HUMAN REVIEW | Vol. 11, No. 2, 2022 | ISSN 2695-9623
International Humanities Review / Revista Internacional de Humanidades
DOI: https://doi.org/10.37467/revhuman.v11.4503

MENTES CONTRA MÁQUINAS
Revisión histórica y lógico-filosófica del argumento gödeliano de Lucas-Penrose

Minds vs Machines
Historical and logical-philosophical review of the Lucas-Penrose's Gödelian argument

KARIM GHERAB
Universidad Rey Juan Carlos, España

KEYWORDS

Gödel's Theorem
Artificial Intelligence
Mechanism
Free Will
Philosophy of Mind
Consciousness

ABSTRACT

This paper presents, from a historical and logical-philosophical perspective, the Gödelian arguments of two Oxford scholars, John Lucas and Roger Penrose. Both have been based on Gödel's Theorem to refute mechanism, computationalism and the possibility of creating an AI capable of simulating or duplicating the human mind. In the conclusions, the growing application of empirical methods in mathematics is mentioned and a possible path that would support Lucas and Penrose's arguments is speculated.

PALABRAS CLAVE

Teorema de Gödel
Inteligencia artificial
Mecanicismo
Libre albedrío
Filosofía de la mente
Consciencia

RESUMEN

Este artículo presenta, desde una perspectiva histórica y lógico-filosófica, los argumentos gödelianos de dos investigadores oxonienses, John Lucas y Roger Penrose. Ambos se han basado en el Teorema de Gödel para refutar el mecanicismo, el computacionalismo y la posibilidad de crear una IA capaz de simular o duplicar la mente humana. En las conclusiones, se menciona la creciente aplicación de métodos empíricos en las matemáticas y se especula con una posible vía que daría apoyo a los argumentos de Lucas y Penrose.

Recibido: 10/08/2022
Aceptado: 27/10/2022

1. Introducción. Teoremas de Gödel

A principios del siglo XX, el lógico austriaco Kurt Gödel presentó sus famosos Teoremas de Incompletitud[1] (Gödel, 1931), donde demostraba el fracaso de cualquier tentativa por sistematizar la aritmética, es decir, mostró con total rigurosidad lógico-matemática que siempre sería posible encontrar al menos un enunciado (construido con los símbolos pertenecientes a la lógica y a la aritmética) cuya verdad o falsedad fuera imposible demostrar con los elementos de la lógica y la aritmética, si bien sí sería factible hacerlo apelando a un sistema lógico-matemático más amplio (una meta-matemática) que, naturalmente, incluyera la lógica y la aritmética. Se trataba de enunciados *indeterminados* o *indecidibles*, esto es, enunciados que no podían ser probados verdaderos, pero tampoco falsos. Seguidamente, Gödel demostró que no era posible demostrar la consistencia de un sistema formal puesto que siempre habría enunciados indecidibles. La aportación de Gödel provocó un cisma en la lógica-matemática del siglo XX, con dramáticas consecuencias para los matemáticos formalistas (Nagel y Newman, 1958):

1) Cualquier sistema formal matemático sería pues incompleto, es decir, siempre sería posible construir un enunciado (a partir del vocabulario de ese sistema) que no podría ser demostrado dentro de dicho sistema;
2) No se podría entonces demostrar la consistencia de un sistema formal que fuera efectivamente consistente a partir de los axiomas y reglas de inferencia de dicho sistema formal, es decir, no se podría asegurar que de los axiomas básicos de la aritmética (o de un sistema formal mayor que incluyera a la aritmética) no fueran a derivarse contradicciones.

2. El argumento gödeliano de John Lucas

Un artículo[2] del filósofo oxoniense John R. Lucas (1961) provocó el primer gran debate filosófico al expresar con claridad y sin tapujos las implicaciones del Teorema de Gödel sobre el mecanicismo (y, por extensión, sobre el computacionalismo y la inteligencia artificial o IA), algo que tanto Gödel como Nagel y Newman habían ya apuntado tímidamente. Estos últimos afirmaron que el Teorema de Gödel conducía "a la cuestión de si podría construirse una máquina calculadora que llegara a equipararse en inteligencia matemática al cerebro humano" (Nagel & Newman, 1958, p. 100), habiéndose dado el caso de que, puesto que "lo que entendemos por proceso de la prueba matemática no coincide con la explotación de un método axiomático formalizado [...], la propia argumentación de Gödel señala [que] no es posible trazar ningún límite a la inventiva de los matemáticos en la ideación de nuevas reglas de prueba" (Nagel & Newman, 1958, p. 99).

Lucas comenzaba su "altamente polémico y retador" (Hofstadter, 1992) artículo del siguiente modo:

> Tengo la impresión de que el Teorema de Gödel demuestra que el Mecanicismo es falso, es decir, que las mentes no pueden ser explicadas como si fueran máquinas. (Lucas, 1961, p. 112)

Y añadía:

> La misma impresión han tenido otras muchas personas: casi todos los lógicos-matemáticos a los que les he planteado el asunto han confesado pensamientos similares, pero se han sentido reacios a comprometerse definitivamente hasta que les fuera expuesto el argumento completo, con todas las objeciones completamente establecidas y debidamente satisfechas. (Lucas, 1961, p. 112)

Los argumentos de Lucas pueden resumirse en los tres siguientes extractos sacados de su artículo:

1) "El teorema de Gödel sostiene que, en cualquier sistema consistente y suficientemente capaz de producir la aritmética simple, hay fórmulas que no pueden ser demostradas dentro del sistema, pero que nosotros podemos ver que son verdaderas";

[1] En este artículo nos referiremos a ellos en singular y de manera escueta como Teorema de Gödel.
[2] El artículo corresponde a una conferencia que J. R. Lucas dio a la Oxford Philosophical Society el 30 de octubre de 1959.

