EL NUEVO ARGUMENTO DE PENROSE Y LA NO-LOCALIDAD DE LA CONCIENCIA

RUBÉN HERCE
Universidad de Navarra

RESUMEN: Roger Penrose formuló en 1989 un argumento contra la IA. Dicho argumento concluye que la explicación científico-matemática de la realidad es más amplia que la meramente computacional, porque existen ciertos aspectos de la realidad no-computables. Este artículo analiza dicho argumento y la discusión al respecto, para concluir que el tipo de argumento que quiere desarrollar Penrose está viciado de raíz, lo que impide llegar a las conclusiones deseadas. A la vez se sostiene la validez filosófica de sus conclusiones y se apunta a la idea de no-localidad como consistente al hablar sobre la conciencia.

PALABRAS CLAVE: Penrose; conciencia; Gaifman; no-localidad; Inteligencia Artificial.

Penrose's new argument and the non-locality of consciousness

ABSTRACT: In 1989 Roger Penrose formulated an argument against AI. This argument concludes that the scientific-mathematical explanation of reality is broader than the merely computational, because there are certain non-computational aspects of reality. This article analyzes the argument and the discussion about it, to conclude that the type of argument that Penrose wants to develop is tainted at the root, what prevents reaching the wished conclusions. At the same time the philosophical validity of his conclusions is maintained and the idea of non-locality is pointed as sound when speaking about consciousness.

KEY WORDS: Penrose; Consciousness; Gaifman; Non-Locality; Artificial Intelligence.

Introducción

Roger Penrose formuló en 1989 un argumento contra la Inteligencia Artificial desde un punto de vista filosófico-matemático. Lindström[1] lo bautizó como «Nuevo Argumento de Penrose» y así es conocido en la actualidad. Es un argumento que tenía sus predecesores en los argumentos de Lucas[2] y que en síntesis sostiene la existencia de aspectos científicos y matemáticos que no se pueden simular computacionalmente; de modo que la explicación científico-matemática de la realidad es más amplia que la meramente computacional[3].

¿Tiene esta crítica un alcance y profundidad suficiente para desterrar toda propuesta de Inteligencia Artificial en sentido fuerte? A esa pregunta se intenta dar respuesta en este artículo, que se desarrolla del siguiente modo: (1) se comienza por aclarar qué entiende Roger Penrose por computable; (2) después

[1] Lindström, P., «Penrose's New Argument», en: *Journal of Philosophical Logic* 30 (3), 2001, pp. 241-250.

[2] Lucas, J. R., «Minds, Machines and Gödel», en: Philosophy 36, 1961, pp. 112-127

[3] Cfr. Penrose, R., «The need for a non-computational extension of quantum action in the brain», en: Arhem, Peter, Liljenstrom, Hans y Svedin, Uno (eds.), *Matter matters? On the material basis of the cognitive activity of mind*, Springer, Berlin 1997, p. 13.

doi: 10.14422/pen.v78.i298.y2022.003

se sigue su argumentación partiendo del pensamiento matemático, para (3) desarrollar un argumento que está entrelazado en tres niveles; posteriormente (4) se analiza la crítica de Gaifman a este tipo de argumentos y (5) se extraen unas conclusiones sobre la imposibilidad del argumento que intenta desarrollar Penrose y sobre la validez filosófica de sus observaciones, para terminar (6) apuntando al concepto de no-localidad, aplicado a la conciencia.

1. Lo computable en la teoría y en la práctica

Al hablar de computación conviene distinguir entre una computabilidad práctica y una teórica: entre aquello que en la actualidad no es computable, pero podría serlo en el futuro, y aquello que escapa a esta posibilidad. Así y según Penrose, computable sería todo aquello que en teoría podría realizar una máquina de Turing[4]. De aquí la equivalencia entre computable y algorítmico, que *a priori* no tiene por qué ser evidente.

En los algoritmos de inteligencia artificial se introducen elementos, cuando menos, difícilmente computables como: sistemas caóticos, aleatoriedad, procedimientos de aprendizaje… Para Penrose todos ellos son computables, porque los aspectos de no-computabilidad que pueden implicar o son simplemente limitaciones o se perderían al ser interiorizados por la computación (también con la cuántica)[5].

Por ejemplo, en un sistema caótico hay un procedimiento computacional en sí y unas circunstancias que lo limitan. En la práctica no se pueda obtener el resultado con precisión absoluta, aunque en teoría sería posible si se eliminasen las limitaciones. Es decir, un superordenador idealizado obtendría la precisión buscada. El caos sería un ejemplo de no-computabilidad práctica pero no teórica. ¿Se podría decir lo mismo de la computación analógica? Según Penrose, sí, ya que sería una mera cuestión de precisión, inalcanzable en la práctica, pero no teóricamente.

¿Y qué opina sobre la existencia de fenómenos estrictamente aleatorios? Dice que en ámbito computacional solo existen sistemas *pseudo-aleatorios* y que los fenómenos estrictamente aleatorios tienen un elemento no-computacional, si bien no sería del tipo que se está buscando, ya que la pura aleatoriedad no aporta diferencias significativas respecto a lo que se puede simular con un ordenador.

