearson, 2020.

3 https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

4 https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

Si bien no es un concepto nuevo, los estudios sobre *sesgo algorítmico* han tomado gran importancia durante los últimos años dada la masiva adopción de los sistemas de IA en la toma de decisiones, particularmente de aquellos basados en *aprendizaje automático*. Desde la asignación de puestos laborales hasta el diagnóstico médico de patologías por medio de imágenes, pasando por la traducción de textos y el otorgamiento de créditos bancarios, en todas estas áreas de aplicación se están comenzando a considerar predicciones generadas por los sistemas de aprendizaje automático para la toma de decisiones. Comencemos entonces por comprender cómo funcionan dichos sistemas y cómo se lleva a cabo su desarrollo, para identificar cuáles son los componentes y etapas que pueden dar origen a dichos sesgos.

Tal como se discutió brevemente en la introducción de este capítulo, uno de los elementos clave en cualquier sistema de aprendizaje automático son los *datos* a partir de los cuales el sistema *aprende* a llevar a cabo una tarea en particular. Dentro de lo que se conoce como el paradigma de *aprendizaje supervisado*, uno de los más utilizados actualmente, cada muestra de nuestra base de datos está acompañada por una *etiqueta*, que indica la salida 'correcta' para esa muestra, dado el problema que queremos resolver. Si queremos entrenar un modelo para clasificar imágenes en función de su contenido y distinguir, por ejemplo, perros de gatos, cada imagen deberá contar con la etiqueta 'perro' o 'gato', previamente asignada de forma manual (es decir, por personas). Durante el proceso de entrenamiento, la idea es ajustar el modelo para que encuentre los patrones en dicha imagen que permiten distinguir una categoría de la otra. Esos patrones están asociados en general a características de la imagen como el color, la textura o las formas. La idea es entonces alimentar el modelo con imágenes e ir ajustando sus parámetros para que la salida se comporte como lo indican las etiquetas; en nuestro caso, que sea capaz de distinguir entre perros y gatos. Claramente, tanto para este aprendizaje como para la posterior validación del modelo, es clave que los datos de entrenamiento sean los adecuados. En la actualidad, generalmente no es un problema el volumen total de datos disponibles, pero es muy común que estos grandes volúmenes de datos estén contaminados (por ejemplo, con imágenes mal capturadas o que no correspondan ni a un perro ni a un gato). También es frecuente que las etiquetas de referencia, que uno supondría a priori correctas, hayan sido asignadas con errores, por ejemplo, indicando cierta proporción de imágenes de gatos con la etiqueta 'perro' y viceversa. A estas contaminaciones más triviales (aunque no por eso fáciles de detectar en grandes volúmenes de datos), se les suman otros problemas que afectan fuertemente el entrenamiento, como el desbalance en las clases de salida (que haya muchos más gatos que perros), o el desbalance en algún otro atributo oculto (y sensible) de los datos, como vamos a analizar a continuación.

Dado que el modelo es entrenado a partir de una base de datos, se genera de alguna manera una dependencia entre aquello que el modelo puede reconocer y los datos observados durante el proceso de aprendizaje. Si bien dichos modelos nunca son entrenados para realizar predicciones a partir de datos que ya han sido observados (¿Qué sentido tendría, si esa misma base de datos de entrenamiento ya contiene las salidas correctas?), la hipótesis subyacente es que los datos nuevos, sobre los que se realizarán predicciones (también conocidos como 'datos de prueba'), seguirán de alguna manera patrones similares a los observados. Pensemos el siguiente ejemplo: nuestro clasificador es entrenado con una base de datos que solo posee gatos negros y perros blancos. Un clasificador ideal para este problema se podría obtener ignorando todas las características del animal, excepto su color. Si es negro, asigna la categoría 'gato'. Si es blanco, asigna 'perro'. Ahora imaginemos que nuestro conjunto de prueba incluye gatos blancos. Es claro que el sistema fallará sistemáticamente en la clasificación de gatos blancos, presentando de esta forma un rendimiento inferior en dicha subpoblación, y por lo tanto resultando en un sistema sesgado.

A partir del ejemplo anterior, se torna evidente que las bases de datos pueden ser una fuente de sesgo algorítmico. Si la base de datos con la que contamos no es lo suficientemente diversa como para representar bien a la población objetivo, entonces es probable que el clasificador resultante presente un rendimiento dispar en la población subrepresentada. Pero los clasificadores de imágenes no son el único ejemplo. Tomemos por caso los sistemas de traducción basados en IA, que permiten traducir automáticamente de un idioma a otro. Abundan ejemplos en la literatura[5] donde se ponen en evidencia casos de sesgo de género en las traducciones: al convertir palabras de un idioma con pronombres neutros como el húngaro a otro como el inglés, el sistema automáticamente perpetuaba estereotipos de género asignando el género femenino a términos como enfermero/a, panadero/a y organizador/a de bodas, pero traduciendo al masculino las palabras 'médico/a', 'científico/a' o 'ingeniero/a'. Ciertamente este comportamiento está relacionado con la frecuencia de aparición de dichos términos en los textos utilizados para entrenar al traductor. Pero por más que dicha frecuencia sea una medición objetiva de las asimetrías propias de una sociedad como la actual, donde lamentablemente aún existen roles de género asociados a distintas actividades, un modelo sesgado por los datos de entrenamiento sólo contribuye a amplificarlas, dado que continúa incorporándolas en futuras traducciones.