2) "El teorema de Gödel debe aplicarse a las máquinas cibernéticas, ya que el ser una ejemplificación concreta de un sistema formal pertenece a la esencia de una máquina";
3) "Resulta de ello que ninguna máquina puede ser un modelo completo o adecuado de la mente y que las mentes difieren esencialmente de las máquinas".

Enseguida surgieron respuestas en contra del argumento de Lucas. Una de ellas señalaba que, aunque ciertamente una máquina no podía demostrar la veracidad de la fórmula de Gödel, sí se le podía enseñar a colegir (gödelizar) dicha fórmula y a incluirla dentro del sistema como un axioma adicional (Good, 1967; Hofstadter, 1992; ver también Penrose, 1994). Lucas (1968a, 1970b) respondió rápidamente que esta estrategia presentaba la siguiente dificultad: a cualquier programa Ω capaz de efectuar gödelizaciones de este tipo se le podía aplicar una nueva fórmula de Gödel que estuviera fuera del alcance de Ω, porque a cada programa le correspondería alguna fórmula de Gödel que no podría derivarse de dicho programa. A pesar de las reiteradas peticiones de sus adversarios (Good, 1967; Lewis, 1969; Hofstadter, 1992), que le exigían presentar un operador de gödelización para demostrar cómo las mentes humanas se las arreglan para hacer uso de este operador a la hora de exponer enunciados de Gödel, Lucas (1968a) se mantuvo firme, al igual que Penrose (1994) muchos años después, señalando que su refutación al mecanicismo no era tanto una demostración como un contraejemplo o una *reductio ad absurdum*. En otras palabras, Lucas insistía en que su artículo no demostraba los límites de las máquinas en general, sino los límites de las máquinas particulares que le presentaran como candidatas a simular (o a duplicar) una mente: a cualquier máquina le correspondería un enunciado gödeliano que no sería capaz de demostrar, mientras que una mente sí podría hacerlo (Lucas, 1961). Cabe mencionar, no obstante, que Lucas (1961) se planteó en su artículo la cuestión de producir dicho operador de gödelización.

Hubo algunas críticas (Whiteley, 1962; Benacerraf, 1967; Lewis, 1969) en tono jocoso acerca de cómo la mente de Lucas se las arreglaría para hacer uso de un operador de gödelización capaz de producir enunciados de Gödel *ad infinitum*. Así, Whiteley (1962) retó a Lucas a pronunciar la siguiente proposición: "Este enunciado no puede ser afirmado de manera consistente por Lucas". Como es obvio, Lucas no podía pronunciar de manera consistente este enunciado, puesto que, si lo hacía, caía inmediatamente en una contradicción. Triquiñuelas de este tenor fueron presentadas posteriormente también a Penrose. Por ejemplo, McCullough (1995) propuso que la fórmula de Gödel G fuera el enunciado siguiente: "Este enunciado no es una indudable creencia de Roger Penrose". Por su parte, Moravec (1995) sugirió el enunciado siguiente: "Penrose debe equivocarse al creer en este enunciado". Este tipo de contrargumentos molestó a Penrose, que respondió a ambos diciendo que "es ciertamente una parodia intentar expresar los puntos esenciales de mi argumento (o en verdad los de Gödel) de esta manera" (Penrose, 1996).

Aunque se encontrara un operador de gödelización Ω (Lucas, 1961; Good, 1967; Hofstadter, 1992; Penrose, 1994) que generara, como pretendía también Alan Turing (1939) en su tesis doctoral, una tras otra, todas las fórmulas de Gödel G_1, G_2, G_3, ..., G_N, G_{N+1}, ... , añadiéndolas al sistema formal a cada paso, siempre habría una nueva fórmula gödeliana G(Ω) cuya validez no podría ser demostrada por ese sistema formal. Se podría entonces especular con la posibilidad de construir un operador de gödelización más potente Ω' capaz de demostrar la validez de G(Ω), pero, de nuevo, una mente humana sería capaz de generar una nueva fórmula de Gödel G(Ω') inalcanzable para el operador de gödelización Ω'. Este procedimiento podría repetirse una y otra vez, pero a cualquier nuevo operador de gödelización $\Omega^{(N)}$ Lucas respondería con una nueva fórmula de Gödel que echaría por tierra la coronación de una máquina como equivalente a una mente. Este juego de ida y vuelta, realmente basado en el procedimiento de diagonalización utilizado por George Cantor en sus descubrimientos de las propiedades de los conjuntos transfinitos, sugirió a I. J. Good (1967, 1969) la posibilidad de que la esencia del argumento de Lucas, que buscaba romper la equivalencia entre las mentes y las máquinas, no se hallaba en el Teorema de Gödel, sino en los métodos de diagonalización y de contaje transfinito. El físico César Gómez también ha considerado irrelevante el Teorema de Gödel para la tesis de Penrose, que, como veremos después, es similar a la de Lucas: "lo que Penrose quiere decir es que existe una diferencia esencial entre *comprensión* y *computabilidad*" (Gómez, 2003). No obstante, en este artículo no entraremos a tratar los conceptos de computabilidad y no-computabilidad tratados por Penrose (1989, 1994).

