LA IMPORTANCIA DEL TEOREMA GÖDELIANO EN EL PENSAMIENTO DE ROGER PENROSE

Daniel Heredia González
Universidad de Sevilla

Resumen: La de Roger Penrose es una postura filosófica bien conocida, en concreto, la referida al debate de la computabilidad o no computabilidad de mente y la consciencia humana. En su primera exposición en *La nueva mente del emperador* encontramos los argumentos en contra de la IA *fuerte*. Pero solo en *Las sombras de la mente* se encuentra la base más firme de su posicionamiento, ya que allí realiza un análisis más profundo del teorema de Gödel, que es clave para entender su rechazo de la computabilidad como respuesta al citado debate.
Palabras clave: Teorema de Gödel, incompletitud, reflexión, máquina de Turing.
The importance of the Gödelian theorem in the thought of Roger Penrose
Abstract: Roger Penrose's position is a well-known one in philosophical matters, specifically the one referred to the debate of computability or non-computability of mind and human consciousness. In his first exposition in *The Emperor's New Mind* we find the arguments against Strong AI. But it is not until *Shadows of the mind* where (in my opinion) the strongest basis of his positioning is found, because it carries out a deeper analysis of the acceptance of Gödel's theorem, which is the key to understanding the rejection of computability in response to the aforementioned debate.
Keywords: Gödel's theorem, incompleteness, reflection, Turing's machine.
Recibido: 17/09/2019 **Aprobado:** 20/02/2019

1. El teorema de Gödel I: Su contexto

El teorema de Gödel en sí mismo no pretendía contribuir al debate de la computabilidad o no computabilidad de la mente y la conciencia humana. Si atendemos fielmente a su formulación vemos que se trata de una respuesta lógica a un problema estrictamente lógico. Pero, ¿de veras estamos

ante un asunto que sólo tiene cabida en dicho ámbito? Roger Penrose sostendrá (y yo me sumo a él) que no es así y que tal teorema viene a ser un punto crucial en contra de aquellos que sostienen la Inteligencia Artificial *fuerte*.

Antes de examinar las implicaciones del teorema de Gödel conviene contextualizarlo, porque ello ayuda a comprender mejor que surgió como la respuesta que el matemático de origen checo ofreció a los distintos movimientos que protagonizaban la filosofía de las matemáticas de su época.

De acuerdo con el estudio de Jean Ladrière (1921-2007), las tres principales posturas que ocupaban la escena matemática eran el *intuicionismo*, la *teoría de la demostración* y, sobre todo, el *formalismo*[1].

¿Qué defiende cada una de estas posturas? En primer lugar, los intuicionistas —corriente cuyo máximo exponente es L. E. J. Brouwer (1881-1966) y sus seguidores— defienden la tradición del número y de la discontinuidad (1969: 44). Para ellos lo matemático no existe por sí mismo, sino que depende de su construcción (constructibilidad en terminología de Ladrière). Su idea fundamental es que el desarrollo de la matemática depende de una intuición originaria, «la de la división de la unidad, fuente de la dualidad, siendo esencialmente la de la estructura del tiempo, que constituye la base de la noción de número entero» (1969: 44). Esta posición con respecto a los números enteros hace que los intuicionistas tengan una perspectiva del infinito distinta a la de los demás, en el sentido de que sólo admiten la naturaleza potencial de esta idea (1969: 44). Uno de las consecuencias más destacables de esta postura es la negación del principio del tercio excluso[2] en conjuntos infinitos. A pesar de que llegaron a tener una influencia notable en las matemáticas, el hecho de no admitir este principio

[1] Sigo esta clasificación de concepciones de fundamentos matemáticos y no la usual (intuicionismo, logicismo y formalismo), porque me he apoyado en Ladrière para la exposición de las ideas principales de esta sección de mi texto.

[2] Este principio defiende que es imposible que dada una disyuntiva con proposición A pueda darse a su vez su contraria [no-A]

fundamental de la lógica les relegó a un segundo plano dentro de la comunidad matemática (1969:45).

La segunda corriente es la denominada teoría de la demostración, liderada por el matemático alemán David Hilbert (1862-1943). La misión que se había autoimpuesto Hilbert era posibilitar la convivencia entre las matemáticas clásicas y la idea constructivista que los intuicionistas aportaban, cuya utilidad resultaba indiscutible. ¿Qué era lo que no compartía Hilbert con la corriente brouweriana? El matemático alemán concebía la matemática con la inclusión de los principios básicos de la lógica, como el de tercio excluso —que rechazaban los intuicionistas— o el de no-contradicción. La admisión de estos principios como parte sustancial de la matemática sitúa a Hilbert en el bando contrario al de Brouwer y sus seguidores, puesto que la no-contradicción presupone una existencia en sí del ente matemático con independencia de los procedimientos del pensamiento que nos ayudan a alcanzarlo (Ladrière, 1969: 45-46). Pero lo que pretende recoger Hilbert de la concepción intuicionista es la idea de constructibilidad, no en el sentido de que la existencia del ente matemático dependa de lo que se construye, sino en el de que la actividad constructora en matemáticas es totalmente imprescindible. La matemática necesita de la intuición para llegar a las soluciones, demostrando su no-contradicción. Visto de este modo, parece que la teoría de la demostración lograría romper barreras, aunque en realidad había conseguido el efecto contrario. Los criterios de esta teoría resultaron más restrictivos que los defendidos por el punto de vista brouweriano, ya que se no aceptaba los razonamientos de inducción completa. Esta forma de entender a los intuicionistas por parte de la escuela de Hilbert fue expuesta por el matemático francés Jacques Herbrand (1908-1931):