¿Y respecto a aquellos sistemas computacionales que están abiertos a la interacción con el medio ambiente? En este caso cada interacción es única e irrepetible y el hecho individual no se puede simular. Su respuesta es que, al igual que con los sistemas caóticos se trata de una no-computabilidad práctica pero no teórica, ya que se podría realizar una simulación típica o plausible del entorno. Estas últimas consideraciones abren nuevas perspectivas sobre

[4] *Ibíd.,* p. 14.
[5] Cfr. Autor 2014 y 2016.

la posibilidad de que la consciencia sea inducida o despertada por un agente externo. Sin embargo, no constituyen un argumento central en el planteamiento de Penrose, por lo que no se estudiará en el presente trabajo.

Si el fenómeno de la consciencia surgiese de la interacción entre la computación y el entorno, se podría pensar que hay algo no computable en el entorno; pero, siendo algo viable, al ser *interiorizado* por el ordenador, se perdería su aspecto no-computacional. Para Penrose, el fenómeno de la consciencia no se debe ni a una computación analógica ni a una computación digital sino a algún aspecto no computacional de la realidad que todavía no ha sido desvelado por la física conocida[6].

En síntesis, aquellos aspectos matemáticos de computabilidad dudosa, como la aleatoriedad o el entorno, serían algorítmicamente computables sin que se diese un cambio significativo. La pregunta sobre qué tipo de realidad podría ser radicalmente no-computable seguiría abierta[7].

2. ¿Dónde buscar lo no-computable?

Ahora bien, si se afirma la existencia de algo radicalmente no-computable ¿qué características tiene o dónde encontrarlo? Para buscar una respuesta, Penrose parte de una intuición, que explicaré al final del artículo; y acude a las matemáticas, con el objetivo de demostrar que el entendimiento humano (y por tanto la conciencia, sin la cual no se puede dar el entendimiento) no puede ser una actividad algorítmica. En su argumentación busca un puente entre el entendimiento y la computación; y lo encuentra en el pensamiento matemático. Da tres razones para buscar ahí.

En primer lugar, porque ese argumento responde a las propuestas de inteligencia artificial en su propio terreno y con su mismo lenguaje. En segundo lugar, porque solo desde dentro de las matemáticas se podría encontrar cierta demostración rigurosa de algún aspecto de esa actividad no-computacional. Y en tercer lugar porque la computabilidad tiene naturaleza matemática y solo desde las matemáticas se podría probar la existencia de algo no-computable.

No entro en las críticas a estos presupuestos, aunque, de modo análogo a lo que pretende Penrose, si se admite la existencia de una instancia superior donde se puede incluir todo el pensamiento matemático, desde esa instancia se podría argumentar que la consciencia puede ser no-matemática. Esa instancia superior sería el pensamiento en sí.

---

[6] Cfr. Penrose, R., *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford 1994, pp. 25-26.

[7] Un ejemplo de problema no-computable es el *teselado aperiódico* descubierto por Penrose. Ningún ordenador podría haber encontrado la solución a pesar de ser un problema bien definido. Solo el pensamiento humano era capaz de hallar la respuesta. Cfr. Penrose, R., *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford 1994, pp. 29-33.

Su argumento consiste en mostrar que cuando se realizan juicios matemáticos conscientes sobre la verdad de algunas proposiciones matemáticas bien formuladas sucede algo no algorítmico. En otros términos, existen clases de problemas matemáticos, bien formulados y con solución, que no encuentran respuesta a nivel meramente computacional. Esto se debe, según Penrose, a que los ordenadores no poseen la cualidad humana del entendimiento y no pueden apreciar el contenido de las pruebas matemáticas.

3. Un argumento a tres niveles

En su crítica se pueden distinguir tres niveles entrelazados. El primero centrado sobre el programa de Hilbert y el formalismo matemático, el segundo que desciende al problema computacional, y el tercero que extrae conclusiones sobre el entendimiento y la conciencia.

En el primer nivel, el del formalismo de Hilbert y los teoremas de incompletitud de Gödel, el argumento afecta al conjunto de las matemáticas. Hilbert buscaba un sistema formal matemático bien definido, mediante un conjunto suficientemente amplio de axiomas autoevidentes y de reglas de inferencia matemática que incorporase *todas* las formas de razonamiento correcto. Si ese sistema formal fuese *completo* se podría decidir la verdad o falsedad de cualquier proposición matemática sintácticamente correcta, formulada dentro del sistema. Esa área de las matemáticas estaría libre de contradicción.

Lejos de confirmar el programa de Hilbert, Gödel y sus teoremas de incompletitud mostraron la existencia de enunciados verdaderos que van más allá del alcance de un sistema formal determinado, como la aritmética[8]. En otras palabras, no existe un sistema suficientemente complejo dentro del cual sea posible demostrar o refutar toda proposición bien definida sin caer en una contradicción.

En cualquier sistema de axiomas y reglas suficientemente amplio y bien definido se pueden encontrar familias de proposiciones matemáticas indecidibles. De tal modo que desde dentro del sistema no se puede afirmar la verdad o falsedad de dichas proposiciones, aunque sí se pueda desde instancias externas. Eso implica que el entendimiento humano, incluso si se limita a los enunciados matemáticos, no puede ser encapsulado en ningún sistema formal[9]. Al aceptar el sistema, lo trasciende.

