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

Bernardo Subercaseaux[1]

Resumen.

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

En cuanto a lo antropológico, en este ensayo propongo que el primer paso en pos de una convivencia saludable con modelos de IA es aceptar la oposición anteriormente mencionada y,

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

basándonos en ella, responder a consciencia la pregunta «¿cuánta interpretabilidad estamos dispuestos a sacrificar en pos de una mejor precisión?». Naturalmente, la respuesta a esta pregunta dependerá de la aplicación y contexto específico en que se utilicen los modelos; pero, en cualquier caso, propongo que plantearse esta pregunta deliberadamente constituye un deber ético para quienes utilizan tales modelos afectando a otros.

1. Ya están entre nosotros

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

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

En resumen, modelos de IA están entre nosotros, no solo en calidad de herramientas sino también en calidad de tomadores de decisiones, un rol que hasta hace poco estaba reservado para los seres humanos. Esta tendencia, además, parece aumentar mes a mes; desde una perspectiva financiera los modelos de inteligencia artificial son cada vez más rentables en tanto que tomadores de decisiones, y pueden utilizarse a escalas masivas que serían imposibles para seres humanos. Por ejemplo, cuando la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile ofrece el concurso literario Beauchef en 100 palabras, un jurado humano debe leer y evaluar cada uno de los cientos de relatos presentados, por lo que logísticamente sería complejo ofrecer un concurso como aquel más de una vez al año. En contraste, un modelo de IA podría evaluar la totalidad de los relatos recibidos en cuestión de segundos, permitiendo ofrecer semejantes concursos a diario. Es de esperarse que, en ausencia de regulaciones que garanticen a las y los ciudadanos un derecho a explicaciones para las decisiones tomadas por IA, como por ejemplo la GDPR en la Unión Europea[2], las empresas, en cuanto agentes que usualmente maximizan sus finanzas, tenderán a utilizar cualquier modelo que les otorguen beneficios financieros, sin importar si tales modelos son interpretables o no. Peor aún, una variedad de modelos de IA utilizados por empresas y gobiernos afecta desproporcionadamente a grupos tradicionalmente marginados, ya sea desde un punto de vista racial, de género, o socioeconómico[3]. Si bien los avances en IA ofrecen un sinfín de oportunidades tecnológicas nuevas, el foco de este artículo es la cara opuesta de tan brillante moneda: la convivencia con estos nuevos actores, pues los modelos de IA presentan desafíos éticos, sociales y matemáticos que no pueden ser ignorados.

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

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

2. Explicabilidad e interpretabilidad

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

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

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

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

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

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

En este caso, la dimensión asociada es $d = 7$, y la evaluación $f\left((1,0,1,1,0,1,0)\right) = 0$ corresponde al rechazo de la solicitud de una solicitante con las características descritas. Un modelo de IA ha de aprender esta función a partir de un conjunto de ejemplos, cada uno de los cuales se compone de la lista de atributos de una solicitud y de la decisión correcta o esperada. Idealmente, el modelo de IA será capaz de generalizar la información que yace implícita en los ejemplos sobre los cuales ha sido entrenado, y tomará decisiones razonables frente a ejemplos nuevos. Naturalmente, el uso de un tal modelo puede representar un ahorro significativo en tiempo y dinero para un banco que recibe un gran volumen de solicitudes de créditos, y que habitualmente utiliza humanos asalariados para evaluar cada una de ellas en ausencia de modelos. La pregunta al corazón de la explicabilidad, o interpretabilidad[5] es:

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

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

La interpretabilidad de un modelo, según Miller (2019), corresponde al grado en el cual un humano determina la causa de sus decisiones. Esta definición, sin embargo, no aclara el sentido en el cual las decisiones son «causadas»; más aún, Miller utiliza la palabra «causa» en singular, asumiendo implícitamente que para la noción de causalidad en juego existirá una única causa para cada decisión[6].