> Entendemos por razonamiento intuicionista el que satisface las siguientes condiciones: siempre se considera un número finito y determinado de objetos y funciones; estas están perfectamente definidas, permitiendo su definición calcular sus valores de manera unívoca; nunca se afirma la existencia de un objeto sin proporcionar el medio de construirlo; no se considera nunca el conjunto de

> todos los objetos *x* de un conjunto infinito; y cuando se dice que un razonamiento (o un teorema) es cierto para todos estos *x*, quiere decirse que para cada *x* considerado particularmente es posible repetir el razonamiento general en cuestión que sólo debe ser considerado como prototipo de estos razonamientos particulares (cit. por Ladrière, 1969: 46).

Más tarde, Hermann Weyl (1885-1955) demostraría que esto no acarrearía los resultados esperados, ya que conducía a una postura más prohibitiva que conciliadora (Ladriere, 1969: 46).

La última de las corrientes matemáticas es el formalismo. También hay formalismo en las dos teorías anteriores, porque tanto del intuicionismo como la teoría de la demostración buscan elaborar un sistema formal. La característica distintiva de los sistemas formales es que intentan acabar con toda discusión. En ellos la meta está claramente determinada (a través de reglas inquebrantables), y de apartarse de ella supone adentrarse en un camino que conduce necesariamente al error. Brouwer y Hilbert pretendían lograrlo de manera diferente. Sin embargo, los sistemas más representativos del formalismo son los propuestos en los *Principia Mathematica* por Alfred North Whitehead (1861-1947) y Bertrand Russell (1872-1970), de un lado; y la denominada axiomática de la teoría de conjuntos de la mano de Ernst Zermelo (1871-1953), Adolf Fraenkel (1891-1965) o John von Neumann (1903-1957) (Ladriere, 1969: 74). La aportación de estos sistemas formales convertía la matemática y la lógica en un fortín inexpugnable.

O eso era al menos lo que se pretendía. En 1931 un joven matemático publicó «Sobre proposiciones formalmente indecidibles de los *Principia Mathematica* y sistema relacionados», trabajo que hizo temblar los cimientos de estos sistemas formales[3]. Ese matemático era Kurt Gödel (1906-1978).

[3] Cabe destacar también el intento en este mismo sentido del matemático germano-suizo Paul Finsler, quien trabajó en la demostración de la existencia de proposiciones indecidibles. Aunque se rige por la misma noción que Gödel (la de derivación en vez de la de certeza o la de definición) es considerado, en comparación con el teorema del matemático austríaco, como un método más incompleto. Finsler no entra en considerar los sistemas propios sobre

¿Qué hizo tan dramática su aparición en escena? El hecho de poner en tela de juicio uno de los cimientos más firmes de los procedimientos dados en los sistemas formales. Se cifraba éste en asegurar la completitud de una teoría que fuese consistente (que no tuviera contradicciones). Su propuesta colocó al matemático[4] austriaco-estadounidense entre los pensadores más brillantes del momento. Y no sólo eso. Kurt Gödel sigue siendo considerado como «uno de los lógicos más grandes de todos los tiempos, ¡probablemente el que más!», en opinión de Penrose (2012: 197).

2. El teorema de Gödel II: Una breve mirada

El teorema de Gödel surge como contraposición a los diferentes sistemas formales, pero sobre todo a los de Russell y Whitehead[5]. En concreto, intenta dar respuesta a la pregunta de si estos sistemas [PM] son capaces de «decidir todas las cuestiones matemáticas que puedan ser formuladas en dichos sistemas» (Gödel, 2006: 54). A esta pregunta el matemático austríaco responde que no. La consecuencia más importante de hacerlo así es la posibilidad de poner en duda la infalibilidad de los sistemas formales.

Como la crítica de Gödel se refiere a los sistemas formales propios en PM, el matemático austríaco encuentra necesario manejar sus mismos conceptos[6]. Pero, ¿cuáles son esos conceptos? Sin entrar en todos los detalles, es necesario comprender los que son sustanciales para el tema que abordamos aquí.

los que se centraban sus críticas. Gödel, en cambio, se mete de lleno en ellas, teniendo como herramienta principal la aritmetización (Ladrière, 1969: 97), proceso que veremos en §2.

[4] Siendo injustos con él por encasillarlo de tal modo, ya que la profundidad filosófica de sus propuestas son indiscutibles.

[5] De aquí en adelante me referiré tanto a los sistemas propios como a la obra en sí misma como PM, tal y como lo hace el mismo Gödel (2006); aunque algunas veces también me referiré a los sistemas como LFG, tal como hace Ladrière.

[6] Al menos sí que respetaba la mayoría de ellos para no dar lugar a confusiones.

El primer concepto a considerar es el de *recursividad*. La recursividad es un término matemático que se define como la capacidad que tiene un procedimiento de definirse a sí mismo. Dentro de los procesos recursivos, los que interesan a Gödel para su teorema son las *funciones recursivas primitivas*[7]. Estas funciones son, a su vez, las que se definen a sí mismas mediante operaciones que usan la recursión y composición de funciones.