5 Prates, Marcelo OR, Pedro H. Avelar, and Luís C. Lamb. "Assessing gender bias in machine translation: a case study with google translate." *Neural Computing and Applications* 32.10 (2020): 6363-6381.

El segundo componente clave en los sistemas basados en IA es el modelo en sí mismo. Como se dijo anteriormente, en la actualidad prácticamente todos estos modelos están basados en aprendizaje automático, y particularmente en los últimos años en lo que se denominó *aprendizaje profundo* o también *redes neuronales profundas*. Para entender los fundamentos de esta reciente revolución en la IA hay que entender primero qué es una neurona artificial, y luego simplemente conectar la suficiente cantidad de neuronas para que, gracias a los grandes volúmenes de datos disponibles y los recursos de cómputo actuales, podamos entrenar una red neuronal tan 'profunda' como requiera el problema a resolver.

La neurona artificial se inspiró inicialmente en la forma en que las neuronas naturales procesan la información que reciben y envían un pulso de salida a través de su axón. Hoy por hoy, ya sin pretender ser un modelo de la neurona biológica, una neurona artificial es básicamente una unidad elemental de cómputo, que de forma similar a la neurona natural: recibe entradas, las procesa, y da una salida. Tanto las entradas como las salidas son números, y por lo tanto el procesamiento que realiza la neurona es de tipo numérico. Las entradas podrían ser, por ejemplo, la temperatura de una persona (37.5 o 38.3, en ºC) y si tiene tos (1 si tiene, 0 si no tiene). La salida podría ser 1 para indicar que la persona tiene síntomas compatibles con determinada enfermedad, o 0 en caso de que no lo sean. Y el procesamiento que hace la neurona es muy simple: sólo multiplica y suma. Básicamente multiplica cada entrada por un *peso sináptico* (un número), suma las entradas pesadas, y si esa suma supera cierto umbral de referencia (otro número más), la neurona da 1 como salida. Por ejemplo, si la persona tiene 38 ºC con tos, y los pesos sinápticos fueran 0.5 y 50, la cuenta sería 38x0.5 + 1x50 = 69. Si suponemos un umbral de referencia es 60, la salida de la neurona será 1. Es decir, la neurona estaría decidiendo que los síntomas son compatibles con la enfermedad. En caso contrario, si no se supera el umbral, la salida será 0 y la neurona estaría decidiendo que los síntomas no son compatibles con la enfermedad a diagnosticar. Vale aclarar que estos valores para los pesos sinápticos y el umbral de salida no son definidos manualmente como hicimos en este ejemplo, sino que cada neurona los tiene que aprender a partir de los datos de entrenamiento.

Es claro que una sola neurona no va a poder resolver problemas de cierta complejidad, como la detección de patologías en imágenes, donde cada píxel es una entrada (un simple número que indica su intensidad) y sabemos que hay millones de píxeles en una imagen capturada por un celular. Entonces el paso siguiente consiste en *conectar* muchas neuronas y así formar *redes* neuronales, que con el tiempo se constituyeron como las principales representantes de lo que se conoce en IA como modelos *conexionistas*. Pero ¿qué es conectar neuronas? Bueno, simplemente hacer que el número 0 o 1 que sale de una neurona sea el número que entra en la otra. De esa forma, al entrar ese 0 o 1 se multiplicará por el correspondiente peso sináptico de la neurona siguiente, se sumará con las otras entradas, y determinará así la salida de esa segunda neurona que ha sido conectada. Sin entrar en los detalles técnicos acerca de todas las formas en que podríamos interconectar a las neuronas, podemos ver entonces que estos modelos no dejan de ser simples operaciones como sumas y multiplicaciones que, al componerlas, permiten realizar tareas más complejas. Como se puede ver, los sistemas basados en IA no son creaciones ajenas a quienes las crean, que pueden tomar decisiones arbitrarias de forma independiente más allá de cómo fueron diseñadas y entrenadas. A fin de cuentas, quienes discriminamos somos las personas, ya sea en forma directa o (potenciados) con los artefactos tecnológicos que diseñamos.