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

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

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

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

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

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

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

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

3. Interpretabilidad desde la complejidad computacional

Esta sección asumirá una familiaridad básica con la teoría de la complejidad computacional[7]. La hipótesis fundamental de esta sección es que la interpretabilidad de una clase de modelos de IA está relacionada con la complejidad computacional de encontrar explicaciones para modelos de esta clase. Para aceptar esta hipótesis pareciera ser necesario aceptar al menos un par de premisas que ilustraré a continuación.

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

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

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

problemas para los cuales contamos con algoritmos que requieren un número polinomial de pasos en el tamaño de la entrada serán resueltos en la práctica[8].

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

4. Interpretabilidad y Precisión

Dado un conjunto $D = \{(\overrightarrow{x_1}, b_1), \ldots, (\overrightarrow{x_n}, b_n), \ldots\}$ de datos $\overrightarrow{x_n}$ etiquetados con su clasificación deseada b_n, la manera tradicional de evaluar la precisión de los modelos que entrenamos sobre ellos está basada en particionar D en dos: un subconjunto $T \subseteq D$ de datos se utilizará para entrenar el modelo, y un subconjunto $E = D \setminus T$ se utilizará para evaluar la precisión del modelo ya entrenado (Hastie et al., 2001). De este modo, si un modelo M_T resulta de entrenar sobre T, definiremos su «precisión» sobre el conjunto E según:

$$p_E = \frac{|\{\overrightarrow{x_m} \in E \; tal \; que \; M(\overrightarrow{x_m}) = b_m\}|}{|E|},$$

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

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

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

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

4.1. La tesis de Rudin

La científica norteamericana Cynthia Rudin plantea en su célebre trabajo *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* (Rudin, 2019) que, en lugar de intentar explicar modelos complejos de IA como son las redes neuronales, deberíamos en su lugar utilizar modelos más interpretables como son los árboles de decisión. A mi juicio hay dos problemas con esta tesis. En primer lugar, como mencioné en la sección 2[9], no es claro que modelos como los árboles de decisión sean efectivamente interpretables en la práctica. Más estudios experimentales en la línea de Piltaver et al. (2016) serán necesarios para esclarecer este problema. En segundo lugar, y posiblemente más importante, la tesis de Rudin se basa en la idea de que es posible obtener modelos interpretables con la misma precisión que aquellos menos interpretables[10]. Dicho en breve, el argumento de Rudin es que el espacio de modelos (entendidos como funciones) cuya precisión es cercana al óptimo (i.e., Conjuntos de *Rashomon*) es suficientemente grande como para esperar que contenga modelos interpretables. En mi opinión, hay dos problemas cruciales con esta hipótesis, que planteo a continuación.

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

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

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

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

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

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

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

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

en problemas complejos[11], de la misma forma en que nadie ha obtenido un programa que resuelva instancias de SAT en tiempo polinomial. Por supuesto, «la ausencia de evidencia no es evidencia de la ausencia», y es posible que en el futuro se descubran modelos interpretables que resuelvan problemas complejos con igual o mayor precisión que las redes neuronales profundas.

5. Contra la objeción de Goodfellow

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

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

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

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

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

6. Preguntas para guiar nuestra convivencia

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

¿cuánta interpretabilidad estamos dispuestos a sacrificar en pos de la precisión?

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

Referencias bibliográficas

Arenas, Marcelo, Báez, Daniel, Barceló, Pablo, Pérez, Jorge and Subercaseaux, Bernardo (2021). «Foundations of Symbolic Languages for Model Interpretability». In *Advances in Neural Information Processing Systems*, volume 34, pages 11690–11701. Curran Associates, Inc.

Arenas, Marcelo, Barceló, Pablo, Romero Orth, Miguel and Subercaseaux, Bernardo (2022). «On Computing Probabilistic Explanations for Decision Trees». In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28695–28707. Curran Associates, Inc..

S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2006. ISBN 978-0-521-42426-4. URL https://theory.cs.princeton.edu/complexity/book.pdf.