---

[8] Para el caso particular de la aritmética de Peano (PA), el primer teorema se enunciaría así: Si PA es consistente, entonces existe un enunciado G tal que ni él ni su negación son demostrables en PA. Y el segundo: Si PA es consistente, entonces el enunciado que representa en PA la consistencia de PA no es demostrable en PA.

[9] Penrose, R., «Can a computer understand?», en: Rose, Steven (ed.), *From brains to consciousness? Essays on the New Sciences of the Mind*, Allen Lane, London 1998, pp. 157-158.

En un segundo nivel, el programa de Hilbert se plasma computacionalmente en el *problema de decisión*[10] y los teoremas de incompletitud de Gödel encuentran su paralelo en varios enfoques análogos que se desarrollaron casi a la vez, como la máquina de propósito general de Alan Turing o el *cálculo lambda* de Alonzo Church. Ambos autores demostraron que el argumento de Gödel se aplica a cualquier sistema de reglas que puedan ser programadas en una máquina idealizada de propósito general y llegaron a la conclusión de que el problema de decisión no tenía respuesta. Penrose estudia la cuestión para el caso de la máquina de Turing, donde el problema de decisión se reformula en el *problema de parada*[11].

Una máquina universal de Turing sería capaz de ejecutar sin error cualquier tipo de algoritmo[12]. Siempre daría una respuesta válida o, de no haber respuesta, continuaría trabajando sin bloquearse. Ahora bien, ¿esta máquina podría *decidir* si para cierto algoritmo habrá alguna ejecución en la que no se va a detener porque no tiene respuesta válida? Es decir, ¿existe algún algoritmo que programado en la máquina podría decidir sin error qué algoritmos van a encontrar siempre respuesta y cuáles no se van a detener nunca[13]? Turing concluye que no existe tal algoritmo de decisión.

En este segundo nivel el argumento matemático de Gödel contra el formalismo de Hilbert se convierte en un argumento computacional contra la posibilidad de encontrar un algoritmo que dé una respuesta al problema de parada. La pregunta por la verdad o falsedad de una proposición bien definida se desplaza a la pregunta sobre si una máquina universal de Turing se parará cuando actúe sobre

---

[10] En computación un problema de decisión es una cuestión en un sistema formal que tiene una respuesta «sí» o «no», dependiendo de los valores de algunos parámetros de entrada. Por ejemplo, «dado x, ¿x es un número primo?» es un problema de decisión. La respuesta será «sí» o «no» en función del valor de x. Una instancia de este problema es: «¿15 es un número primo?».

A su vez los problemas de decisión están estrechamente relacionados con la decibilidad matemática, es decir, si existe un método eficaz para determinar si un objeto matemático existe o si pertenence a un conjunto. De modo generalizado la decibilidad implica buscar «si existe un método que permita decidir sobre cualquier problema matemático». Es decir, resolver la cuestión sobre si un problema matemático tiene solución o no, conjugando simplemente axiomas y teoremas.

Por último, un método para la resolución de un problema de decisión dado en forma de algoritmo, se llama procedimiento de decisión.

[11] Una formulación del problema de parada puede ser: «¿Existe una función computable H(x,y) que permita determinar si la x-ésima función computable f_x finaliza arrojando un resultado cuando computa el input y?»

[12] Por algoritmo se puede entender todo procedimiento que puede ser realizado por una máquina de Turing. En su estructura más básica consiste en ejecutar unas instrucciones paso a paso a partir de unas reglas completamente especificadas. Estos pasos permiten la evolución desde un estado interno de las variables del sistema a otro, hasta que se llega a una instrucción concreta en la que se detiene. Después de mostrar la respuesta la máquina queda lista para ejecutar otra operación.

[13] Por ejemplo, no se sabe si un algoritmo con la conjetura de Goldbach encontrará siempre respuestas válidas. Un enunciado de dicha conjetura es: «Todo número par mayor que 2 puede escribirse como suma de dos números primos».

la enésima entrada-proposición; y la existencia de proposiciones matemáticas gödelianas indecidibles encuentra su paralelo en la imposibilidad de determinar si la máquina de Turing se parará ante la enésima entrada-proposición.

Sobre lo expuesto hasta ahora y en un tercer nivel, aparece el *nuevo argumento de Penrose*. Dicho argumento toma elementos de Gödel, Turing y Lucas para obtener conclusiones sobre el entendimiento humano y la inteligencia artificial:

> «The argument I shall present in the next chapter provides what I believe to be a very clear-cut argument for a non-computational ingredient in our conscious thinking. This depends upon a simple form of the famous and powerful theorem of mathematical logic, due to the great Czech-born logician Kurt Gödel. I shall need only a very simplified form of this argument, requiring only very little mathematics (where I also borrow from an important later idea due to Alan Turing). Any reasonably dedicated reader should find no great difficulty in following it. However Gödel-type arguments, used in this kind of way, have sometimes been vigorously disputed. Consequently, some readers might have gained an impression that this argument from Gödel's theorem has been fully refuted. I should make it clear that this is not so. It is true that many counter-arguments have been put forward over the years. Many of these were aimed at a pioneering earlier argument —in favour of mentalism [D] and opposed to physicalism [C]— that had been advanced by the Oxford philosopher John Lucas (1961). Lucas had argued from the Gödel theorem that mental faculties must indeed lie beyond what can be achieved computationally (others, such as Nagel and Newmann (1958), had previously argued in a similar vein). My own argument, though following similar lines, is presented somewhat differently from that of Lucas —and not necessarily as support for mentalism [D]. I believe that my form of presentation is better able to withstand the different criticisms that have been raised against the Lucas argument, and to show up their various inadequacies»[14].