En el proceso de creación de estos sistemas basados en IA, y particularmente en aprendizaje automático, además de los ya mencionados datos de entrenamiento, existen varios aspectos asociados al modelo que pueden dar como resultado un sistema de IA sesgado. Más precisamente, se trata de aspectos que si no son considerados adecuadamente en el proceso de diseño, seguramente van a generar modelos sesgados, e incluso no seremos capaces de detectar tales sesgos en el sistema. En este proceso de diseño y creación se puede identificar: i) la arquitectura, que incluye particularidades acerca de cómo computan las neuronas, cómo se conectan entre ellas y otras unidades de cómputo que pueden incluirse en el modelo; ii) el aprendizaje, que determina la forma en que se adaptan los parámetros del modelo (pesos sinápticos en el caso de las redes neuronales) para que resuelvan correctamente el problema que tienen que resolver; y iii) la selección del modelo definitivo y su puesta en funcionamiento, lo que involucra principalmente a los esquemas de validación y las medidas de desempeño adecuadas. En estos tres niveles se desarrolla actualmente una intensa tarea de investigación, proveyendo nuevas arquitecturas y algoritmos de entrenamiento robustos a los sesgos por desbalances (explícitos o implícitos) de variables sensibles en los datos, así como también metodologías que permitan validar correctamente los modelos y medir los sesgos resultantes, más allá de las tradicionales métricas de clasificación y calibración.

Cuando estos sistemas basados en IA comienzan a ser desplegados en ámbitos como la justicia, la salud o la selección de personal para puestos laborales, resulta simple imaginar las consecuencias inmediatas de dichos sesgos, especialmente cuando las asimetrías y desigualdades de nuestra propia sociedad se cuelan por medio de los datos y las decisiones de diseño (muchas veces inconscientes) de

aquellas personas que llevan adelante estos desarrollos. Los datos son entonces importantes, y construir bases de datos diversas que representen al conjunto de la población puede ser una estrategia de mitigación de sesgos. Pero como vimos, no son la única causa. A fin de cuentas, la decisión de qué incluir o excluir en una base de datos es una decisión tomada por personas. Estas personas son quienes llevan adelante no sólo la construcción y curado de las bases de datos, sino que implementan y supervisan el proceso de entrenamiento de los modelos, eligen las tareas a resolver, ponen los sistemas con IA en funcionamiento y monitorean su rendimiento a lo largo del tiempo. En todas estas etapas que constituyen el ciclo de vida de desarrollo de un sistema de IA, son las personas quienes toman las decisiones, y muchas de esas decisiones pueden generar sesgos algorítmicos, o permitir que sean detectados a tiempo y mitigar sus impactos. Como veremos en la próxima sección, medir si un sistema de IA es justo o no, no es tarea simple. Las definiciones formales de justicia algorítmica tienden a ser mutuamente excluyentes, en el sentido de que no todas pueden ser satisfechas al mismo tiempo, y por tanto nuevamente las decisiones humanas, sobre qué criterios de justicia han de ser priorizados, se vuelven relevantes. Por esa razón, contar con equipos diversos, con integrantes que manifiesten distintos puntos de vista, que puedan auditar tanto los datos como los modelos, antes, durante y después del proceso de desarrollo, constituye una herramienta fundamental en la construcción de sistemas de IA más justos. Lamentablemente, todavía estamos lejos de esta realidad. Tanto a nivel nacional[6] como internacional[7], las cifras reflejan que la composición de la comunidad CTIM (ciencia, tecnología, ingeniería y matemática) en general, y la de informática en particular[8], cuenta con una terrible subrepresentación de mujeres y diversidades en su composición.

6 Ana Inés Basco, Cecilia Lavena y Chicas en Tecnología: «Un potencial con barreras. La participación de las mujeres en el área de Ciencia y Tecnología en Argentina», Nota Técnica No IDB-TN-01644, bid, 2019.

7 Sarah Myers West, Meredith Whittaker y Kate Crawford: «Discriminating Systems: Gender, Race and Power in AI», AI Now Institute, 4/2019

8 Informe de la Red Disciplinar de Informática y Comunicaciones, CONICET, 2019: https://proyectosinv.conicet.gov.ar/redes-disciplinares/

El ojo del amo...: ética y responsabilidad

A medida que ampliamos las funciones que involucran a los sistemas con inteligencia artificial (SIA) en asuntos prácticos, como los vehículos autónomos, los sistemas de diagnóstico y de apoyo para toma de decisiones, las consideraciones éticas surgen inevitablemente. La ética, como rama de la filosofía, pretende analizar y responder la siguiente pregunta: ¿Qué debo hacer? Permite delimitar lo "correcto" de lo que no lo es, lo "bueno" de lo "malo", de acuerdo a un proceso de reflexión que determina las acciones de un agente, o un grupo de ellos, en base a un conjunto de valores, principios y propósitos. Uno de los principales desafíos en relación a la ética es determinar qué valores deben considerarse y cómo deben priorizarse en caso que haya interacción.

