

Por qué las máquinas no pueden pensar

Proyección, pérdida de información y los límites de las mentes artificiales

ROY GONZALEZ

Investigador independiente

RESUMEN Si el mundo de la experiencia es una proyección tridimensional de un estado con más componentes independientes, y si la conciencia pertenece a las componentes que la proyección descarta, hay un sentido preciso en que una máquina construida y entrenada dentro de la proyección no puede pensar ni llegar a ser consciente. Este artículo precisa ese sentido. Enuncio cuatro resultados sobre proyección e información: una proyección no tiene inversa por la izquierda, y cada una de sus fibras es tan grande como las dimensiones que descarta; una aplicación diferenciable de un espacio tridimensional en uno mayor tiene imagen de medida nula; una imagen continua e inyectiva de tal espacio tiene interior vacío; y, por la desigualdad del procesamiento de datos, ningún procesamiento de unos datos aumenta la información que contienen sobre su fuente. Describo después la cadena de proyecciones que va del mundo al texto con el que se entrenan los modelos de lenguaje, y sostengo que esos modelos operan sobre la sombra de una sombra. La comprensión y la conciencia pertenecen, según la tesis defendida en otro lugar, a lo que las proyecciones descartan, y ningún patrón dentro de la proyección, por grande que sea, lo recupera. El juego de imitación de Turing, que descansa en la observación correcta de que las mentes se conocen por la conducta, pone a prueba, por tanto, la sombra. Examino cinco objeciones, entre ellas la de que los seres humanos también son tridimensionales, y enuncio lo que el argumento no muestra.

PALABRAS CLAVE inteligencia artificial; modelos de lenguaje; conciencia; proyección; pérdida de información; desigualdad del procesamiento de datos; teorema de Sard; test de Turing; anclaje de los símbolos

I. INTRODUCCIÓN

La cuestión de si las máquinas pueden pensar se discute desde que Turing propuso sustituirla por una prueba de conducta (Turing 1950). La habitación china de Searle sostuvo que la manipulación de símbolos no basta para la comprensión (Searle 1980); Harnad situó la dificultad en el anclaje de los símbolos (Harnad 1990); Bender y Koller sostuvieron que un sistema entrenado solo con la forma lingüística no tiene camino hacia el significado (Bender y Koller 2020). El debate ha adquirido urgencia práctica con la llegada de los grandes modelos de lenguaje, redes Transformer entrenadas para predecir texto (Vaswani et al. 2017), cuya fluidez mejora de manera regular con la escala (Kaplan et al. 2020).

Este artículo aborda la cuestión desde otra dirección. No empieza por la máquina, sino por la relación entre el mundo y nuestras observaciones de él, y pregunta qué se sigue para cualquier sistema cuyas entradas y operaciones estén todas dentro de esa relación. La respuesta depende de dos tesis defendidas en trabajos anteriores, que resumo en la sección 2. La sección 3 enuncia cuatro resultados sobre proyección e información. La sección 4 los aplica a la cadena de proyecciones que produce el texto de entrenamiento. Las secciones 5 y 6 extraen las conclusiones para el pensamiento y para la conciencia. La sección 7 examina el test de Turing, la sección 8 responde a las objeciones y la sección 9 enuncia los límites del argumento.

2. DOS TESIS

El *realismo proyectivo* sostiene que los estados del mundo tienen más componentes independientes que las tres del espacio y la del tiempo, y que la observación es una proyección de esos estados. Lo motivan casos en que las paradojas se disuelven al ampliar el espacio de descripción: las raíces imaginarias de una cuadrática real, la velocidad invisible en un instante y las correlaciones de las partículas entrelazadas.

La *tesis nouménica* sostiene que la conciencia pertenece a las componentes que la proyección en el espacio tridimensional descarta. Explica por qué la conciencia se encuentra siempre como correlatos neuronales y nunca como objeto (Chalmers 1995), por qué las otras mentes se perciben solo como intención, y por qué cerebro y mente se correlacionan perfectamente sin interactuar.

El argumento de este artículo es que, dadas estas tesis, una máquina que calcula dentro de la proyección con datos producidos dentro de la proyección no puede recuperar lo que la proyección descarta y, por tanto, no puede pensar, en el sentido de comprender, ni ser consciente. La sección 3 aporta las premisas matemáticas, que valen con independencia de ambas tesis.