Barceló, Pablo, Monet, Mikaël, Pérez, Jorge and Subercaseaux, Bernardo. «Model Interpretability through the lens of Computational Complexity». In *Advances in Neural Information Processing Systems*, volume 33, pages 15487–15498. Curran Associates, Inc., 2020a.

Barceló, Pablo, Pérez, Jorge and Subercaseaux, Bernardo. *Foundations of Languages for Interpretability and Bias Detection*. AFCI, 2020b.

Arriagada Bruneau, Gabriela (2024). *Los sesgos del algoritmo*. La Pollera Ediciones: Santiago. ISBN 9789566267256.

Buolamwini, Joy and Gebru, Timnit. «Gender shades: Intersectional accuracy disparities in commercial gender classification». In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24, Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Burke, Lilah. U of Texas will stop using controversial algorithm to evaluate Ph.D. applicants — insidehighered.com. https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd, 2020. [Accedido el 15-05-2024].

Cobham, Alan. The intrinsic computational difficulty of functions. In Yehoshua Bar-Hillel, editor, *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress (Studies in Logic and the Foundations of Mathematics)*, pages 24–30. North-Holland Publishing, 1965.

Darwiche, Adnan and Hirth, Auguste. «On the Reasons Behind Decisions». In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), pages 712–720, 2020. doi: 10.3233/FAIA200158.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Markus D. Dubber, Frank Pasquale, and Sunit Das, editors. *The Oxford Handbook of Ethics of AI*. Oxford University Press, July 2020. ISBN 978-0-19-006739-7. doi: 10.1093/oxfordhb/9780190067397.001.0001.

Edmonds, Jack. *Paths, trees, and flowers*. Canadian Journal of Mathematics, 17:449–467, 1965. doi: 10.4153/ CJM-1965-045-4.

Hinton, Geoffrey. Question — twitter.com @geoffreyhinton. https://twitter.com/geoffreyhinton/status/1230592238490615816?lang=en, 2020. [Accedido el 15-05-2024].

Goodfellow, Ian. *How important is interpretability for a model in Machine Learning?* — quorasessionwithiangoodfellow.quora.com. https://quorasessionwithiangoodfellow.quora.com/How-important-is-interpretability-for-a-model-in-Machine-Learning?ch=1&share=b5056dcf, 2017. [Accedido el 19-05-2024].

Goodman, Bryce and Flaxman, Seth. *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*. AI Magazine, 38(3):50–57, October 2017. doi: 10.1609/aimag.v38i3.2741.

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001). *The Elements of Statistical Learning. Springer Series in Statistics*. Springer New York Inc., New York, NY, USA.

Izza, Yacine, Ignatiev, Alexey and Marques-Silva, Joao (2020). *On explaining decision trees*.

Kirchner, Lauren, Larson, Jeff, Angwin, Julia and Mattu, Surya (2016). *How We Analyzed the COMPAS Recidivism Algorithm* — propublica.org. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accedido el 15-05-2024].

Kirchner, Lauren, Angwin, Julia, Larson, Jeff and Mattu, Surya (2016). *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks*— propublica.org. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accedido el 16-05-2024].

Karatsuba, Ekaterina. Fast computation of the riemann zeta function. s/ for integer values of s. Problems of Information Transmission, 31:353–362, 01 1995.

Kelly, Jack (2024). *Your Next Job Interview May Be With 'Alex,' The AI Interviewer* — forbes.com. https://www.forbes.com/sites/jackkelly/2024/05/10/your-next-job-interview-may-be-with-alex-the-ai-interviewer/?sh=7cac4c4f76f2. [Accessed 15-05-2024].

Knuth, Donald (2014). *Twenty Questions for Donald Knuth | | InformIT* — informit.com. https://www.informit.com/articles/article.aspx?p=2213858&WT.mc_id=Author_Knuth_20Questions. [Accedido el 18-05-2024].