En la sección anterior vimos que los sesgos son la principal fuente de comportamientos discriminatorios, al violentar los derechos humanos y la privacidad. Por tanto los sistemas que tomen decisiones sesgadas son éticamente cuestionables. En esta sección nos ocuparemos de describir distintos aspectos sobre cómo evitar comportamientos antiéticos y de qué manera garantizar la responsabilidad en caso que a pesar de nuestro esfuerzo no puedan ser evitados. Dado que los documentos de las otras academias abordarán los fundamentos empíricos y epistémicos de la ética, aquí sólo nos referiremos a ella como un concepto sellado y para explorarlo sugerimos la lectura de esos documentos. Lo que sí consideraremos aquí es que detrás del desarrollo de cualquier SIA están los seres humanos y por lo tanto somos nosotros los responsables últimos de "asegurar" el resguardo de los aspectos éticos.

En ese sentido consideramos que los sistemas informáticos no son herramientas puras y neutras, sino productos de su contexto socio-técnico, y deben ser considerados como tales. Asumiremos que los SIA siempre pueden entenderse en un nivel superior, intensionalmente en términos de sus diseños y objetivos operativos, y extensivamente en términos de sus entradas y salidas. Concluiremos que a menos que un SIA sea auditado correctamente no debería ser puesto en funcionamiento.

Para derivar eso, plantearemos qué valores/parámetros deberían ser auditados. Al respecto cabe aclarar que si bien en los últimos años se ha desarrollado una plétora[9] de valores y principios que los SIA deben promover y respetar, al día de hoy existe un acuerdo amplio de que deben incluirse *justicia* y *equidad*, *transparencia*, *confiabilidad/confianza*, *explicabilidad*, *rendición de cuentas/justificación* y *responsabilidad*. Independientemente de cuál es el conjunto preciso de valores y principios que debemos tener en cuenta, hay un fuerte consenso hacia dónde dirigirnos:

9 Ver https://ssrn.com/abstract=3518482

el objetivo es crear y promover una IA basada en valores humanos. Esto implica por un lado un intento sistemático de incluir valores de importancia ética en todo el ciclo de vida de un SIA y por otro hacernos responsables de los artefactos que construimos. Pero para responsabilizarnos de una pieza de tecnología necesitamos comprender para qué fue diseñada, cómo fue delineada para hacer eso y por qué fue diagramada de esa manera en particular en lugar de otra alternativa. Los productos de software, y en particular los que se desarrollan en base a encontrar patrones en los datos utilizando técnicas como el aprendizaje automático, no son diferentes. Comprender cómo se diseñan y construyen los sistemas informáticos, incluída la comprensión de cuándo quienes los diseñaron, hicieron concesiones entre objetivos en competencia y por qué, permite disipar gran parte de la inescrutabilidad que pueden tener los sistemas cuando se los ve como un todo o desde la perspectiva de alguien afectado por tal situación.

Hacia una IA responsable

La realidad socio-tecnológica en la que vivimos plantea una comprensión distinta de la ética respecto al control y la autonomía. Los SIA son un producto humano, están pensados, diseñados, construidos y usados por humanos, y por tanto, una aproximación basada en valores implica un compromiso de responsabilidad de todos los involucrados a través de todo el ciclo de vida de los SIA.

Responsabilidad implica necesariamente la habilidad de prevenir y medir.

1) Prevenir

Dignum[10] propone un enfoque a la ética de la IA apuntalado en tres dimensiones: ética en el diseño, ética por diseño, y ética para diseñadores. En esta sección pretendemos mostrar cómo a *priori* desde el diseño de los sistemas de aprendizaje automático es posible despejar la creencia de que estos son necesariamente inescrutables y que es imposible revertir esa condición de opacidad.

En la sección siguiente se expondrá cómo la explicación es una forma externa para lidiar a *posteriori* con dicha opacidad. Aquí abordaremos el problema a través del concepto de *desarrollo responsable*. En particular mostraremos cómo tales sistemas de IA pueden y deben entenderse en términos de sus objetivos de diseño y los mecanismos de su construcción y operación.

10 Responsible Artificial Intelligence: Designing Ai for Human Values. Dignum, Virginia. ITU Journal Special Issue No. 1. Sept. 2017

La ética por diseño se refiere a los métodos, algoritmos y herramientas necesarias para dotar a sistemas autónomos de la capacidad de razonar sobre los aspectos éticos de sus decisiones y los métodos, herramientas y formalismos para garantizar que el comportamiento de los mismos permanece dentro de límites morales que determinamos.

Una discusión sobre la primera parte queda fuera del alcance de este artículo, por lo que nos enfocaremos en la segunda. Asegurar que se mantiene esta dimensión humana en los resultados de los sistemas autónomos implica que todos los actores en el ciclo de vida de los mismos, incluyendo investigadores y profesionales de la IA, deberán tener la capacidad de tomar en cuenta valores morales, sociales y legales. Para poder lograr este objetivo se necesita trabajar en la forma de obtener, representar y traducir estos valores en técnicas y requisitos del sistema y poder demostrar que las soluciones de diseño efectivamente verifican los valores deseados.