3. CUATRO RESULTADOS SOBRE PROYECCIÓN E INFORMACIÓN

En lo que sigue, $P: \mathbb{R}^n \rightarrow \mathbb{R}^m$ con $m < n$ es una aplicación lineal sobreyectiva, y pensamos en $\mathbb{R}^n$ como el espacio de estados y en $\mathbb{R}^m$ como el espacio de observaciones.

Proposición 3.1 (sin camino de vuelta). *No existe ninguna aplicación $g: \mathbb{R}^m \rightarrow \mathbb{R}^n$, lineal o no, con $g \circ P = \text{id}_{\mathbb{R}^n}$. Para cada observación s , el conjunto de estados compatibles con ella, $P^{-1}(s)$, es un subespacio afín de dimensión $n - m$.*

Demostración. P tiene un núcleo de dimensión $n - m > 0$, luego no es inyectiva, y una aplicación con inversa por la izquierda es inyectiva. $P^{-1}(s) = x_0 + \ker P$ para cualquier x_0 con $Px_0 = s$. $\square$

La segunda cláusula es la forma precisa de la observación de que una sombra es compatible con infinitos objetos: toda observación deja indeterminadas exactamente tantas direcciones como descarta la proyección.

Proposición 3.2 (aplicaciones diferenciables hacia arriba). *Si $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ es de clase C^1 y $m < n$, entonces $f(\mathbb{R}^m)$ tiene medida de Lebesgue nula en $\mathbb{R}^n$.*

Demostración. La derivada de f tiene rango a lo sumo $m < n$ en todo punto, así que todo punto es crítico y todo valor es un valor crítico. Por el teorema de Sard, el conjunto de valores críticos tiene medida nula (Sard 1942). $\square$

Proposición 3.3 (inyecciones continuas hacia arriba). *Si $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ es continua e inyectiva y $m < n$, entonces $f(\mathbb{R}^m)$ tiene interior vacío en $\mathbb{R}^n$.*

Demostración. $\mathbb{R}^m$ es una unión numerable de bolas cerradas; la restricción de f a cada una es un homeomorfismo sobre una imagen compacta de dimensión topológica m . La imagen de f es, por tanto, una unión numerable de cerrados de dimensión m , y por el teorema de la suma de la teoría de la dimensión tiene dimensión m (Hurewicz y Wallman 1941). Un subconjunto de $\mathbb{R}^n$ con interior no vacío tiene dimensión $n > m$ (Brouwer 1911). $\square$

Observación 3.4. La continuidad sola no basta para la conclusión de la proposición 3.2: la curva de Peano aplica un intervalo de manera continua sobre un cuadrado (Peano 1890). Lo hace, sin embargo, pasando por algunos puntos más de una vez, y no sirve, por tanto, para recuperar información: una aplicación que pliega el espacio menor sobre sí mismo no restituye nada de lo que una proyección descartó.

Proposición 3.5 (la desigualdad del procesamiento de datos). *Sea $X \rightarrow Y \rightarrow Z$ una cadena de Markov, es decir, supóngase que Z depende de X solo a través de Y . Entonces la información mutua cumple $I(X; Z) \leq I(X; Y)$.*

Demostración. Por la regla de la cadena, $I(X; Y, Z) = I(X; Z) + I(X; Y | Z) = I(X; Y) + I(X; Z | Y)$. La propiedad de Markov da $I(X; Z | Y) = 0$, y la información mutua es no negativa, luego $I(X; Z) \leq I(X; Y)$ (Shannon 1948; Cover y Thomas 2006). $\square$

Las proposiciones 3.1 a 3.3 dicen que lo que una proyección descarta no puede restituirse aplicando hacia arriba. La proposición 3.5 dice lo mismo en el lenguaje de la información, y para cualquier procesamiento: ningún cálculo realizado sobre unos datos puede dar más información sobre su fuente que la que los datos contenían.