¿Cómo pretende Gödel demostrar la indecidibilidad de una proposición que pertenece a un sistema LFG? La clave se encuentra en la noción de *derivación*[8]. Lo que intenta obtener Gödel es una proposición que afirme su

[7] Gödel utiliza como funciones básicas para las funciones recursivas primitivas las siguientes:
- *la función sucesor Scc*;
- *la función constante Cnt*, cuyo valor es siempre 0 (*Cnt* $\alpha = 0$);
- *la función de selección Sel* $^{(in)}$, que despeja la i-ésima variable (*Sel* $^{(in)}$ $X_1 X_2 \ldots X_n = X_1$).

Una función recursiva primitiva es la que puede obtenerse partiendo de estas tres funciones básicas mediante las tres operaciones siguientes:

1) aplicación de una función a una serie de argumentos;

2) aplicación del *esquema de sustitución*:

$F X_1 X_2 \ldots X_n = G (H_1 X_1 X_2 \ldots X_n) (H_2 X_1 X_2 \ldots X_n) \ldots (H_m X_1 X_2 \ldots X_n)$, donde $G, H_1, H_2, \ldots H_m$ son funciones previamente definidas.

3) Aplicación del *esquema de recursión primitiva*.

Todas las funciones elementales de la Aritmética entran en esta clase (suma, diferencia, multiplicación, elevación a potencia, etc.) (Ladrière, 1969: 88).

Lo verdaderamente importante del teorema de Gödel con respecto a las demás posturas que tuvieran como meta la demostración de la existencia de indecidibilidad en algunas proposiciones, era la inclusión de la Aritmética en ello. A través de esta se puede profundizar en asuntos que son inalcanzables para la mayoría de los métodos que puedan utilizarse.

[8] Incluso los poco doctos en el campo de la lógica saben que este método consiste —diciéndolo mal y pronto— en llegar a una conclusión a través de premisas que responden a unas reglas lógicas determinadas, tipo:

1. $\neg (p \rightarrow r)$ (P1)
2. $p \rightarrow (q \rightarrow r)$ (P2)
3. p hipótesis
4. q hipótesis
5. $q \rightarrow r$ ($E \rightarrow$), 2, 3

propia inderivabilidad. Resulta esencial para su propósito que la proposición indecidible pertenezca al sistema LFG (Ladrière, 1969: 95). Así consigue en una crítica inmanente a los sistemas formales y no establecida fuera de ellos. Para introducir su proposición dentro de los sistemas formales es imprescindible realizar otro proceso: la *aritmetización* de dichos sistemas formales. El método de aritmetización utilizado por Gödel se aplica a un sistema formal concreto, pero, en realidad, su aplicación es susceptible a cualquiera de ellos (Ladrière, 1969: 97). Este procedimiento lo define Ladriere del siguiente modo:

> Para aritmetizar un sistema formal, se hace corresponder a cada uno de sus signos (componentes primitivas) un número entero determinado. Después se adopta un procedimiento que haga corresponder un número entero perfectamente determinado a toda serie de signos o expresiones de LF.
> Sea *Sy* una expresión de LF constituida por una sucesión de signos *k*, y sean n_1, n_2,…, n_k los números enteros que se hace corresponder a estos signos.
> El número correspondiente a *Sy* será:
> $2n_1 \times 3n_2 \times \ldots \times p^{nk}{}_{k,}$
> donde p_k representa el k-ésimo primer número diferente de 1.
> Sea igualmente *Sp* una serie de *k* expresiones de LF (por ejemplo una derivación) y sean n_1, n_2,… n_k los números correspondientes a estas expresiones.
> El número correspondiente a *Sp* será:
> $2n_1 \times 3n_2 \times \ldots \times p^{nk}{}_{k.}$
> En virtud del teorema sobre la unicidad de la descomposición de un número en sus factores primos, la correspondencia así establecida es perfectamente biunívoca: a todo elemento de LF (signo, expresión o serie de expresiones) corresponde un número entero, y sólo uno, y a un número entero corresponde un elemento de LF cuando más.

6. r (E $\rightarrow$), 4, 5
7. p $\wedge$ r (I $\wedge$), 3, 6
8. q $\rightarrow$ (p $\wedge$ r) (I $\rightarrow$), 4, 7

Donde 1 y 2 son las premisas, 3 y 4 son hipótesis y del 5 al 7 vemos la aplicación de las reglas (introducción de implicación - ($\rightarrow$) - y conjunción – ($\wedge$) -) para llegar a la conclusión 8 (problema de lógica proposicional tomado de Manzano Arjona, Huertas Sánchez, 2011: 164).

> Llamaremos a tal correspondencia *correspondencia de Gödel*, y al número entero correspondiente a un elemento *Ec* de LF por una *correspondencia de Gödel*, número de Gödel o número-G de *Ec* (Ladrière, 1969: 97-98).