Sin embargo, la crítica más original y penetrante contra Lucas ha sido la de Paul Benacerraf (1967), que reconstruyó el argumento de Lucas mediante lógica simbólica. Benacerraf sostenía que existen tres hipótesis diferentes lógicamente consistentes con el Teorema de Gödel, y no sólo una como creía Lucas. Y afirmaba que se debe partir de la forma más general posible, que consiste en considerar las tres premisas siguientes:

a) '$Q \subseteq C_q$' $\in S^*$

b) '$C_q \subseteq S^*$' $\in S^*$

c) $S^* \subseteq C_q$

donde Q representa un sistema formal adecuado de la aritmética, S^* se define como $S^*=[x/\ S \vdash x]$, siendo $S=[x/$ puedo demostrar $x]$ y '$\vdash$' el símbolo correspondiente a la expresión "produce por lógica de primer orden", y donde C_q es un conjunto recursivamente enumerable (es decir, un conjunto cuyos miembros pueden ser generados uno tras otro mediante un procedimiento computacional) que hace las veces de una máquina de Turing particular cuyo programa (o tabla de máquina) es identificable (y recuperable) mediante el sufijo (número entero) q.

Expresado en lenguaje menos técnico: S representa al conjunto de procedimientos de que disponen los matemáticos humanos para demostrar convincentemente enunciados verdaderos y S^* representa al conjunto que engloba a S y a todos los enunciados verdaderos deducibles mediante el uso de la lógica de primer orden a partir de S que, ya sea porque son deducciones demasiado largas o porque son demasiado complicadas, no son alcanzables *en la práctica* por los matemáticos humanos.

El enunciado a) dice que los matemáticos humanos pueden *probar* (no en el sentido formal del término 'demostrar', sino que se refiere a los procedimientos heurísticos de que disponen los matemáticos humanos para derivar convincentemente enunciados verdaderos) que un conjunto recursivamente enumerable, generado por un procedimiento computacional C_q, es suficientemente extenso como para incluir como subconjunto a la aritmética (el conjunto Q) entre sus axiomas y reglas de inferencia de teoremas. El enunciado b) señala que los matemáticos humanos pueden *probar* que los procedimientos computacionales forman parte (esto es, son un subconjunto) de los procedimientos de que disponen los matemáticos humanos para demostrar convincentemente enunciados verdaderos. Y finalmente, el enunciado c) afirma que los procedimientos de que disponen los matemáticos humanos para demostrar convincentemente enunciados verdaderos forman parte (esto es, son un subconjunto) de los procedimientos computacionales (de una máquina de Turing), es decir, que los matemáticos humanos son máquinas de Turing.

En su artículo, Benacerraf mostraba que la combinación de estos tres supuestos conduce a una contradicción, a saber, que tanto la fórmula de Gödel G como su negación ~G son teoremas de S^*, lo cual significa que S^* es inconsistente. Se obtendría así lo contrario a como se había definido S^*, esto es, como el conjunto de procedimientos disponibles para los matemáticos humanos para producir únicamente enunciados verdaderos, siendo esto posible sólo si S^* es consistente, ya que, de lo contrario, los matemáticos humanos producirían tanto enunciados verdaderos como sus negaciones (enunciados falsos). Benacerraf afirmaba, por consiguiente, que uno de los tres supuestos debía ser descartado para no caer en contradicción. Y añadía que Lucas había rechazado el enunciado c) en su polémico artículo y que ello implicaba que las mentes de los matemáticos humanos podían *probar* fórmulas que las máquinas no pueden *demostrar*. Sin embargo, partiendo del caso más general posible, Lucas podría haber rechazado una de las otras dos premisas, a) o b), dejando inmune el supuesto c) que afirma que todos los procesos desempeñados por la mente pueden ser replicados por alguna máquina de Turing aún por descubrir.

Las otras dos alternativas posibles serían entonces:

1) que, siendo máquinas de Turing, los matemáticos humanos no pueden probar que las computaciones C_q son adecuadas para la aritmética, es decir, no son suficientemente extensas para contener la aritmética;
2) que, siendo máquinas de Turing, los matemáticos humanos no son capaces de demostrar que son máquinas de Turing.