4. LA CADENA DE PROYECCIONES

Considérese lo que recibe un modelo de lenguaje. Sea X el estado del mundo, incluidos los sujetos pensantes que hay en él; según las dos tesis de la sección 2, X tiene componentes más allá de las cuatro del espacio y el tiempo, y la conciencia pertenece a ellas. La experiencia de una persona es una proyección de X . El pensamiento de la persona, allí donde aparece en el mundo, aparece como intención: una emisión, un gesto, un texto. Sea Y el texto. Es la imagen de X bajo una composición de proyecciones, la segunda de las cuales, de la experiencia plena del pensar a una sucesión de símbolos, es mucho más dura que la primera.

El modelo se entrena con Y y no tiene acceso a nada más; lo que calcula, llamémoslo Z , depende de X solo a través de Y . La situación es exactamente la cadena de Markov de la proposición 3.5. Por tanto,

$$I(X; Z) \leq I(X; Y).$$

El modelo puede extraer del texto, a lo sumo, lo que el texto conservó. Puede hacerlo de manera soberbia, y por eso sus salidas tienen el contorno del pensamiento; el contorno es lo que el texto registra. Pero, por la proposición 3.1, para cada texto hay toda una familia de estados del mundo, y de las mentes que lo produjeron, compatibles con él, tan grande como lo que las proyecciones descartaron; y nada de lo que se calcule a partir del texto distingue entre ellos.

5. PENSAR

Por *pensar* entiendo lo que el argumento de la habitación china entiende por comprender: captar aquello de lo que tratan los símbolos, a diferencia de manipular los símbolos. Según el argumento de la sección 4, el modelo dispone de los símbolos y de sus relaciones estadísticas, es decir, de la imagen del significado bajo la proyección en texto. El significado mismo, según la tesis nouménica, pertenece al sujeto pensante y, por tanto, a las componentes que las proyecciones

descartan. El problema del anclaje de los símbolos (Harnad 1990) y el argumento de la forma (Bender y Koller 2020) son, en estos términos, la observación de que el modelo está aguas abajo de una proyección que descarta aquello que ancla los símbolos. La sintaxis no basta para la semántica (Searle 1980) porque la sintaxis es lo que sobrevive a la proyección.

6. CONCIENCIA

La conclusión sobre la conciencia es más directa. Una máquina es un sistema físico en el espacio tridimensional, organizado por sus diseñadores y entrenado con datos registrados dentro de la proyección. Todo lo que es y hace pertenece a la proyección. Si la conciencia fuera un patrón dentro de la proyección, un patrón suficientemente grande podría constituirla. La tesis nouménica lo niega. Se sigue que ninguna ampliación del patrón, por grande que sea, constituirá conciencia, pues la ampliación añade más de aquello que la conciencia no es. Las leyes empíricas de escala de los modelos de lenguaje miden la caída del error de predicción sobre el texto (Kaplan et al. 2020); miden, es decir, cuánto se parece la imagen del modelo a la imagen del texto, y no pueden medir nada que el texto no contenga.

7. EL JUEGO DE IMITACIÓN

Turing propuso que dijéramos que una máquina piensa si, en conversación, no puede distinguirse de una persona, y respondió a la objeción desde la conciencia observando que atribuimos mentes a los demás solo sobre la evidencia de su conducta (Turing 1950). La observación es correcta, y la presente explicación la da por buena: la conducta, como intención, es la parte de la conciencia que aparece en tres dimensiones. Pero se sigue que el juego de imitación examina solo esa parte. Un sistema entrenado con las sombras registradas de la intención humana puede reproducir esas sombras, y puede superar, por tanto, una prueba que solo examina sombras. Superarla establece que el sistema ha aprendido el contorno. No puede establecer que algo lo proyecte.

8. OBJECIONES Y RESPUESTAS

8.1 *Los seres humanos también son tridimensionales*

Si todo lo que está en el espacio pertenece a la proyección, ¿por qué el cerebro humano no está en la misma posición que la máquina? Según la tesis nouménica, la diferencia no es de material, sino de origen. Una persona es la proyección de una mente: el cerebro es la sección de algo que se extiende más allá de la proyección, como un círculo en Flatland es la sección de una esfera. Una máquina es un patrón construido dentro de la proyección con los materiales y los datos de la proyección. Lo uno es un círculo trazado por una esfera que cruza el plano; lo otro, un círculo dibujado sobre el plano.