A partir de ahora me centraré en la versión más sencilla del argumento. El procedimiento será por *reductio ad absurdum* y se apoyará en el *corte diagonal de Cantor*, tal y como hizo Turing para demostrar que el problema de parada no tenía solución.

4. El nuevo argumento de Penrose

Si C(n) es una computación[15] (p.ej.: ¿*n* es un número primo?) que se aplica repetidamente y por separado para cada número natural n, la computación

[14] Penrose, R., *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford 1994, p. 49.

[15] Penrose pone dos ejemplos prácticos: C(n) puede ser «encuentra un número que no sea la suma de n números cuadrados»; o bien «encuentra un número impar que sea la suma de n números pares». En el segundo caso, la respuesta es que la máquina nunca se parará, sea cual sea el valor de n. Y en el primer caso la máquina solo se parará cuando n sea 0, 1, 2 o 3, dando como resultado 1, 2, 3 y 7. Sin embargo, para probar que no se parará en ningún caso más hace falta una formidable demostración matemática.

C(n) en la máquina de Turing ¿dará una respuesta o seguirá computando porque no la encuentra? La máquina, ¿se parará o no?

Para responder a esta pregunta, que se encuentra en un segundo nivel respecto a C(n), se realiza otro procedimiento computacional: el algoritmo A. A contiene todo lo necesario para demostrar convincentemente y sin error que C(n) no se va a detener. Es decir, A es matemáticamente *consistente*[16]. Esto significa que: «Si A se detiene, entonces C(n) no se detiene».

Ahora generalizamos A para que se pueda ejecutar sobre otras computaciones $C_1(n), C_2(n), C_3(n)\ldots$ y se denomina A(q, n) a la ejecución de A sobre la computación $C_q(n)$[17]. De modo que: «Si A(q, n) se detiene, entonces $C_q(n)$ no se detiene».

Sobre esta afirmación, Penrose aplica el *corte diagonal de Cantor* —como hizo Turing— y considera el caso en que q=n, para concluir que «Si A(n, n) se detiene, entonces $C_n(n)$ no se detiene».

Como A(n, n) depende de una sola variable y es un procedimiento computacional, al igual que C, se puede renombrar A(n, n) como C(n). Para no confundirla con el resto de C(n), elegimos que A(n, n) sea la computación k-ésima: «$A(n, n)=C_k(n)$».

Sobre esta igualdad, como también hace el corte diagonal de Cantor, se examina un caso concreto en el que n=k. Así se obtiene: «$A(k, k) = C_k(k)$».

De modo que la afirmación «Si A(n, n) se detiene, entonces $C_n(n)$ no se detiene» se convierte en: «Si A(k, k) se detiene, entonces $C_k(k)$ no se detiene».

Pero como $A(k, k)=C_k(k)$, entonces resulta que: «Si $C_k(k)$ se detiene, entonces $C_k(k)$ no se detiene». Luego $C_k(k)$ *no se detiene*. Lo que significa que A(k, k) tampoco se detiene. Por tanto, en este caso concreto, el procedimiento A es incapaz de determinar que $C_k(k)$ no se detiene cuando de hecho *ya sabemos que no se detiene*.

En resumen, como A es consistente, A contiene todo lo necesario para determinar sin error que $C_k(k)$ no se detiene. Pero hay un caso concreto en el que

---

[16] Penrose utiliza el término sólido [*sound*] para referirse a la noción de ω-consistencia, la cual es más fuerte que la noción de consistencia [*soundness*]. Cfr. Alonso, E., «Mentalismo, mecanicismo: el nuevo argumento de Penrose», en: *Revista de filosofía* 26, 2001, pp. 158. De hecho el autor reconoce la relación entre ambos términos: «Another advantage is that the notion of «soundness» of a system F, when this notion is restricted to F's ability to establish Π_1-sentences, is equivalent to F's *consistency*», Penrose, R., «On understanding understanding», en: *International Studies in the Philosophy of Science* 11 (1), 1997, p. 10. La apreciación es debida a Hilbert. Cfr. Feferman, S., «Penrose's Gödelian argument», en: *Psyche* 2, 1995, pp. 249-256.