Con este proceso de aritmetización, Gödel está tratando de indicar que la Aritmética que se está construyendo es también recursiva. Es decir, lo que puede demostrarse en ella responde a un número finito de pasos (anteriormente establecidos, claro está) a realizar[9]. Además del gran paso que en sí misma representa, la aritmetización sirve para sentar la *consistencia lógica*. ¿Qué es la consistencia lógica? Es la propiedad de un sistema formal que consiste en establecer la imposibilidad de la aceptar un sistema concreto y su contradicción al mismo tiempo. Esta noción posee gran trascendencia para el desarrollo de la solución gödeliana al problema de la indecidibilidad de los procesos algorítmicos.

Dejando a un lado los conceptos preliminares, pasemos al grueso del contenido del teorema que lleva el nombre del matemático austríaco.

Gödel quiere encontrar una proposición perteneciente a un sistema formal PM que se declare a sí misma como indecidible y, como consecuencia, que contenga la prueba de su indeducibilidad, pudiendo ser traducible a cualquier sistema formal. Ladrière divide esta demostración en dos etapas: la construcción de dicha proposición (que el autor belga denomina como J*[10]) y la demostración de su carácter indecidible. Con la división ladriereana se manifiesta que para contruir la proposición J* Gödel sigue los procedimientos propios de los sistemas formales PM[11]. A finde esquivar los

[9] En la obra de Ladrière (1969) se expone cada uno de los pasos lógicos que se siguen en el teorema gödeliano. Para no entretenerme en nombrarlos todos, y alejarme de ese modo de lo que pretendo en este punto, recomiendo véase para aclaración de dicho análisis las páginas 100-124 del libro citado.

[10] Lo cual comento porque me referiré a ella de tal modo más delante.

[11] De nuevo evito reproducir los pasos lógicos de Gödel (o los de Ladrière) para, de este modo, evitar desviarme demasiado. Las obras que he seguido son las anteriormente citadas (Gödel, 2006; Ladrière, 1969).

detalles lógicos, me serviré de la explicación que da Penrose tanto de la construcción como de la demostración:

> Hemos numerado todas las funciones proposicionales que dependen de una sola variable, de modo que la que acabamos de escribir debe tener asignado un número. Escribamos este número como *k*. Nuestra función proposicional es la k-ésima de la lista. Por consiguiente:
> $\neg\exists x\,[\prod_x \text{demuestra } P_w\,(w)] = P_k\,(k)$
> Examinaremos ahora esta función para el valor *w* particular: $w = k$. Tenemos:
> $\neg\exists x\,[\prod_x \text{demuestra } P_k\,(k)] = P_k\,(k)$.
> La proposición específica P_k (k) es un enunciado aritmético perfectamente bien definido (sintácticamente correcto). ¿Tiene una demostración dentro de nuestro sistema formal? ¿Tiene demostración se negación $\neg P_k$ (k)? La respuesta a ambas preguntas debe ser «no». Podemos verlo examinando el significado subyacente en el procedimiento de Gödel. Aunque P_k (k) es sólo una proposición aritmética, la hemos construido de modo que afirma lo que se ha escrito en el lado izquierdo: no existe demostración, dentro del sistema, de la proposición P_k (k)». Si hemos sido cuidadosos al establecer nuestros axiomas y reglas de inferencia, y suponiendo que hayamos hecho bien nuestra numeración, entonces no puede haber ninguna demostración de esta P_k (k) dentro del sistema. En efecto, si hubiera tal demostración, el significado del enunciado que P_k (k) realmente afirma, a saber, que no existe demostración, sería falso, de modo que P_k (k) tendría que ser falsa como proposición aritmética. Nuestro sistema formal no debería estar tan mal construido como para permitir que se demuestren proposiciones falsas. Por consiguiente, debe ser el caso que, de hecho, no hay demostración de P_k (k). Pero esto es precisamente lo que P_k (k) está tratando de decirnos. Por lo tanto, lo que afirma P_k (k) debe ser un enunciado verdadero, de modo que P_k (k) debe ser verdadera como proposición verdadera que no tiene demostración dentro del sistema (Penrose, 1996: 102).

¿Qué establece entonces la prueba gödeliana? A grandes rasgos podemos decir que, dada la proposición J* y siguiendo la aritmética recursiva de este proceso, resulta que es verdadera pero su demostración queda fuera del alcance del sistema formal en que nos movemos (PM). Por tanto, el

teorema (dada una proposición aritmética recursiva consistente, sólo demuestra su incompletitud) dicta que los sistemas formales resultan insuficientes tanto para probar como para refutar dicha sentencia. Si este sistema logra volver sobre sus pasos sólo lo podrá hacer al modo de círculo vicioso lógico (o dialelo). Es decir, las premisas, al ser cerradas (como requiere todo sistema formal) no permiten demostrar más allá de ellas mismas. No existe, de esta forma, un proceso de reflexión (volver a sí mismo, de un modo más profundo y no simplemente volver sobre los pasos dados) que sí se da en la consciencia humana. Este veredicto es el pilar fundamental de las implicaciones que Penrose extrae del teorema gödeliano, como veremos en el siguiente punto.

3. El teorema gödeliano como apoyo a la postura penroseana

La lógica, sus sistemas formales y la aritmetización de estos en el teorema gödeliano, indican que las matemáticas[12] no funcionan tal y como creen los defensores de la IA *fuerte*.

Los seguidores de las ideas de Turing defienden que nuestro pensamiento responde a un tipo de computación (o algoritmo) de una complejidad enorme, complejidad culpable de que dicha computación no haya podido ser abarcada aun. Sin embargo, esa situación no sería definitiva, porque las soluciones a sistemas cerrados suelen darse tarde o temprano (tanto si es para confirmar como para refutar aquello que se busca).