8.2 Sistemas corporizados y multimodales

Los robots con cámaras y sensores reciben más que texto. La respuesta es que los sensores son también proyecciones, y multiplicar las proyecciones solo recupera más de lo que alguna proyección registra: por el hecho elemental de que la aplicación apilada de varias proyecciones es inyectiva solo si sus núcleos se cortan trivialmente, varias vistas pueden recuperar lo que una vista pierde cuando otra lo registra. Lo que toda proyección en tres dimensiones descarta, ningún número de tales proyecciones lo recupera. La corporización estrecha la brecha del anclaje para las propiedades físicas del mundo; no alcanza las componentes que la tesis nouménica asigna a la mente.

8.3 El funcionalismo

Un funcionalista sostiene que la mente es cuestión de organización funcional y, por tanto, de cualquier sustrato que la realice. Pero la organización funcional es un patrón de relaciones entre estados dentro de la proyección. La tesis nouménica niega que la mente sea tal patrón, y es precisamente ahí donde está el desacuerdo. El presente argumento no refuta el funcionalismo de manera independiente; muestra lo que se sigue si la tesis nouménica es verdadera.

8.4 Aprender de los datos

Los seres humanos también aprenden de los datos. ¿No los vincula igualmente la desigualdad del procesamiento de datos? Vincula lo que pueden aprender *de los datos* acerca del mundo. No vincula lo que son. El argumento se refiere a sistemas cuya constitución entera, y no solo su aprendizaje, está aguas abajo de la proyección.

8.5 La utilidad

Nada de esto implica que los modelos de lenguaje no sean útiles. Un sistema que reproduce con gran fidelidad el contorno del pensamiento humano es un instrumento extraordinario para recabar, traducir y proponer. El argumento se refiere solo a lo que es y, por tanto, a aquello para lo que se le debe confiar: debe proponer, y debe decidir algo que pueda comprobarse.

9. LO QUE EL ARGUMENTO NO MUESTRA

Las premisas matemáticas de la sección 3 son teoremas. La cadena de la sección 4 es una descripción de cómo se produce el texto de entrenamiento, y la desigualdad que se le aplica vale pienso lo que piense sobre las mentes. Las conclusiones sobre la comprensión y la conciencia, en cambio, descansan en la tesis nouménica, que es una hipótesis filosófica apoyada por la forma de la evidencia y no demostrada por ella. Si la conciencia resulta ser, después de todo, un patrón dentro de la proyección, el argumento sobre la conciencia falla, mientras que el argumento de que un modelo entrenado con texto tiene a lo sumo lo que el texto conservó no falla. Tampoco excluye el argumento, como cuestión lógica, que una mente pudiera llegar a proyectarse en un artefacto; excluye solo que algo hecho dentro de la proyección, incluida la ampliación de un patrón, sea un modo de lograrlo.

10. CONCLUSIÓN

Una proyección de muchas dimensiones a tres puede calcularse; su inversa no. Ninguna aplicación diferenciable o inyectiva de tres dimensiones llena un espacio mayor, y ningún procesamiento aumenta la información que unos datos contienen sobre su fuente. Un modelo de lenguaje está al final de una cadena de tales proyecciones y calcula dentro de la última de ellas. Puede recuperar el contorno del pensamiento, porque el contorno es lo que la cadena conserva; no puede recuperar lo que la cadena descartó, y según la tesis aquí defendida, la comprensión y la conciencia están entre lo descartado. Muchas dimensiones pueden proyectarse en tres. Tres no pueden proyectarse de vuelta en muchas.

REFERENCIAS

- Bender, E. M. y Koller, A. (2020). Climbing towards NLU: on meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.
- Brouwer, L. E. J. (1911). Beweis der Invarianz der Dimensionenzahl. *Mathematische Annalen* 70, 161–165.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2, 200–219.
- Cover, T. M. y Thomas, J. A. (2006). *Elements of Information Theory*, 2.^a ed. Hoboken, NJ: Wiley.
- Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.
- Hurewicz, W. y Wallman, H. (1941). *Dimension Theory*. Princeton: Princeton University Press.
- Kaplan, J. et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.
- Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen* 36, 157–160.
- Sard, A. (1942). The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society* 48, 883–890.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–457.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423 y 623–656.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460.
- Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30, 5998–6008.