Esto también implica un cambio cultural en el proceso de desarrollo de estos sistemas. La concepción de un sistema de IA pensado como un sistema socio-tecnológico requiere que mejorar el desempeño no sea el único objetivo conductor del proceso de diseño y desarrollo, sino que se consideren como básicas propiedades que tienen que ver con la interpretabilidad del modelo, la transparencia de todo el proceso, la capacidad de poder explicar las acciones o decisiones que el sistema toma y el establecimiento de una cadena de responsabilidad adecuada en relación a las mismas, como medios base para la implementación sistemática de los valores y principios de la IA.

Desarrollar de manera responsable sistemas de IA requiere desarrollar e implementar, a priori, medios que permitan vincular las decisiones del sistema de IA con el uso justo de los datos y las acciones de las distintas partes interesadas involucradas, directa o indirectamente, en los resultados que el sistema ofrece.

2) Medir

Si bien el asegurar que los sistemas de IA son diseñados de manera responsable favorece la confianza que podemos tener en ellos, eso no suele ser suficiente desde el punto de vista legal.

Además, aún cuando la información sobre el proceso de diseño casi siempre existe y puede mejorar la comprensión de un sistema informático, la misma generalmente no está disponible para las personas que la necesitan para su revisión. Y aunque pueda estarlo, quizás resulta poco útil para auditar un SIA, dado que suele ser difícil predecir completamente en la etapa de diseño cómo este interactuará con su contexto. Por ambas razones, se hace imprescindible disponer de métricas

y/o herramientas que permitan valorar la confiabilidad de un sistema en funcionamiento. Como ya ha sido mencionado en el capítulo anterior los sesgos pueden generarse en cualquiera de las etapas del ciclo de vida de un SIA: recolección de datos de entrenamiento o de conocimiento del dominio, en la generación del modelo y en la utilización del mismo. Para cada una de dichas etapas existen métricas y metodologías para evaluar y mitigar los comportamientos injustos en los SIA. Recientemente, han surgido un conjunto de herramientas de evaluación de equidad que son de código abierto con el fin de hacerlas ampliamente accesibles. Lo interesante de estas herramientas es que también pueden integrarse al proceso de desarrollo, pudiendo identificar los problemas en forma temprana. Algunos de esos modelos son: AIF 360 (IBM), Fairlearn (Microsoft), What-If (Google), Audit-AI (PyMetrics), Aequitas (UChicago).

Demos un ejemplo para clarificar la cuestión central de cómo actúan estos modelos. Supongamos que una entidad bancaria desea automatizar el otorgamiento de créditos personales hasta un cierto monto. Supongamos que recaba todas las solicitudes de los últimos 10 años para generar el modelo buscado. La primera cuestión que aparece es si los distintos segmentos poblacionales están bien representados, si hay atributos sensibles o protegidos (género por ejemplo), si los propios datos tienen "encriptado" un tratamiento injusto (porque la práctica histórica incluye factores ocultos), etcétera. La resolución de la injusticia que produciría entrenar un modelo con esos datos puede ser por reducción de la discriminación (ocultando los atributos sensibles en la generación del modelo) o por generar igualdad de oportunidades (ecualizando la representación de los distintos segmentos poblacionales). En la literatura, la forma de verificar la justicia de otorgamiento del préstamo se focaliza en tres medidas: la paridad demográfica que muestra que dentro de los casos que recibieron un crédito los distintos grupos tienen la misma presencia, la paridad predictiva que señala que el otorgamiento del crédito es independiente al grupo de pertenencia, y la paridad ecualizada determina que fijados ciertos atributos "relevantes" todos los grupos tienen la misma posibilidad de obtener el beneficio. Lamentablemente, hay un resultado teórico de imposibilidad que establece que esas tres métricas son mutuamente excluyentes, lo que "imposibilitaría" garantizar la completa equidad de un SIA. Esa es una de las razones por la cual muchas de las herramientas mencionadas arriba complementan su evaluación con métodos cualitativos de completar planillas de características. En particular, esa es la base del Marco de la OCDE para su metodología de clasificación de SIA.

Volviendo al ejemplo, supongamos que después de tratar y "limpiar" la base de entrenamiento, procedemos a generar el modelo que igualmente genera decisiones

inequitativas, por ejemplo, que las mujeres son sistemáticamente relegadas a pesar que el género no fue un atributo visible durante el entrenamiento. Eso puede pasar porque otros atributos son predictores indirectos del género. A esta altura tenemos identificado un efecto no deseado (que hasta pocos años atrás era invisible) que eventualmente se puede mitigar (hay algunos métodos conocidos) o en su defecto ser consciente de su imperfección y usar el SIA con cautela.