Pero una vez examinado el teorema de Gödel, hay un fuerte soporte para pensar que la pretensión de los partidarios del punto de vista A[13] no resulta tan clara. Los mismos procedimientos matemáticos en sí mismos no son infalibles, ya que su alcance depende de algo que está más allá de ellos. Lo que plantea Gödel es que aquello que está más allá de los procesos y posibilita que alcancen sus metas es, precisamente, la intervención de la mente

[12] Que es «donde nuestros procesos mentales alcanzan su forma más pura» (Penrose, 2012: 78)

[13] IA *fuerte*, según la clasificación establecida por Penrose en *Las sombras de la mente* 2012: 26).

humana. Penrose encuentra en el argumento del matemático austríaco una base firme en la que sostenerse para situarse en contra de la IA *fuerte*[14]. También me ha servido personalmente para inclinarme hacia el lado de la balanza en el que se encuentra el matemático británico sobre este asunto.

El teorema constituye una herramienta[15] muy útil si lo que se quiere es rebatir la posibilidad de la computabilidad de la mente y la consciencia humana. Y no es para menos. Ateniéndonos a lo que nos dice el teorema, colegimos que da una respuesta contundente acerca de la computabilidad (más bien de la no-computabilidad) de los procesos mentales. ¿Cómo podemos deducir de un discurso abstracto, como la aritmetización de los sistemas formales tipo PM, un argumento en contra de aquello que defienden los partidarios de la IA *fuerte*? Planteada así la pregunta puede sospecharse que Penrose fuerza el razonamiento gödeliano para que entre dentro de su propia constelación mental. Pero no es así. La propuesta de Gödel encaja perfectamente en la de Penrose sin necesidad de compeler una con otra.

La clave para poder ver la relación de la idea de Gödel con nuestro autor, se encuentra en un concepto esbozado en el apartado anterior: la *reflexión*.

¿En qué consiste la reflexión? En el ámbito de la filosofía existen muchas posturas para responder a esto[16]. Algunas de ellas nos llevarían hacia terrenos considerablemente alejados del que nos ocupa. Pero en la inmensa mayoría de esas corrientes la idea que subyace es la de volver sobre uno

[14] Esta base tiene que ver con la relación que tenemos los seres humanos con el mundo matemático platónico que el mismo Penrose defiende que existe en su teoría de los tres mundos. Esto es algo que no se desarrollará aquí. Para dicho tema en concreto véase (Penrose, 2012: 433-442; Penrose, 2004: 17-21).

[15] Da la impresión de que para Penrose es la herramienta con mayúsculas.

[16] Me permito una brevísima mención a las filosofías en las que estoy pensado: creo que podemos observar tanto en la filosofía platónica como la aristotélica ese interés por la reflexión. Por ello también es tan fácil advertir que la tendencia a tratar este tema ocupa un lugar importante dentro de la filosofía medieval (tanto en la tradición seguidora de Porfirio y Plotino, como la correspondiente a la de Jámblico y Proclo).

mismo[17]. Un proceso algorítmico cuando vuelve sobre sí mismo entra en un bucle infinito del cual es imposible salir. En cambio, los seres humanos no entramos en ese callejón sin salida cuando volvemos hacia nosotros mismos, sino que tenemos la posibilidad de encontrar una luz que nos permite salir del atolladero. A fin de cuentas, se trata de la facultad creadora que tiene la mente humana en contraste con las capacidades de las computadoras. Supone *algo más*, que desborda las posibilidades de éstas. De todos modos, hablar de «facultad creadora» puede llevar a confusión, en el sentido de que podría pensarse que la actividad matemática es un invento del hombre.

No defiendo eso. La «facultad» a la que me refiero es a la capacidad que tiene la mente humana de poder ir más allá de lo que tiene *de serie*. Defender o insinuar que la matemática es un invento humano heriría la sensibilidad de no pocos matemáticos. Entre ellos el mismo Penrose:

> Debería señalar en primer lugar que cuando los matemáticos elaboran sus minuciosas cadenas de razonamiento consciente para establecer verdades matemáticas, no *piensan* que estén siguiendo ciegamente reglas inconscientes que son incapaces de conocer y creer. Ellos piensan que están basando sus argumentos en lo que son verdades incuestionables —en definitiva, esencialmente «obvias»— y que están construyendo sus cadenas de razonamiento a partir únicamente de tales verdades. Y aunque estas cadenas puedan a veces ser extraordinariamente largas, difíciles o conceptualmente sutiles, el razonamiento es, en principio y de raíz, incuestionable, firmemente creído y lógicamente impecable. No tienden a pensar que estén actuando en realidad de acuerdo con ciertos procedimientos completamente diferentes, desconocidos o no creídos que, quizá «entre bastidores», guían sus creencias de maneras incognoscibles (Penrose, 2012: 142).