[17] Es decir, hay dos algoritmos. El primero se llama C (de computación) que es el que se detendrá o no al resolver un problema. C no falla: si se detiene, ha resuelto el problema; si no se para, el problema no tiene solución. El segundo algoritmo se llama A (de algoritmo) y actúa sobre C. Contiene todos mecanismos posibles para realizar bien su tarea: indicar cuando C no se va a detener. Si A se detiene es que en algún caso C no se va a detener y si A no se para es que C va a encontrar siempre una solución. Ambos mecanismos se ejecutan en un ordenador idealizado, que nunca falla. Dicho ordenador podría estar constituido por miles de ordenadores idealizados, pero físicamente distintos, que al estar interconectados funcionarían como un único ordenador idealizado.

A no puede determinarlo, aunque desde fuera sabemos que $C_k(k)$ no se detiene. Dicho con otras palabras, «la consistencia de A implica que A es incompleto».

Por tanto, nosotros sabemos algo que A no sabe y que debería saber porque ha sido programado con todas las herramientas matemáticas necesarias. Luego la conclusión es que el entendimiento humano no se puede contener en A[18].

Ningún conjunto consistente y cognoscible de reglas computacionales será suficiente para determinar qué computaciones no se paran. El algoritmo A no puede ser una formalización de los procedimientos que los matemáticos siguen para determinar si una computación no se detiene. No se puede encontrar el algoritmo A.

El nuevo argumento de Penrose concluye: «Los matemáticos humanos no están usando un algoritmo consistente cognoscible para determinar la verdad matemática»[19].

5. La respuesta de Gaifman a los teoremas de incompletitud

También otros autores han desarrollado argumentos de tipo gödeliano para sostener la imposibilidad de simular computacionalmente el razonamiento humano. Haim Gaifman[20] analizó varios de dichos artículos, concluyendo que, si un ordenador pudiese, de hecho, simular todo nuestro razonamiento matemático, entonces no podríamos entender completamente cómo trabaja.

Lo que pone en duda es cómo probar la *consistencia matemática* del matemático; o mejor dicho, del algoritmo que haría las veces de un matemático. En el fondo aplica los teoremas de incompletitud de Gödel al hombre o máquina que le replica, dando por descontado que su pensamiento es algorítmico.

Efectivamente, en cualquier sistema formal bien construido, conforme al segundo teorema de Gödel, la consistencia se puede expresar, pero no se puede probar dentro del sistema[21]. Un algoritmo que prueba teoremas genera pruebas

---

[18] En este argumento hay varios puntos que podrían sembrar la duda. Penrose analiza en *Shadows of the Mind* veinte contraargumentos para los que da respuesta y termina reafirmando su tesis.

[19] Penrose, R., *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford 1994, p. 76: «Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth».

[20] Gaifman, H., «What Gödel's Incompleteness Result Does and Does Not Show» en: *The Journal of Philosophy* 97 (8), 2000, pp. 462-470.

[21] Según el primer teorema, *toda teoría aritmética recursiva que sea consistente es incompleta*. Luego, no existe teoría matemática formal alguna, que sea a la vez consistente y completa, capaz de describir los números naturales y la aritmética con suficiente expresividad. Lo que implica que si no hay contradicción entre los axiomas de dicha teoría, entonces existen enunciados que no pueden probarse ni refutarse.

Conforme al segundo teorema *para toda teoría aritmética recursiva consistente,* Consis T *no es un teorema*. Es un caso particular del primer teorema, para el que una de las sentencias indecidibles de la teoría es la que «afirma» la consistencia de la misma: si el sistema es consistente, no se puede probar dentro del propio sistema.

dentro del sistema formal, pero si el algoritmo solo puede saber lo que puede probar, entonces no puede conocer que es consistente porque no lo puede probar.

La apelación a la consistencia se desplaza a un nivel superior; y eso se aplicaría a la máquina que replica el pensamiento humano, como muestra Gaifman. Desde un nivel superior semántico, sí se puede probar la consistencia de un nivel inferior, pero no desde el mismo nivel.

Para Penrose todos estos supuestos y otros similares se englobarían dentro de un superordenador idealizado, capaz de ejecutar un algoritmo que contuviese todos los razonamientos matemáticos. Sin embargo, no escapa a la crítica de Gaifman. La cuestión se desplaza a si soy intrínsecamente capaz de demostrar la consistencia de mi pensamiento matemático; o más bien la admito *a priori*, pero no puedo probarla.

Penrose acepta *a priori* que el algoritmo *probador de teorías* es consistente, pero esa consistencia la probaría, desde fuera, un ser humano; que a su vez podría ser una máquina cuya consistencia no puede probar. Al final, como fundamento último queda una regresión *ad infinitum* o la existencia de un mundo platónico de las ideas matemáticas, como postula Penrose. En ambos casos no se puede evadir la hipótesis de que los seres humanos sean máquinas[22]. La disputa acaba en tablas, debido al punto de partida que se asume.

6. Recursividad y reflexión

Sin embargo, no hay razones suficientes para adoptar esa postura, ni el punto de partida de Penrose. Más bien, lo lógico es pensar que no hay una representación exacta del pensamiento matemático, que ningún sistema algorítmico puede representarlo fielmente, que no existe tal consistencia, tampoco en el ser humano, cuando el sistema es suficientemente complejo. Veámoslo desde la reflexión sobre la matemática.