Para terminar esta sección dejaremos algunas preguntas de investigación en torno a la revisión de sistemas diseñados de manera responsable y correctamente auditados. Por ejemplo, incluso los sistemas que sortean todo el proceso propuesto de creación y evaluación pueden, en algunos casos, equivocarse o causar resultados negativos para sus sujetos. Para estos casos, es importante desarrollar una teoría de negligencia de software para igualar los regímenes de negligencia en otros campos como la medicina, el derecho y la ingeniería civil. Es importante destacar que el mero hecho de identificar un error no es suficiente para definir negligencia; más bien, la negligencia implica situaciones en las que los malos resultados podrían haberse evitado mediante un comportamiento más responsable por parte de quienes controlan un sistema. Hasta ahora, la cuestión de qué constituye, en concreto, un comportamiento suficientemente responsable está casi por completo inexplorada.

IA: si sos tan inteligente, ¡explicámelo! Explicabilidad en los sistemas con inteligencia artificial

Los SIA son ubicuos en nuestra vida cotidiana. Sin embargo, el tamaño y la complejidad de estos sistemas muchas veces dificultan que una persona pueda entender los mecanismos por los que una máquina toma una decisión. Este problema es más preocupante en el caso de los sistemas basados en aprendizaje automático. Especialmente en casos donde los riesgos asociados son grandes, como el uso de vehículos autónomos o el diagnóstico de una enfermedad, es importante que los sistemas empleados puedan explicar sus decisiones de manera comprensible y acorde al contexto de uso.

La explicabilidad debe hacer a un modelo más predecible y controlable que si no fuera explicable y esto debe ayudar a aumentar las capacidades humanas a la hora de tomar decisiones. Cuando los SIA pueden explicar sus decisiones se puede alcanzar mayor transparencia para derivar responsabilidades hacia las personas involucradas: desde quienes los desarrollan y auditan, hasta quienes los usan, según corresponda. Especialmente en contextos de riesgo o cuando se ven afectados los derechos de las personas.

Los desarrollos de modelos de aprendizaje automático más exactos y complejos suelen priorizar su rendimiento con respecto a alguna métrica de negocio antes que su capacidad de ser explicados, basándose en la presunción de que un mayor desempeño implica mayor complejidad y, por lo tanto, menor explicabilidad. Sin embargo, hay trabajos interesantes que muestran que esta idea es infundada y que la explicabilidad puede y debe ser desarrollada a la par de los nuevos sistemas.

Aquí llamaremos explicabilidad a la capacidad de un SIA de comunicar de forma eficaz a una persona las razones por las que tomó una determinada decisión. La explicabilidad de los SIA es un área activa de investigación donde confluyen y dialogan, entre otras, la computación, las ciencias sociales y la filosofía, además de los campos específicos donde se aplican estos sistemas como medicina, economía, derecho, biología, física, etcétera. En esta sección discutiremos la explicabilidad en cuanto a sus posibles definiciones, sus problemáticas actuales y sus perspectivas a futuro.

¿Qué es la explicabilidad?

La palabra *explicar* deriva del latín *explicare*[11], que significa *desdoblar* o *desplegar* —un concepto, una idea, un conocimiento complejo—. El modelo que usamos en una explicación busca facilitar la comprensión de quien la recibe. Entonces, podemos decir que una explicación involucra al menos dos nociones fundamentales: una intención y una interpretación. En el campo de la Inteligencia Artificial, la intención de una explicación debe ser pensada y discutida desde el desarrollo mismo del sistema. Esta intención debe tener en cuenta el público al que apunta y el contexto en el que será usado el SIA. Por ejemplo, al utilizarse en un contexto de diagnóstico por imágenes médicas, un SIA debe explicar sus decisiones a profesionales de la salud utilizando información específica del campo de la medicina y no del campo de las matemáticas. Por otro lado, la interpretación está sujeta a la persona que recibe la explicación: sus conocimientos previos, sus propias intenciones, sus sesgos, etcétera. Los métodos de explicabilidad se suelen utilizar para generar conocimiento, encontrar fallas en un sistema, justificar y mejorar los modelos.