A pesar de la importancia que tiene el teorema de Gödel para el pensamiento de Penrose, el matemático inglés no comparte la totalidad del pensamiento del lógico austríaco. ¿En qué se alejan las pretensiones de uno

[17] Que es al fin y al cabo la noción que hemos heredado nosotros y cómo podemos entenderla hoy en día.

y otro? Gödel podría aceptar la idea de que la mente de los matemáticos humanos en realidad responde al dictamen de un algoritmo muy complejo del que no son conscientes. Pero esto no se podría demostrar. Tal defensa situaría a Gödel como partidario del punto de vista D[18] (2012: 143). Penrose, al adoptar el punto de vista C, sostiene que sí es demostrable la no existencia de tal algoritmo. Es algo que podría conseguirse si se llevara a cabo una modificación en la física actual.

Entre estos dos pensamientos se encuentra el mío. Comparto con Penrose la creencia de no conceder a un algoritmo (por muy sutil que este sea) la exclusividad de poder explicar los procesos mentales humanos. Pero la precaución de Gödel al defender que la demostración de ello es imposible me lleva algo más lejos que el positivismo del matemático inglés. He de añadir que mi formación en materia científica no constituye una credencial fiable para rebatir a un autor del peso de Penrose. Pero no pretendo entablar un debate a nivel científico. Mi intención es mantenerme en el terreno de la filosofía, donde uno puede permitirse del lujo de codearse de igual a igual con pensadores de la talla de nuestro autor.

4. Cómo conectar el teorema gödeliano y la máquina de Turing

La conexión entre el teorema gödeliano y la máquina de Turing tiene un sencillo punto de unión. Uno y otra constituyen respuestas al *Entscheidungsproblem* de Hilbert. Por otra parte, su mutua conexión tiene que ver con las desavenencias directas entre ambos autores. Dichas discrepancias pueden considerarse, de manera transversal, como el inicio del debate de la computabilidad o no computabilidad de la mente y la consciencia humana. Gödel y Turing nunca debatieron de forma directa acerca de este asunto. Sus posturas no fueron construidas en los términos considerados en este trabajo. No obstante, teniendo en cuenta el desarrollo posterior de sus

[18] Sigo con la clasificación penroseana de *Las sombras de la mente* citada anteriormente (2012: 26).

ideas, tiene pleno sentido situar sus contribuciones en el inicio de la polémica.

Penrose destaca una de las críticas que Alan Turing lanza directamente a la propuesta de Gödel (2012: 144). En ella el matemático inglés quiere aclarar que el enfoque del austríaco es diferente del suyo propio. Pero, ¿en qué sentido? Turing observa que el planteamiento de Gödel[19] está referido a máquinas que sean infalibles. Este enfoque deja fuera a las que no son infalibles, que son las que interesan a Turing. Si lo que se quiere es igualar las capacidades humanas, sabemos de sobras que estas no son, ni mucho menos, perfectas, en sentido matemático. Esto provoca que el teorema gödeliano lleve hacia un hermetismo conceptual que le impide ver más allá de sí mismo y de aquello que dicta.

En mi opinión esto no es correcto. El teorema se desarrolla en el sentido contrario. En un principio está dirigido hacia un sistema formal concreto (PM). Pero no deja de ser cierto que puede acabar siendo extendido a cualquier sistema de este tipo. Así que aquello que recrimina Turing al teorema de Gödel es inexacto. De todos modos, la diferencia entre Turing y Gödel va más allá. Penrose se percata de ello y señala la sutilidad de la diferencia entre ambos:

> Los «teoremas» que tenía en mente (Turing)[20] son indudablemente el teorema de Gödel y otros relacionados con él, tal como su propia versión «computacional» del teorema de Gödel. Así pues, parece que él ha considerado la

[19] O más bien de los seguidores de este, ya que hemos visto que él mismo no procuraba entrar en el debate de la computabilidad de forma íntegra, sino que su exposición estaba dirigida hacia el terreno de la lógica y la matemática. Su posición acerca de la posibilidad de la adquisición por parte de las máquinas de las capacidades mentales humanas era más bien precavida (como vimos Penrose lo sitúa dentro del punto de vista D). Aun así, es plenamente lícito que Turing le pida explicaciones —en sentido figurado, claro está— a lo que Gödel defiende en su teorema.

[20] Paréntesis añadido por mí.

> imprecisión del pensamiento matemático humano como algo esencial, permitiendo que la (supuesta) acción algorítmica imprecisa de la mente proporcione una potencia mayor que la que sería alcanzable por medio de cualquier procedimiento algorítmico completamente válido. En consecuencia, sugirió un modo de evitar las conclusiones del argumento de Gödel: el algoritmo del matemático sería técnicamente no válido, y ciertamente no sería «cognosciblemente válido». Así pues, el punto de vista de Turing sería consistente con G[21], y parece probable que hubiera estado de acuerdo con el punto de vista A (Penrose, 2012: 114).

Penrose siente la necesidad de aclarar qué tiene de cuestionable tal argumento:

> Como parte de la exposición siguiente, voy a presentar mis razones para rechazar que la «no validez» en un algoritmo de un matemático pueda ser la explicación *real* de lo que está pasando en la mente del matemático. Existe, en cualquier caso, cierta inverosimilitud intrínseca en la idea de que lo que hace a la mente superior a un ordenador preciso es la imprecisión de la mente —especialmente cuando estamos interesados, como aquí, en la capacidad del matemático para *percibir la verdad matemática incuestionable*, más que en la originalidad o la creatividad matemática. Es un hecho sorprendente que cada uno de estos dos grandes pensadores, Gödel y Turing, se viese llevado, por consideraciones tales como G, a lo que muchos podrían considerar un punto de vista de algún modo inverosímil. Es interesante especular sobre si se hubieran visto llevados en esta dirección si hubieran estado en situación de contemplar seriamente la posibilidad de que la acción física pudiera, en su raíz, ser algunas veces no computable —de acuerdo con el punto de vista C que estoy proponiendo aquí— (Penrose, 2012: 144).