Para Benacerraf, la primera alternativa parecía muy inverosímil y además no era defendida por los expertos, mientras que la segunda implicaba dramáticas consecuencias (ver también Lucas, 1968b; y Penrose, 1994) para los defensores de la IA porque sugería que, aún en el caso de que se consiguiera construir una máquina que duplicara la mente humana, el algoritmo correspondiente sería incognoscible para siempre. Sin embargo, no todo el mundo ha estado de acuerdo con esto, ya que John Searle (2000), por ejemplo, ha señalado que "del hecho de que nuestro conocimiento de [...] verdades [tipo Gödel] no venga de un algoritmo demostrador de teoremas no se sigue que no usemos algoritmos para llegar a esas conclusiones" (p. 66). Benacerraf (1967) resumía del siguiente modo el impacto de su argumento sobre la Psicología:

> Una persona a la que se lo expliqué concluyó que la Psicología tal y como la conocemos es entonces imposible. Veamos: si en el mejor de los casos no somos máquinas de Turing, entonces es imposible; y si lo somos, entonces hay cosas que no podemos saber sobre nosotros mismos ni sobre otros con los mismos comportamientos que nosotros. (p. 30)

Lucas (1968b) concedió cierto grado de verosimilitud a la propuesta de Benacerraf, pero la rechazó por vacía e intrascendente. Así, en relación con la dramática segunda alternativa, puntualizó con ironía a Benacerraf (y a los mecanicistas y defensores de la IA) indicando que sólo lograba proclamar un triste, vacío y tautológico: "Yo soy Yo". Y esto es algo que ya sabíamos todos… ¡antes incluso de nacer Gödel!

3. El argumento gödeliano de Roger Penrose

De ser verdad que somos máquinas de Turing y que no podemos demostrar qué máquina de Turing somos, resultaría que no podemos saber qué sistema formal nos representa y, por consiguiente, no podemos asegurar (Chalmers, 1995; McCullough, 1995) nuestra propia consistencia. Supongamos que somos máquinas de Turing y que F es el sistema formal que nos representa: en virtud del Teorema de Gödel, el enunciado de Gödel asociado a F sólo es verdadero si F es consistente y, además, no podemos demostrar su consistencia sin salirnos de F, lo cual implicaría, como se lamentaba Lucas (1961), que "un hombre no puede afirmar su propia consistencia". Por tanto, F puede ser inconsistente sin nosotros saberlo, es decir, que los seres humanos nos quedaríamos satisfechos con nuestras propias contradicciones puesto que seríamos, en ese caso, sistemas inconsistentes. Pero esto no parece ser lo que ocurre (Lucas, 1961; Penrose, 1994, 1996) puesto que cuando descubrimos contradicciones en nuestros argumentos, inmediatamente las corregimos. Cometemos errores y somos inconsistentes ocasionalmente, pero ello no significa que seamos sistemas inconsistentes puesto que los errores acaban siendo corregidos mediante una revisión detallada del argumento, ya sea gracias a una reflexión individual crítica y penetrante al problema o bien debatiendo y buscando el consenso (Penrose, 1994) con otros expertos del ramo. Además, tal y como señaló Putnam (1997), si fuéramos intrínsecamente inconsistentes:

> la evolución no debería haber mantenido seres inteligentes que tuvieran posibilidades de sobrevivir con esquemas de razonamiento que conducen a contradicciones en la práctica cotidiana. (p. 31)

Puesto que los defensores de la IA aducen que nuestras capacidades mentales pueden ser capturadas por un sistema formal F que no puede demostrar que F es consistente (es decir, que no puede demostrar su propia consistencia), la estrategia de Penrose (1994) consistió en mostrar precisamente, al igual que Lucas (1968b), cómo una mente humana reducible a un sistema formal F puede proclamar inapelablemente que F es consistente:

a) Sabemos que somos consistentes;

b) Sabemos que F captura nuestras capacidades racionales;

c) Por tanto, sabemos que F es consistente.

David Chalmers (1999), que consideraba más amenazante para los intereses de la IA el *argumento de la habitación china* de Searle (1980) que los argumentos de Lucas y Penrose, criticó la premisa (b) que utilizaba Penrose para mostrar que podemos demostrar que cualquier F que capture nuestras

capacidades mentales ha de ser necesariamente consistente. Su argumento, poniendo como ejemplo los sistemas conexionistas (tales como las redes neuronales artificiales), es que F podría ser un sistema computacional que no fuera reducible a axiomas y reglas de inferencia. Al igual que hizo Benacerraf con Lucas, Chalmers (1995, 1999) enfatizó que Penrose no tendría modo alguno de saber qué F captura nuestras capacidades mentales debido a que F podría ser extraordinariamente complicado (Hofstadter, 1992; Chalmers, 1995; Putnam, 1997; Searle, 2000; Gómez, 2003), o incluso no ser un sistema formal al uso (Hofstadter, 1992; Chalmers, 1995), o sea, reducible a axiomas y reglas de inferencia; y, por consiguiente, el programa de la IA seguiría estando a salvo. Sin embargo, Chalmers encontró más interesante otro argumento de Penrose, que Chalmers (1995) ha reconstruido del modo siguiente:

a) Yo sé que soy consistente;
b) Sea un sistema formal F tal que F captura mis capacidades mentales, lo cual se resume en el enunciado "yo soy F";
c) F es consistente, ya que yo soy consistente y "yo soy F";
d) Sea F' un sistema tal que F' es F implementado con el enunciado "yo soy F"; entonces F' es consistente puesto que F es consistente y "yo soy F" es un enunciado verdadero;
e) Yo sé que el enunciado de Gödel G(F') es verdadero, ya que F' es consistente;
f) Yo soy equivalente a F', puesto que yo soy equivalente a F implementado con el conocimiento de que yo soy F ("yo soy F");
g) Sin embargo, en virtud del teorema de Gödel, F' no puede demostrar que G(F') es verdadero, lo cual implica que yo no soy F' porque, en virtud de e), yo sé que G(F') es verdadero.
h) Por tanto, puesto que F' no captura mis capacidades mentales y puesto que F' es F implementado con "yo soy F", entonces la hipótesis inicial de que F captura mis capacidades mentales es falsa.