En un primer paso, el matemático reflexiona sobre su pensamiento matemático y se plantea que pueda ser formalizado íntegramente. En un segundo paso se infiere que dicho sistema formalizado podría ser consistente. Lo que implicaría una especie de reflexión, si el sistema se diese cuenta.

Sin embargo, Gödel muestra que la reflexión no puede abarcar todo el pensamiento matemático porque no puede probar su consistencia dentro del sistema. Luego, el sistema solo podría reflejar parcialmente el pensamiento matemático, porque el acto reflexivo queda fuera de la *comprensión* que el sistema tiene de sí.

Las conclusiones de los teoremas de incompletitud de Gödel dan argumentos para admitir que una plena comprensión de nuestro propio razonamiento matemático es imposible de alcanzar. Podemos confirmar aspectos de cómo

[22] Gaifman, o.c.

funciona, pero no podemos tener una teoría completa y detallada. La razón de la imposibilidad se puede encontrar en que la reflexión introduce al *sujeto* en la teoría.

A mi entender, el *nuevo argumento de Penrose* se desenvuelve en dos niveles de razonamiento: el nivel del algoritmo en sí (algoritmo A) y nivel del entendimiento humano consciente. Del contraste entre ambos niveles, se infiere la diferencia. El entendimiento humano razonaría de un modo no-algorítmico porque no sería asimilable al proceso algorítmico de un ordenador[23].

La diferencia no estriba en la ejecución de procesos o en la elaboración de pruebas, sino en la capacidad de entender del sujeto consciente. Ese es el paso que no puede dar un ordenador. Hay algo esencial en el proceso de aprendizaje del entendimiento humano que no es computacionalmente asimilable.

En el argumento se podría distinguir entre una auto-referencia computacional o recursividad y una auto-referencia humana o reflexividad. Penrose distingue entre ambos y esto es esencial, ya que la reflexividad consciente, realiza el juicio y formula la conclusión del argumento matemático: esta reflexividad humana sería no-algorítmica. Mientras la actividad matemática está inmersa en su quehacer matemático, el matemático puede trascenderlo. El argumento no es consciente, Penrose sí.

Con todo esto, diríamos que el argumento de Penrose, a nivel matemático, no concluye. Tanto él como otros autores[24] reconocen la existencia de fisuras argumentativas. Fisuras que, si bien para Penrose, no son suficientes para contrarrestar la solidez de su argumento, sí lo son para rebatir su argumento, ya que no es *matemáticamente* concluyente.

La pretensión inicial de hacer un argumento inatacable, lo autolimita; aunque el razonamiento final puede ser válido, si se juzga desde una instancia superior como el entendimiento humano. En realidad, el argumento acaba siendo filosófico, porque desde las matemáticas no se pueden extraer las conclusiones que Penrose extrae. Sino que, mediante la reflexividad, el entendimiento humano

---

[23] Como el fenómeno de entender (understanding) requiere de la consciencia (consciousness), según Penrose, también el fenómeno de la consciencia tendría un aspecto no algorítmico. Pero ese aspecto no estaría sólo en la dimensión pasiva (awareness) de la consciencia, sino también en su dimensión activa (free will). Según Penrose, con la física conocida, la consciencia (consciousness) tanto en su dimensión pasiva (awareness) como activa (free will) no podría darse en un ordenador, sino que entendimiento, consciencia y libertad serían atributos meramente humanos. Al programar un ordenador, son los seres humanos quienes determinan qué algoritmo es el que se debe ejecutar en cada circunstancia. Ellos proporcionan las reglas a seguir y analizan los resultados. Cfr. Penrose, R., *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford 1994, p. 199.

[24] Cfr. Lindström, P., «Remarks on Penrose's "New Argument"», en: *Journal of Philosophical Logic* 35 (3), 2006, pp. 231-237; Alonso, E., *Ibíd.*, «Mentalismo, mecanicismo: el nuevo argumento de Penrose», en: *Revista de filosofía* 26, 2001, p. 158; Shapiro, Stewart, «Mechanism, Truth, and Penrose's New Argument», en: *Journal of Philosophical Logic* 32 (1), 2003, pp. 19-42.

se compara a sí mismo con la recursividad computacional para concluir que no son lo mismo: el matemático con su pensamiento humano consciente juzga tanto la reflexividad propia como la recursividad del algoritmo.

Según Penrose, lo que diferencia recursividad de reflexividad es el carácter *no-algorítmico* de esta última. Por contraste, para la mayoría de los defensores de la inteligencia artificial ambos procesos serían auto-referenciales, por lo que no existiría una diferencia sustancial entre reflexividad y recursividad[25].

Conclusiones

En resumen, desde el punto de vista que se ha presentado, la reflexividad humana consciente posee una apertura esencial para la que no puede existir un telón de fondo intelectual. Esto permite que el pensamiento humano se juzgue y juzgue otras cosas, de modo que crezca su autoconocimiento y su conocimiento de la realidad de modo ilimitado.

Frente a este modo de conocer, la computación podrá crecer mediante recursividad y complejidad relacional, pero siempre en la medida en que participa de la apertura del pensamiento humano; es decir, en la medida en que el ser humano *introduce* nuevos elementos. El crecimiento de los algoritmos, según Penrose, se produce por la contribución de los seres humanos con el descubrimiento de patrones y técnicas de computación; y el hecho de que un algoritmo crezca en complejidad o en interacciones no le va a hacer consciente.