[21] Este símbolo tiene el significado del teorema de Gödel (junto con el de Turing) que Penrose desarrolla en *Las sombras de la mente* (pp. 87-91) y que viene a decir básicamente que «los matemáticos humanos no están utilizando un algoritmo cognosciblemente válido para asegurar la verdad matemática» (Penrose, 2012: 90). Esto es lo que nos dice tanto el teorema de Gödel como Turing (este último en su respuesta al problema de la parada).

Es decir: la «no validez»[22] de la que habla Turing (o mejor dicho, la que Penrose interpreta que estaría incluida en la defensa de Turing) supone la llave que da con la diferencia entre lo que puede postularse a través del teorema de Gödel y lo que te garantiza seguir las ideas de Turing[23]. Pero, aparte de ello, también destaca otra característica: la *cognoscibilidad*[24]. Según Penrose:

> Debemos distinguir claramente entre tres puntos de vista distintos con respecto a la cognoscibilidad de un supuesto procedimiento algorítmico *F* subyacente en la comprensión matemática, sea válido o no válido. En efecto *F* podría ser:
> I. conscientemente cognoscible, y tal que su papel como el algoritmo real subyacente en la comprensión matemática es también cognoscible;
> II. conscientemente cognoscible, pero su papel como el algoritmo real subyacente en la comprensión matemática es inconsciente y no cognoscible;
> III. inconsciente y no cognoscible (Penrose, 2012: 146).

Tras un análisis lógico sobre las premisas de la conclusión G y del primer punto de vista de la cita, Penrose determina que un procedimiento algorítmico que sigue las condiciones de **I** no puede considerarse como «una

[22] Entendamos la «validez» o «no validez» en el sentido a como lo entendimos anteriormente (en §2), cuando hablábamos de la consistencia. Es decir, dado un sistema consistente, es imposible deducir una contradicción dentro de este. Como nos dicen Manzano y Huertas (2011: 8): «la *consistencia lógica* o coherencia de un conjunto de creencias significa para nosotros *compatibilidad* de creencias».

[23] Es conveniente recordar que estoy siguiendo la línea argumental de Penrose, quien defiende que tanto Gödel como Turing ponen el acento en la verdad incuestionable de las matemáticas, cometiendo ambos el mismo error. Y considera que en un grado más significativo lo hace el lógico austríaco, ya que su teorema no lleva necesariamente a esa conclusión sino que existen alternativas (precisamente la postura de Penrose pretende ser una de ellas).

[24] Entendida en su sentido más amplio, es decir, la capacidad que tiene algo —en este caso hablamos del algoritmo o la computación que buscamos— para ser conocido.

posibilidad seria» (Penrose, 2012: 147). Siguiendo el teorema de Gödel, eso implica que sus convicciones[25] no sean relevantes para subyazcan en la comprensión matemática de la que se habla.

Con respecto a **II** y **III**, el matemático inglés ofrece unos argumentos ampliamente elaborados por los que acaba rechazándolos también. Para no dejarlos en el aire, reseñaré algunos de sus más destacables aspectos. En el caso de **II**, Penrose no puede rechazarlo lógicamente del mismo modo que a **I**. Es más, si seguimos un razonamiento de tipo Gödel, no hay «ninguna manera evidente de descartar el caso **II** solamente sobre fundamentos lógicos rigurosos» (Penrose, 2012: 148). ¿Deberíamos entonces aceptar el punto de vista **II**? La respuesta es no. Para mostrar por qué, Penrose recurre a las matemáticas, aunque sin dejar de lado los procedimientos lógicos:

> ...ninguno de estos nos dice que tal *F* —la supuesta «máquina de demostrar teoremas» de Gödel— es una imposibilidad, pero, a partir del punto de vista de nuestras comprensiones matemáticas, su existencia parece muy poco probable. En cualquier caso, no existe por el momento la más mínima sugerencia sobre la naturaleza de un *F* plausible semejante, ni hay ningún indicio de su existencia. Podría ser solamente una *conjetura*, en el mejor de los casos —y aun así una conjetura indemostrable. (¡Demostrarla la contradiría!). Me parece que sería extremadamente precipitado para cualquier defensor de la IA (ya sea A o B) tener esperanzas en encontrar semejante procedimiento algorítmico, encarnado por *F*, cuya misma existencia es dudosa en extremo y, en cualquier caso, si existiera su construcción real estaría más allá del ingenio de cualquiera de los matemáticos o lógicos actuales (Penrose, 2012: 153).

En definitiva, no es posible rechazar de manera tajante lo defendido en **II**, pero tampoco resulta conveniente aceptarlo.

[25] Penrose en este caso particular hace referencia a la validez de un sistema formal al que denomina **F**, cuyo desarrollo corresponde a la «máquina de demostrar teoremas» de Gödel.