Para Chalmers, el punto importante está en que, según e), yo sé que G(F') es verdadero porque se supone que yo sé que soy F, pongamos que porque he comprobado mediante un *test de Turing* muy avanzado que el sistema computacional que los ingenieros han creado es idéntico a mí. El hecho de saber que yo soy F es lo que me permite reconocer la consistencia de F, por extensión la consistencia de F' y, por el teorema de Gödel, la verosimilitud de G(F'), conduciéndome esto último a la contradictoria circunstancia de afirmar que soy F sin serlo en realidad. De aquí se concluiría que los ingenieros de la IA no podrían descubrir, ni siquiera empíricamente, algún sistema F que duplicara la mente humana, ya fuera F un sistema formal reducible a axiomas y reglas de inferencia o bien un sistema conexionista o una máquina de otro tipo. Pero, como ha matizado acertadamente Searle (2000), no es necesario que yo sepa que soy F, es decir, no es necesario que implemente F con el enunciado "yo soy F", puesto que yo puedo ver que G(F') es verdadero únicamente *suponiendo* que soy F (siendo F consistente por ser yo consistente) porque eso es lo que, por ejemplo, me insisten los ingenieros de la IA que soy. La conclusión de Penrose seguiría entonces siendo la misma: para cada F que me presenten los ingenieros, F seguirá sin poder capturar todas mis capacidades mentales.

Tanto para Chalmers como para McCullough, la grieta en el último argumento de Penrose se encontraba en la suposición de que los seres humanos saben que son consistentes. Con un sutil razonamiento, McCullough (1995) mostraba que "si Roger Penrose se considera consistente, entonces es realmente inconsistente". Los pasos de su argumentación son los que se indican a continuación:

a) Sea G el enunciado de Gödel siguiente: "Este enunciado no es indudablemente verdadero para Roger Penrose";
b) Si G es un enunciado indudablemente verdadero para Penrose, entonces G es necesariamente falso, luego Penrose (para permanecer consistente) no debe reconocer G como indudablemente verdadero;
c) Podemos entonces reescribir el paso b) de otro modo, construyendo el enunciado siguiente: "Si Roger Penrose es consistente, entonces G es verdadero";

d) Si Penrose se considera a sí mismo consistente y considera asimismo que G es verdadero en virtud de c), entonces Penrose se contradeciría y sería inconsistente ya que, según a), Penrose no puede considerar G como verdadero;
e) Si Penrose considera que el enunciado de c) (que dice que "Si Roger Penrose es consistente, entonces G es verdadero") es verdadero entonces, de acuerdo con el punto b), G es falso aun cuando Penrose seguiría pensando que G es verdadero;
f) Por tanto, llegamos al siguiente enunciado: "Si Roger Penrose se considera consistente, entonces es realmente inconsistente".

El argumento anterior parecía mostrar que no podemos construir ningún argumento que se fundamente en nuestra consistencia puesto que, en ese caso, dejaríamos de ser consistentes y nuestro argumento perdería toda su fuerza. Penrose (1996) consideraba que este tipo de enunciados son falaces y pueden ser descartados en un sistema válido como es la mente humana. Desde luego, parece haber algo extraño en el hecho de que, siendo consistentes, nos convirtamos inmediatamente en inconsistentes si nos declaramos consistentes o, como muestra un curioso experimento mental de McCullough (1995), únicamente lo pensemos. Realmente, al final no queda claro en qué situación queda la discusión ya que, por ejemplo, si por un lado Chalmers apoya el argumento de McCullough, por el otro le cuesta negar, al igual que Penrose, que Putnam y que cualquiera (¡incluido McCullough!) que presente argumentos con la convicción de que son válidos, la consistencia de la mente humana puesto que podría responderse (quien sabe si consistente o inconsistentemente) a McCullough que sus argumentos carecen de valor si duda de su propia consistencia. De hecho, podríamos afirmar que si McCullough es inconsistente, entonces lo que dice no tiene sentido puesto que diría cosas incoherentes y se estaría contradiciendo constantemente, y, además, podríamos añadir que si McCullough es equivalente a un sistema formal entonces lo que dice tampoco tiene sentido puesto que sus "argumentos" serían el resultado de diversos cómputos y no serían interpretables semánticamente, sino sólo sintácticamente, careciendo sus palabras de cualquier tipo de intencionalidad. Recuérdese que, siguiendo a Franz Brentano (1874), los filósofos de la mente consideran la *intencionalidad* como una de las propiedades esenciales que caracteriza *lo mental* frente a *lo* (meramente) *material*.