Así como en la vida cotidiana, los SIA pueden exhibir distintos tipos de explicaciones. Ejemplos frecuentes son reglas de decisiones, explicaciones contrafácticas y las explicaciones basadas en ejemplos. Las reglas explicitan las relaciones entre características de un caso y la decisión que se tomó para ese caso, como una serie de pasos condicionados: si se cumple esta condición, luego seguir por este paso; si

11 En inglés, la traducción de *explicar* es *to explain* (del latín *ex-planare*, hacer plano). *Plain* se puede traducir como *simple*, *claro*; *to explain* sería simplificar o clarificar.

no, seguir por ese otro. De hecho, las reglas son un mecanismo básico para automatizar conocimiento experto y un modelo básico de aprendizaje automático. Consideramos que las reglas pueden ser buenas explicaciones si la cantidad de reglas y el detalle de las características son las adecuadas para la persona que requiere entender la decisión. Las explicaciones contrafácticas cuentan qué decisión se habría tomado de haber sido diferentes ciertas condiciones: por ejemplo, un valor de entrada diferente. Las explicaciones basadas en ejemplos resultan más fáciles de entender para cierto tipo de decisiones como las imágenes, en las que las características de los casos son presimbólicas (píxeles, conjuntos de píxeles). Entre las más efectivas, las explicaciones basadas en ejemplos adversarios nos ayudan a entender una decisión aportando casos con cambios mínimos para los que el SIA toma una decisión distinta. Al analizar las alternativas, una persona puede darse una intuición de cómo el SIA toma sus decisiones.

Interpretabilidad y transparencia

La explicabilidad está íntimamente relacionada con la transparencia (o su contraparte, la opacidad), la comprensibilidad y la interpretabilidad. En algunos trabajos, la transparencia indica la medida en que una persona puede ver o entender qué es lo que el sistema está haciendo. Un SIA completamente transparente es aquel que brinda toda la información necesaria para que una persona por su cuenta pueda llegar a las mismas respuestas a partir de los mismos datos. Dependiendo de la situación, esto en la práctica puede ser sencillo, como en un modelo simple con pocos datos (por ejemplo, si hace de 20 grados y sopla viento, me pongo abrigo), pero en la mayoría de los casos es irrealizable, por el tamaño de los modelos o su complejidad, especialmente en el caso de los modelos inferidos mediante aprendizaje automático.

En contraposición, un sistema puede ser opaco por diferentes razones: ser demasiado complejo como para que una persona lo pueda entender, o ser propietario, es decir que solo conocen y controlan su funcionamiento interno las personas a quienes pertenece. En ambos casos, para poder entender cómo el SIA llegó a la conclusión que tomó, se puede utilizar un modelo sustituto cuyo fin es explicar los pasos que tomó el primero, obteniendo así un sistema paralelo al original que resulte comprensible. Esta solución es eficaz pero puede ser problemática, ya que el segundo modelo no necesariamente es completamente fiel al primero y sus explicaciones pueden ser engañosas. En cambio, cuando el SIA por sí mismo es transparente, se dice que es interpretable.

Argumentos, responsabilidad y confianza

Otro concepto importante relacionado al de explicación es el de argumento. Mientras que una explicación intenta desplegar un fenómeno o una idea para hacerla más entendible, un argumento intenta dar razones pertinentes y convincentes por las que algo es o debe ser de determinada manera. Una explicación no necesariamente incluye un argumento en sus razonamientos o una causalidad explícita. Un SIA puede informar en su explicación qué relación hay entre los datos que usó como entrada y los que predijo como salida, pero esto no quiere decir que haya causalidad entre esas variables, ni que ese resultado sea justo o ético. En este sentido, podríamos pensar que una explicación es aceptable si se ajusta a los mecanismos de justificación que esperamos de otros seres humanos.

Tanto la comprensibilidad como la interpretabilidad se definen y adquieren su valor en la interacción de un sistema con una persona determinada, en un contexto en particular. Es decir, están sujetas a la capacidad de un individuo de entender lo que está sucediendo y de que esa explicación tenga sentido en el contexto en el que se esté usando el SIA. Sin embargo, incluso los SIA más avanzados solo procesan formas, pero no sentido. Por ejemplo, una computadora es capaz de procesar miles de libros e incluso mantener una conversación con un humano acerca de estos, pero es incapaz de entender que las palabras o las imágenes apuntan a objetos, ideas o fenómenos externos a ellos[12]. Entonces, para obtener explicaciones valiosas es necesario tener en cuenta la persona y el contexto en que se va a interpretar esa explicación.

Cuándo una explicación es válida y cuándo una decisión es justa también dependen del contexto en el que estén enmarcadas, y esto debe tenerse en cuenta a la hora de emplear un SIA, sobre todo cuando los riesgos de su uso son altos. En línea con lo indicado en las secciones de ética y sesgos, la explicabilidad debe utilizarse como una manera de facilitar la auditoría de un SIA durante su desarrollo y en su uso final. Este abordaje incluye no sólo a quienes desarrollan software, sino también a organismos libres de conflictos de interés, como un comité de ética, que incluya personas expertas en el campo de uso y personas que representen a quienes consintieron el uso de sus datos para la generación del SIA, entre otras. La explicabilidad entonces no debe entenderse como una manera de adjudicar la responsabilidad solamente a quien use un SIA para asistir a la toma de decisiones.