Kozachinskiy, Alexander (2023). *Inapproximability of sufficient reasons for decision trees*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto.

Lipton, Zachary C. *The mythos of model interpretability*. Queue, 16(3):31–57, 2018.

Malmon, Judy (2023). *How Are Algorithms Used in the Criminal Justice System?* — superlawyers.com. https://www.superlawyers.com/resources/criminal-defense/how-are-algorithms-used-in-the-criminal-justice-system/. [Accedido el 15-05-2024].

Marques-Silva, Joao (2023). Logic-based explainability in machine learning.

Marques-Silva, Joao and Ignatiev, Alexey. *Delivering Trustworthy AI through Formal XAI*. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12342–12350, June 2022. doi: 10.1609/aaai.v36i11.21499.

Marques-Silva, Joao and Ignatiev, Alexey. *No silver bullet: Interpretable ML models must be explained*. Frontiers in Artificial Intelligence, 6, 2023. ISSN 2624-8212.

Miller, George A. *The magical number seven, plus or minus two: Some limits on our capacity for processing information*. Psychological Review, 63(2):81–97, 1956. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/h0043158.

Miller, Tim. *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence, 267: 1–38, 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007.

Molnar, Christoph (2022). *Interpretable Machine Learning*. 2 edition. URL https://christophm.github.io/interpretable-ml-book.

O'Neil, Cathy (2016*). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, USA. ISBN 0553418815.

Pearl, Judea and Mackenzie, Dana (2018). *The Book of Why: The New Science of Cause and Effect.* Basic Books, Inc., USA, 1st edition. ISBN 046509760X.

Piltaver, Rok, Luštrek, Mitja, Gams, Matjaž and Martinčić-Ipšić, Sanda (2016). *What makes classification trees comprehensible? Expert Systems with Applications*, 62:333–346, 2016. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.06.009.

Pringle, Eleanor. *Bumble founder says your dating 'AI concierge' will soon date hundreds of other people's 'concierges' for you* — fortune.com. https://fortune.com/2024/05/10/bumbles-whitney-wolfe-herd-dating-concierge-artificial-intelligence/, 2024. [Accedido el 15-05-2024].

Romera-Paredes, Bernardino, Barekatain, Mohammadamin, Novikov, Alexander, Balog, Matej, Kumar, M. Pawan, Dupont, Emilien, J. R. Ruiz, Francisco, Ellenberg, Jordan S., Wang, Pengming, Fawzi, Omar, Kohli, Pushmeet and Fawzi, Alhussein. *Mathematical discoveries from program search with large language models*. Nature, 625(7995): 468–475, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06924-6.

Roughgarden, Tim (2021). *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press.

Rudin, Cynthia. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/ s42256-019-0048-x.

Sapia.ai. Sapia.ai | Hire top talent, faster, with AI Smart Interviewing — sapia.ai. https://sapia.ai/, 2024. [Accedido el 15-05-2024].

Semenova, Lesia, Rudin, Cynthia and Parr, Ronald. *On the existence of simpler machine learning models*. In 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533232. URL http://dx.doi.org/10.1145/3531146.3533232.

Talently.ai. Talently.ai: Your AI Interviewer — interview.talently.ai. https://interview.talently.ai/, 2024. [Accedido el 15-05-2024].

Wegener, Ingo. *Branching Programs and Binary Decision Diagrams*. Society for Industrial and Applied Mathematics, January 2000. doi: 10.1137/1.9780898719789.

Papers with Code. Papers with Code - ImageNet Benchmark (Image Classification) — paperswithcode.com. https://paperswithcode.com/sota/image-classification-on-imagenet, 2024a. [Accedido el 19-05-2024].

Papers with Code. Papers with Code - CIFAR-10 Benchmark (Image Classification) — paperswithcode.com. https://paperswithcode.com/sota/image-classification-on-cifar-10, 2024b. [Accedido el 19-05-2024].