Por su parte, Searle pensaba haber refutado el último argumento de Penrose aduciendo que F no tendría por qué ser reducible únicamente a métodos de demostración matemática y que podría estar más allá de lo que los razonamientos matemáticos pueden alcanzar. Parece ser que Putnam, que ha denunciado el argumento de Penrose como "un caso palmario de falacia matemática" (Putnam, 1994), no descartaba completamente que la mente humana pudiera ser inconsistente, pese a que, como advertía Chalmers, parece absurdo requerir (como algo esencial) que el sistema F buscado sea inconsistente para poder ver que los enunciados de Gödel son verdaderos. Putnam ha cambiado de opinión a este respecto. En todo caso, según Abner Shimony (1999), Putnam se ha unido a los defensores de la IA al reprochar a Penrose que no contemple como una posibilidad lógica que pueda existir una máquina de Turing que contemple las capacidades humanas pero cuyo programa maestro sea tan extremadamente complejo que no pueda ser entendido por las mentes humanas.

La indignada[3] respuesta de Penrose (1999) es que "las críticas de Putnam eran una parodia y resultaban particularmente irritantes porque no daba la impresión de haber leído siquiera aquellas partes del libro que estaban dirigidas a los mismos puntos que él planteaba" (Penrose, 1999, p. 136). Desde luego, como el propio Penrose admitía, existe la posibilidad lógica que aducía Putnam, pero también es cierto que, por muy complicado que sea el programa maestro, debe ser entendible *en principio* por parte de una mente humana si se sigue paso por paso, aunque en la *práctica* no sea posible como consecuencia de la brevedad de la vida; y si algún "F semejante, humanamente inespecificable, formara parte del sistema de control de semejante robot matemático [...] parecería implicar que una estrategia IA de tal alcance eventual es imposible" (Penrose, 1994).

Si el programa fuera humanamente inespecificable, cabrían dos alternativas según Penrose (1994):

[3] Además de la respuesta que hace en Penrose (1999), y que citamos a continuación, Penrose (1994, 1995) contestó indignado a Putnam a través de dos cartas publicadas en *The New York Times Book Review*, 8 de diciembre de 1994, p. 39, y 15 de enero de 1995, p. 31.

a) El programa sería un acto de Dios; una suerte de "Programa Maestro" según Putnam (1988), o "mente como cristal" en palabras de Daniel Dennett (1988);

b) El programa sería alcanzable por selección natural; utilizando el método de "ensayo y error" de un latonero (Putnam, 1988), o "mente como caos" en palabras de Dennett (1988).

Tales posibilidades lógicas respaldarían lo dicho por Putnam (1997) cuando advertía que Penrose había intentado presentar

> lo que en el mejor de los casos es un argumento filosófico discutible como si fuera un resultado matemático [...], [lo cual] ofende a cualquiera que conozca la diferencia entre las matemáticas y la filosofía. (p. 31)

En todo caso, para Penrose, la alternativa a) tiene un evidente carácter místico y su defecto consiste en ser irrefutable científicamente (o sea, no falsable). Dicha alternativa no admitiría discusión y dejaría el debate en una mera cuestión de Fe y, como ha apuntado Putnam (1988),

> El resultado final es por cierto pesimista: si no hay Programa Maestro, luego nunca iremos demasiado lejos en términos de la simulación de la inteligencia artificial. (p. 273)

La alternativa b) sí estaría abierta al debate científico, según Penrose. Por ejemplo, podría especularse (Hofstadter, 1992; Penrose, 1994; Dennett, 1988) con la posibilidad de que el programa capaz de capturar las capacidades de la mente humana estuviera codificado en el ADN y que sus propiedades fueran cambiantes en función de las mutaciones y de las presiones selectivas impuestas por el entorno (George, 1962). Posiblemente, esta perspectiva resultaría atrayente para los neurocientíficos y los defensores de modelos conexionistas.

Daniel Dennett ha sido uno de mayores defensores de la opción del proceso selectivo en IA, bajo el paraguas del paradigma evolutivo, para alcanzar una máquina con las capacidades mentales humanas. Pero entonces el asunto parecería más una cuestión de azar (que lográsemos implementar el algoritmo correcto que es capaz de hacer emerger la consciencia) que una cuestión de análisis racional. La solución al problema se asemejaría a la tarea de poner un parche tras otro en algún programa, el cual no sería otra cosa que "la expresión de billones de bits de 'hojalatería'" (Putnam, 1988). Desde luego, sorprende pensar que el ADN, con todas sus desventuras y circunstancias históricas producidas por las presiones selectivas del entorno, haya conservado como una ventaja selectiva la capacidad humana de hacer matemáticas abstractas. En particular, resulta chocante imaginar, por ejemplo, que el ADN contenga en su hilera aperiódica de bases el sistema formal matemático de, pongamos, Zermelo-Fraenkel. Y nos viene a la memoria el relato de una anécdota que el físico alemán Werner Heisenberg (1972) ha contado sobre el matemático húngaro John Von Neumann, en la que éste respondió a un biólogo evolucionista lo siguiente:

> El biólogo era un seguidor convencido del moderno darwinismo. Von Neumann era escéptico. El matemático [Von Neumann] llevó al biólogo a la ventana de su cuarto de estudio y le dijo: "¿Ve usted allá arriba, sobre la colina, aquel hermoso caserío blanco? Ha surgido al azar. A lo largo de millones de años se ha ido formando la colina, a través de procesos geológicos; crecieron árboles, se pudrieron, cayeron y volvieron a erguirse; más tarde, el viento cubrió fortuitamente la cima de la colina con arena; probablemente, un proceso volcánico lanzó las piedras sobre el pasaje, y, por casualidad también, quedaron éstas ordenadas por estratos. Y así siguió adelante el proceso. Evidentemente, en el curso de la historia de la Tierra se han ido originando otras cosas merced a estos desordenados procesos fortuitos. Pero he aquí que una vez, después de mucho, muchísimo tiempo, surgió también el caserío, a continuación entraron los hombres en él, y ahora son ellos sus habitantes". (p. 142)

Naturalmente, el biólogo no quedó satisfecho con tal explicación. Sin embargo, aquí creemos que algo parecido a lo que dijo Von Neumann es lo que pensaba Penrose (1996) cuando escribió lo siguiente:

> De acuerdo, éstas [Programa Maestro y programa evolucionado por selección natural] son posibilidades lógicas, ¿pero son realmente explicaciones plausibles? (Penrose, 1996, 4.5)

Aquí, nos posicionamos del lado de Lucas y Penrose, de modo que pensamos que, siendo posibles (o incluso plausibles), dichas explicaciones son improbables.

4. Consideraciones finales

Hemos visto la manera en que el filósofo oxoniense John Lucas (1961) formuló su polémico argumento contra el mecanicismo y, por extensión contra el computacionalismo y la IA. Se apoyó en el Teorema de Gödel para sostener que la mente humana no es equivalente a una máquina de Turing (o sea, una computadora). Su artículo provocó un acalorado debate académico que todavía dura actualmente, ya que se trata de un dilema no resuelto. Si Lucas tiene razón, el proyecto de la IA de crear mentes artificiales equivalentes a las naturales (humanas) sería un espejismo. Pero también hemos visto varias objeciones a las tesis de Lucas, algunas de las cuales abogan por considerar las mentes humanas como consistentes y otras que se inclinan por presentarlas como inconsistentes; y en ambos casos los argumentos de Lucas pueden fallar. Tras varios años de cierta calma, otro oxoniense, el físico-matemático Roger Penrose (1989, 1994), recuperó las tesis de Lucas para armar un nuevo ataque contra la IA. Penrose introdujo nuevos conceptos y argumentos en favor de la tesis anti-mecanicista: introdujo la mecánica cuántica en el debate, sugirió la posible conexión de la volición humana con procesos físicos no-computables, e incluso propuso (en colaboración con Stuart Hameroff) un lugar en el interior de las neuronas para buscar las bases biológicas de la consciencia (los microtúbulos y las tubulinas). En este artículo no hemos tratado estas novedosas propuestas de Penrose, sino que nos hemos centrado en los aspectos lógico-filosóficos que más se asemejaban a los argumentos utilizados previamente por Lucas y sus detractores: la consistencia o inconsistencia del sistema estudiado, los mecanismos de gödelización, el origen divino o darwiniano de eventuales algoritmos pensantes, etc.

En este debate, nos posicionamos del lado de Lucas y Penrose porque consideramos que el Teorema de Gödel conduce a una disyuntiva: o bien la mente humana no es una máquina de Turing, o bien existen problemas matemáticos irresolubles (tanto para mentes naturales como para mentes artificiales). Creemos firmemente que muchos de los problemas matemáticos (si no todos) que hoy en día se consideran indecidibles serán resueltos en el futuro mediante el descubrimiento de sorprendentes fenómenos físicos y el desarrollo de técnicas y tecnologías punteras. Pensamos que las matemáticas pueden desarrollarse (de hecho, lo hacen ya) mediante métodos empíricos (Echeverría, 1996; en prensa, previsto 2023). Recordemos que el matemático húngaro George Pólya (1965) llegó a definir las matemáticas por su capacidad de resolver problemas. Si otro húngaro, el matemático John Von Neumann, inició el camino hacia una nueva praxis matemática mediante el uso de potentes computadoras, actualmente con individuos como Stephen Wolfram (y su WolframAlpha[4]) o empresas multinacionales como Google, IBM y Microsoft (y sus revolucionarios computadores cuánticos), parece sólo una cuestión de tiempo que se logren resolver, mediante procedimientos empíricos, algunos de los problemas matemáticos considerados hoy irresolubles. No es descartable que logremos resolver problemas matemáticos mediante el uso de nuevos y sorprendentes fenómenos físicos (por ejemplo, cuánticos, usando el entrelazamiento o la contrafacticidad) que nunca lleguemos a comprender y/o a traducir a categorías que sean inteligibles para nuestras mentes. Desconocer los mecanismos que subyacen a una herramienta física que, al mismo tiempo, nos permite resolver problemas matemáticos que considerábamos irresolubles o indecidibles, sí podría suponer la *muerte filosófica* (que no *tecnológica* o *comercial*) definitiva del mecanicismo, del computacionalismo y de la IA.

[4] https://www.wolframalpha.com/.