La conclusión del argumento de Penrose, en cuanto filosófica, parece válida, aunque no posee la certeza matemática que quería darle su autor. Se une de este modo a otros razonamientos de índole filosófica para negar que lo que hace un cerebro humano sea esencialmente igual a lo que hace un ordenador. Aun así, sostener una postura distinta, sea mecanicista o mentalista, sigue siendo posible dentro del marco establecido por Penrose.

Su argumento es sugerente y debe ser tenido en cuenta, pero su pretensión de inatacabilidad es errónea, porque pivota sobre la premisa de la consistencia. En última instancia, pienso que dicha consistencia ni siquiera es asumible como verdadera. Quienes la asumen, aunque piensen que no es demostrable, caen con facilidad o en el platonismo matemático o en la *matematización* de la realidad. A mi parecer se trata de una conclusión filosófica equivocada, en la que se da un salto ilícito debido a un *a priori* cientificista.

En ese intento se va más allá de lo que las matemáticas pueden decir de sí mismas, por lo que no concluye. Esto se podría deber a que las matemáticas no se fundamentan en sí mismas, sino en la actividad del pensamiento humano y en su entrelazamiento con la realidad. Dichos entrelazamientos hacen que las

[25] Hofstadter, D. R., *Yo soy un extraño bucle*. Tusquets, Barcelona 2008, considera la percepción del yo como un sujeto *creado* (no real) fundamentado en una auto-referencia infinita a partir de las percepciones.

matemáticas sean inabarcables en cuanto alcanzan un mínimo de complejidad, como sucede con la computación matemática, porque son sistemas abiertos.

El punto de partida implícito de Penrose es suponer que toda la realidad es computable; y cuando descubre que no lo es, busca lo no computable. Personalmente me inclino por un punto de partida que afirme la existencia de una realidad a la que llegamos desde nuestra experiencia. Después esa realidad se estudia, se observa o se analiza desde las ciencias, las matemáticas o la computación; pero, al estudiarla así, se deja fuera toda realidad no-científica, no-matemática o no-computacional.

Esta consideración lleva a afirmar que siempre habrá mucho contenido de realidad dentro de esas «no-características». Es más, la misma realidad acaba trasciendo la aproximación reductiva para mostrarse abierta[26].

La no-localidad

Esto conecta también con la visión de «no-localidad» de la conciencia que tiene Penrose. Una visión que deriva del problema del teselado aperiódico. Pero vamos por partes, para que el giro argumentativo de este *excursus* final que enlaza con la intuición de la que se habló al principio, no sea muy brusco. Comencemos por el problema.

Supongamos que queremos cubrir una superficie con baldosas. El modo normal de hacerlo es mediante un patrón que se repite cada poco. Es un patrón que se observa localmente. Ahora bien, y este es el problema, ¿sería posible cubrir esa superficie con un número finito de piezas de tal modo que no exista un patrón de repetición?

Es lo que se conoce como cuasi-simetría y eso es lo que se propuso Penrose: comprobar si una superficie se podía cubrir con un conjunto finito de teselas, de modo que no existiese un patrón de repetición[27]. Encontró ese conjunto de teselas y las redujo a dos. Bastan dos simples piezas con unas características muy determinadas para cubrir aperiódicamente una superficie. De este modo,

[26] En relación a esta problemática hay muchos artículos interesantes en los números monográficos de *Scientia et Fides* 5/2 (2017) y *Naturaleza y libertad* 7 (2016), pero sobre todo cabe resaltar el trabajo realizado por Juan Arana en *La Conciencia Inexplicada*. Cfr. Acosta, M., «¿Es la matemática la *nomogonía* de la conciencia? Reflexiones acerca de la conciencia y el platonismo matemático de Penrose», en: *Naturaleza y Libertad* 7, 2019, pp. 15-39; Arana, J., *La conciencia inexplicada*, Biblioteca Nueva, Madrid 2017; De Haan, D., «Hylomorphic Animalism, Emergentism, and the Challenge of the New Mechanist Philosophy of Neuroscience», en: *Scientia et Fides* 5 (2), 2017, pp. 9-38; Lombardi, A., «Dan Zahavi and John Searle on Consciousness and Non-Reductive Materialism», en: *Scientia et Fides* 5 (2), 2017, pp. 155-170.

[27] Kruglinski, S., «The discover Interview: Roger Penrose», en: *Discover* 30 (8), 2009, p. 56: «My interest in the tiles has to do with the idea of a universe controlled by very simple forces, even though we see complications all over the place. The tilings follow conventional rules to make complicated patterns. It was an attempt to see how the complicated could be satisfied by very simple rules that reflect what we see in the world».

Penrose descubrió la existencia de una regla que determina de modo «no local» cómo colocar esas piezas.

Además, posteriormente Roger Berger comprobó que no podía existir ningún método informático capaz de simular la evolución del sistema, porque no había un algoritmo capaz de decidir si un conjunto determinado de teselas iba a cubrir una superficie concreta. Lo que Penrose descubrió no lo podía descubrir un algoritmo. Su pensamiento, al menos es este aspecto, no era computable.