12 Para un ejemplo clásico, ver el argumento de la habitación china, de John Searle.

Cuando se puede facilitar una explicación pertinente y convincente para la persona, se avanza en la confianza de esa persona en la decisión automática, que es uno de los objetivos últimos en la integración de los mecanismos de inteligencia artificial en la sociedad.

Actualidad y perspectivas a futuro

A pesar de que la explicabilidad de los SIA se investiga desde hace varias décadas, los últimos cinco años han sido explosivos, con mayor profundidad teórica, más métodos disponibles y mayor discusión de sus usos que nunca. Aún así, persiste en algunas comunidades una falsa dicotomía entre rendimiento y explicabilidad. Afortunadamente, la explicabilidad de los SIA está en vías de convertirse en un derecho a nivel internacional, pero quedan muchos desafíos por delante para garantizar un uso ético, seguro, justo y confiable de esta tecnología.

Reflexión y alerta final

En este capítulo hemos intentado desmitificar a la Inteligencia Artificial para mostrar que no se trata de una tecnología abstracta que viene dada sino que está sustentada por técnicas computacionales concretas, que aunque algunas veces pueden ser complejas, siempre están mediadas por profesionales de la informática que diseñan sistemas basados en inteligencia artificial y que codifican sesgos y consideraciones éticas con más o menos conciencia de ello.

Desmitificar es un primer paso para entender, y una sociedad que comprende de qué trata una tecnología y es capaz de dominarla (y esto difiere enormemente de ser simplemente usuaria) tiene herramientas para moldear su propio futuro, desarrollándose social y económicamente.

Y en este sentido hacemos un llamado de alerta. La comunidad académica de ciencias de la computación (disciplina que ha sido motor decisivo de la transformación de nuestro mundo en estas últimas décadas, y que apuntala muchas de las nuevas tecnologías que continuarán esa transformación, incluyendo a la Inteligencia Artificial) está debilitándose en Argentina debido a la enorme demanda del sector industrial. La capacidad de formar y retener investigadores jóvenes en ciencias de la computación viene disminuyendo principalmente por la abrumadora brecha salarial que existe con lo ofrecido por la industria local y extranjera. La pérdida de estos recursos humanos ya está afectando la disponibilidad de docentes universitarios en informática con impacto inevitable sobre la calidad y cantidad

de personas egresadas de tecnicaturas, licenciaturas, ingenierías y doctorados, que puedan insertarse en la industria del software local. La falta de recursos humanos altamente capacitados ya está afectando la productividad del sector de software[13] en Argentina, y el panorama a mediano plazo es muy desalentador.

La Argentina necesita políticas científicas diferenciadas para sostener e incrementar la actividad científica en el área de ciencias de la computación.

13 Reporte de coyuntura 2020 y expectativas 2021, Cámara de la Industria Argentina del Software - CESSI.

ACADEMIA NACIONAL DE CIENCIAS DE BUENOS AIRES

Las tecnologías inteligentes: múltiples aspectos de su impacto

Juan Carlos Ferreri

Introducción

Los sistemas inteligentes comienzan su historia oficial luego de la segunda guerra mundial, como producto de la reubicación de la mano de obra científica en el ámbito privado. El desarrollo del conocimiento que había sido estratégico cambia de rumbo hacia una producción industrial. Nacen así las tecnologías inteligentes. Esta amalgama produce un cambio profundo que comienza en algunas partes específicas de la tecnología industrial y robótica, pero a lo largo de 20 años se enraíza en la sociedad productiva y comienza a infiltrar todos los ámbitos de la vida cotidiana. Este florecer de las Tecnologías Inteligentes (TIs) da origen a una vasta cantidad de áreas, diversificándolas y colocándolas es una posición estratégica para el desarrollo de la humanidad. Entre otras, se han destacado la Minería de Datos, el Análisis de tendencias y la causalidad para la predicción de sus consecuencias y el tratamiento inteligente de grandes volúmenes de datos, conocidos como Big Data. La lista es larga y crece constantemente. Todo este cambio ha llevado a la tecnología a diseñar un futuro conocido como Industria 4.0, donde el hombre, la sociedad y la naturaleza estarán entrelazados y sostenidos por dos grandes ejes: las comunicaciones y los sistemas inteligentes. Por otra parte, surgen aspectos nuevos a considerar, en particular las cuestiones éticas vinculadas a los desarrolladores de las TIs, la preservación de los neuro-derechos de los individuos y las posibilidades reales de control y regulación del uso de las "armas inteligentes" o autónomas. La robótica-nova y las implicancias de su desarrollo, serán también objeto de análisis en este trabajo. El análisis estará centrado en las cuestiones relacionadas con los impactos sociológicos, éticos y jurídicos. Un ejemplo también relevante es el impacto en el procesamiento y creación en las artes en general. Los aspectos mencionados y otros que, por limitación de longitud, no han sido considerados, serán motivo de un volumen especial de la ANCBA.