Referencias

Benacerraf, P. (1967). God, the Devil and Gödel. *The Monist*, *51*(1), 9-32. https://doi.org/10.5840/monist196751112.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Duncke & Humblot. [Traducido parcialmente al español: *Psicología*. Schapire, 1946]

Chalmers, D. J. (1995). Minds, machines and mathematics. A review of *Shadows of the Mind* by Roger Penrose. *PSYCHE*, *2*.

Dennett, D. C. (1988). When Philosophers Encounter Artificial Intelligence. *Daedalus*, *117*(1), 283-295. https://www.jstor.org/stable/20025148.

Echeverría, J. (1996). Empirical Methods in Mathematics. A Case-Study: Goldbach's Conjecture. In: G. Munévar, (Eds), *Spanish Studies in the Philosophy of Science* (pp. 19-55). *Boston Studies in the Philosophy of Science*, 186. Springer. https://doi.org/10.1007/978-94-009-0305-0_2.

Echeverría, J. (en prensa, previsto para 2023). Tecnomatemáticas experimentales y sociedad: von Neumann y Wolfram. *Estudios Filosóficos*, *72*.

George, F. H. (1962). Minds, machines and Gödel: another reply to Mr Lucas. *Philosophy*, *37*(139), 62-63. https://doi.org/10.1017/S0031819100030898.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik*, *38*, 173-198. https://doi.org/10.1007/BF01700692.

Gómez, C. (2003). *Significado y libertad. Un ensayo en Filosofía del Lenguaje*. Siglo XXI.

Good, I. J. (1967). Human and machine logic. *The British Journal for the Philosophy of Science*, *18*(2), 144-147. https://doi.org/10.1093/bjps/18.2.144.

Good, I. J. (1969). Gödel's theorem is a red herring. *The British Journal for the Philosophy of Science*, *19*(4), 357-358. https://doi.org/10.1093/bjps/19.4.357.

Graubard, S. R. (Ed.) (1999). *El nuevo debate sobre la inteligencia artificial: sistemas simbólicos y redes neuronales*. Gedisa.

Heisenberg, W. (1972). *Diálogos sobre la física atómica*. La Editorial Católica.

Hofstadter, D. R. (1992). *Gödel, Escher, Bach: un Eterno y Grácil Bucle*. Tusquets.

Lewis, D. (1969). Lucas against mechanism. *Philosophy*, *44*(169), 231-233. https://doi.org/10.1017/S0031819100024591.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112-127, https://doi.org/10.1017/S0031819100057983.

Lucas, J. R. (1968a). Human and machine logic: a rejoinder. *British Journal for the Philosophy of Science*, *19*(2), 155-156. https://doi.org/10.1093/bjps/19.2.155.

Lucas, J. R. (1968b). Satan stultified: a rejoinder to Paul Benacerraf. *The Monist*, *52*(1), 145-158. https://doi.org/10.5840/monist196852111.

Lucas, J. R. (1970). Mechanism: a rejoinder. *Philosophy*, *45*(172), 149-151. https://doi.org/10.1017/S0031819100009815.

McCullough, D. (1995). Can humans escape Gödel? A review of *Shadows of the Mind* by Roger Penrose. *PSYCHE*, *2*(4), April 1995.

Moravec, H. (1995). Roger Penrose's gravitonic brains. A review of *Shadows of the Mind* by Roger Penrose. *PSYCHE*, *2*(6), May 1995.

Nagel, E. & Newman, J. R. (1958). *Gödel's proof*. Routledge.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press. [Traducido al español: *La nueva mente del emperador*, Grijalbo Mondadori, 1991]

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford University Press. [Traducido al español: *Las sombras de la mente: hacia una comprensión científica de la consciencia*, Grijalbo Mondadori, 1996]

Penrose, R. (1996). Beyond the doubting of a shadow: a reply to commentaries on *Shadows of the Mind*. *PSYCHE*, *2*(23), January 1996.

Penrose, R. (1999). *Lo grande, lo pequeño y la mente humana*. Cambridge University Press.

Pólya, G. (1965). *How to Solve It*. Princeton University Press.
Putnam, H. (1988). Much Ado about Not Very Much. *Daedalus*, *117*(1), 269-281. https://www.jstor.org/stable/20025147.
Putnam, H. (1994). Recensión de *Shadows of the mind*, *The New York Times Book Review*, 20 de noviembre de 1994, p. 7.
Putnam, H. (1997). Acerca de un mal uso del teorema de Gödel en la especulación sobre la mente. *Revista de Libros*, *3*, marzo 1997, 30-32.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-457. https://doi.org/10.1017/S0140525X00005756.
Searle, J. R. (2000). *El misterio de la conciencia*. Paidós.
Shimony, A. (1999). Sobre mentalidad, mecánica cuántica y la actualización de potencialidades. En: Penrose, R. (1999). *Lo grande, lo pequeño y la mente humana* (pp. 115-125). Cambridge University Press.
Turing, A. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, *45*(1), 161-228. https://doi.org/10.1112/plms/s2-45.1.161.
Whiteley, C. H. (1962). Minds, machines and Gödel: a reply to Mr Lucas. *Philosophy*, *37*(139), 61-62. https://doi.org/10.1017/S0031819100030886.