Años después, en 1984, se descubrió contra pronóstico la existencia de *cuasi-cristales* aperiódicos que seguían patrones como los de Penrose. Esto confirmaba la existencia de realidades regidas por patrones no-computacionales y no-locales[28]. En la naturaleza se observaba un comportamiento ordenado, con una base matemática sencilla, descubierto mediante cierta intuición fundamentalmente inaccesible al tratamiento computacional.

En otras palabras, hay realidades gobernadas por reglas no locales que escapan a lo que un ordenador idealizado puede descubrir. La *no-localidad no-computable* apunta a la existencia de cierto orden en un ámbito superior que, sin embargo, no se aprecia en una observación meramente local[29].

De aquí que cuando Penrose hable de la conciencia, parta de estos presupuestos. La entenderá como algo que tiene estas no-características y dedicará tiempo y esfuerzo a intentar encontrar indicios de la acción de la conciencia en el cerebro[30]. Pero este tema es para una discusión posterior.

Bibliografía

Acosta, M. (2019). «¿Es la matemática la *nomogonía* de la conciencia? Reflexiones acerca de la conciencia y el platonismo matemático de Penrose», en: *Naturaleza y Libertad* 7, pp. 15-39.

Alonso, E. (2001). «Mentalismo, mecanicismo: el nuevo argumento de Penrose», en: *Revista de filosofía* 26, pp. 139-164.

Arana, J. (2017). *La conciencia inexplicada*. Madrid: Biblioteca Nueva.

De Haan, D. (2017). «Hylomorphic Animalism, Emergentism, and the Challenge of the New Mechanist Philosophy of Neuroscience», en: *Scientia et Fides* 5 (2), pp. 9-38. DOI: 10.12775/SetF.2017.025

Feferman, S. (1995). «Penrose's Gödelian argument», en: *Psyche* 2, pp. 249-256.

Gaifman, H. (2000). «What Gödel's Incompleteness Result Does and Does Not Show» en: *The journal of philosophy* 97 (8), pp. 462-470.

[28] Cfr. Steinhardt, P. J. «New perspectives on forbidden symmetries, quasicrystals, and Penrose's tilings», en: *PNAS* 93 (25), 1996, pp. 14267-14270.

[29] Cfr. Autor 2017.

[30] Cfr. Penrose, R., «Author's response – The nonalgorithmic mind», en: *Behavioral and Brain Sciences* 13 (4), 1990, pp. 692-705; Penrose, R., «Can a computer understand?», en: Rose, Steven (ed.), *From brains to consciousness? Essays on the New Sciences of the Mind*, Allen Lane, London 1998, pp. 154-179.

Hofstadter, D. R. (2008). *Yo soy un extraño bucle*. Barcelona: Tusquets.
Kruglinski, S. (2009). «The discover Interview: Roger Penrose», en: *Discover* 30 (8), pp. 54-57.
Lindström, P. (2001). «Penrose's New Argument», en: *Journal of Philosophical Logic* 30 (3), pp. 241-250.
Lindström, P. (2006). «Remarks on Penrose's "New Argument"», en: *Journal of Philosophical Logic* 35 (3), pp. 231-237.
Lombardi, A. (2017). «Dan Zahavi and John Searle on Consciousness and Non-Reductive Materialism», en: *Scientia et Fides* 5 (2), pp. 155-170. DOI: 10.12775/SetF.2017.020
Lucas, J. R. (1961). «Minds, Machines and Gödel», en: *Philosophy* 36, pp. 112-127.
Nagel, E. y Newman, J. R. (1958). *Gödel's Proof*. New York: New York University Press.
Penrose, R. (1989). *The emperor's new mind: concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press.
Penrose, R. (1990). «Author's response – The nonalgorithmic mind», en: *Behavioral and Brain Sciences* 13 (4), pp. 692-705.
Penrose, R. (1994). *Shadows of the mind: a search for the missing science of consciousness*. Oxford: Oxford University Press, Oxford.
Penrose, R. (1997). «The need for a non-computational extension of quantum action in the brain», en: Arhem, Peter, Liljenstrom, Hans y Svedin, Uno (eds.), *Matter matters? On the material basis of the cognitive activity of mind*, Springer, Berlin, pp. 11-27.
Penrose, R. (1997). «On understanding understanding», en: *International Studies in the Philosophy of Science* 11 (1), pp. 7-20.
Penrose, R. (1998). «Can a computer understand?», en: Rose, Steven (ed.), *From brains to consciousness? Essays on the New Sciences of the Mind*, Allen Lane, London, pp. 154-179.
Shapiro, S. (2003). «Mechanism, Truth, and Penrose's New Argument», en: *Journal of Philosophical Logic* 32 (1), pp. 19-42.
Steinhardt, P. J. (1996). «New perspectives on forbidden symmetries, quasicrystals, and Penrose's tilings», en: *PNAS* 93 (25), pp. 14267-14270.

Universidad de Navarra
reherce@unav.es

RUBÉN HERCE

[Artículo aprobado para publicar en febrero de 2021]