Por último, Penrose encuentra en **III** una premisa que difícilmente puede ser admitida, ya que, de hacerlo, se estaría siguiendo los pasos propios del punto de vista A, aunque por otro lado no estaría respondiendo a la semejanza de la máquina con el ser humano:

> Supongamos que, de acuerdo con **III**, la especificación de un *F* semejante está más allá de las capacidades humanas. ¿Qué nos diría esto sobre la perspectiva de una estrategia IA completamente exitosa (de acuerdo con la IA «fuerte» o «débil» —los puntos de vista respectivos de A y B)? Quienes creen en sistemas IA controlados por ordenador (ciertamente bajo el punto de vista A, y quizá también bajo B) podrían prever que las creaciones robóticas que pueden surgir con el tiempo como resultado de esta estrategia deberían ser capaces de alcanzar y quizá superar las capacidades matemáticas humanas. En consecuencia, debería darse el caso, si aceptamos **III**, que algún algoritmo *F* semejante humanamente inespecificable formara parte del sistema de control se semejante robot matemático. Esto parecería implicar que una estrategia IA de tal alcance eventual es imposible. Pues si se necesitara un *F* inespecificable para conseguir sus propósitos, no habría ninguna esperanza de que los seres humanos lo pusiesen alguna vez en acción (Penrose, 2012: 159).

Vemos entonces que lo que conecta el teorema de Gödel y la máquina de Turing consiste en un error. Concretamente el error de situar la importancia para la demostración del pensamiento computacional o algorítmico en las verdades incuestionables de las matemáticas, cuando, en realidad, estas no son relevantes. Hemos visto que esta condición es necesaria en **I**, y de un modo bastante claro tanto en **II** como en **III**:

> ...los procesos de esta naturaleza no escapan realmente al problema: si los mismos procesos mediante los que se establece inicialmente una estrategia IA son algorítmicos y cognoscibles, entonces cualquier *F* resultante debería ser también cognoscible. De este modo, el caso III se reduciría o bien a I o bien a II, casos que fueron descartados (Penrose, 2012: 160).

En última instancia, Penrose está hablando de la pertenencia a los diferentes puntos de vista (perteneciendo Gödel al D, pudiendo aceptar como válido **II**; Turing al A, quien admitiría **I**; y Penrose a C —de manera fuerte—, que rechazaría los tres). ¿En qué grado podemos decir que el teorema gödeliano influye en el pensamiento del matemático inglés? Negar cualquier tipo de conexión sería ir en contra de lo que el mismo Penrose defiende. Y no es que condene ese tipo de posicionamientos, ya que yo mismo me atrevería a ir un poco más allá. Defiendo que en este apartado en concreto el teorema es más decisivo de lo que pretende exponer el autor de *Las sombras de la mente*. No obstante, los argumentos que pudiera ofrecer estarían muy lejos de ser decisivos.

Bibliografía empleada

P. Álvarez, J.A. Bañares, P. Latorre, S. Velilla, *Programación*, Zaragoza, C.P.S. Universidad de Zaragoza, 2005.

J. Arana, La conciencia inexplicada: Ensayo sobre los límites de la comprensión naturalista de la mente, Madrid, Biblioteca Nueva, 2015.

—, Los sótanos del universo: La determinación natural y sus mecanismos ocultos, Madrid, Biblioteca Nueva, 2012, pp. 175-241.

K. Gödel, *Obras completas*, trad. por Jesús Mosterín, Madrid, Alianza Editorial, 2006, pp. 53-89.

R. Herce Fernández, De la física a la mente: El proyecto filosófico de Roger Penrose, Madrid, Biblioteca Nueva, 2014.

J. Ladrière, Limitaciones Internas de los Formalismos: Estudio sobre la significación del Teorema de Gödel y teoremas conexos en la teoría de los fundamentos de las matemáticas, trad. por José Blaso, Madrid, Tecnos, 1969.

J. O. La Mettrie, *El hombre máquina*, trad. por Ángel J. Cappelletti, Buenos Aires, Eudeba, 1961, pp. 21-29.

M. Manzano Arjona / A. Huertas Sánchez, *Lógica para principiantes*, Madrid, Alianza Editorial, 2011, pp. 3-20, pp. 165-184.

J. McCarthy, "Ascribing mental qualities to machines", Stanford (CA), Stanford University, 1979.

R. Penrose, *La nueva mente del emperador*, trad. por José Javier García Sanz, México D.F., Fondo de Cultura Económica, 1996.

—, *Las sombras de la mente*, trad. por José Javier García Sanz, Barcelona, Crítica, 2012.

—, (con A. Shimony, N. Cartwright y S. Hawking), *Lo grande, lo pequeño y la mente humana*, trad. por José Javier García Sanz, Madrid, Cambridge University Press, 1999, pp. 11-14.

—, The road to reality: A Complete Guide to the Laws of the Universe, London, Jonathan Cape, 2004, pp. 7-25.

F. Rodríguez Valls [ed.], La inteligencia en la naturaleza: Del relojero ciego al ajuste fino del universo, Madrid, Biblioteca Nueva, 2012.

A. Turing, *Maquinaria computacional e Inteligencia,* trad. por Cristóbal Fuentes Barassi, Santiago de Chile, Universidad de Chile, 2010.

Daniel Heredia González
dani_hergon@hotmail.com