Daimon. Revista Internacional de Filosofía, en prensa, aceptado para publicación tras revisión por pares doble ciego.
ISSN: 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.681441

Gödel, Tarski y la paradoja del mentiroso

Gödel, Tarski and the Liar Paradox

VIOLETA CONDE[1]

Resumen: El presente artículo tiene como finalidad explorar la posible relación entre el concepto de *verdad* de Alfred Tarski y el primer teorema de incompletitud de Kurt Gödel. Para ello, se analizarán los diferentes procedimientos llevados a cabo para la definición de la noción de verdad, por un lado, y la construcción de la prueba de incompletitud, por otro, de manera que puedan salir a la luz las motivaciones que las alientan y las implicaciones que se derivan de ellas. Nuestra hipótesis de partida defiende que la cuestión de la autorreferencialidad en los lenguajes -ejemplificada en la clásica *paradoja del mentiroso*- será el alfa y omega de ambas cuestiones, así como su punto de convergencia.

Palabras clave: Autorreferencialidad, verdad, prueba, completitud, Gödel, Tarski

Abstract: This article attempts to explore a possible connection between Alfred Tarski's notion of truth and Kurt Gödel's first incompleteness theorem. To this end, we will analyze the different procedures carried out in order to define the concept of truth, on the one hand, and to construct the proof of incompletness, on the other, so the reasons that have encouraged them and the derived difficulties we shall face could be revealed. The assumption at the basis of our examination supports that the presence of self-referentiality in languages -exemplified in the Liar paradox- will be the alpha and omega of both issues, as well as their convergence point.

Keywords: Self-referentiality, truth, proof, completness, Gödel, Tarski

1. Introducción

El siglo pasado fue, sin duda, un periodo agitado para la disciplina matemática. En 1903, el filósofo y matemático inglés Bertrand Russell anunciaba el descubrimiento de la célebre paradoja que lleva su nombre, la cual evidenciaría los problemas presentes en la axiomatización original de la teoría de conjuntos propuesta por George Cantor y

Recibido: 26/09/2025. Aceptado: 02/11/2025.
[1] La autora es actualmente Profesora Sustituta de Lógica en la Universidad de Salamanca (USAL). Correo electrónico: violecon@usal.es. Líneas de investigación: metafísica y epistemología de la modalidad, fundamentalidad y modalidad, lógica modal, filosofía de la lógica y lógica. Publicaciones destacadas: Conde, V. [Forthcoming]. "On Individuating Contingently Non-Concrete Objects". *Metaphysica. International Journal for Ontology and Metaphysics*;Conde, V. (2023). "Necessitism and Unrestricted Quantification", Teorema. Revista Internacional de Filosofía, Vol. XLII/2.

Gottlob Frege. Fue este uno de los primeros grandes resultados que limitaba la matemática y ponía coto a sus aspiraciones, pero no fue el único. Algunos años después, en 1931, un jovencísimo Kurt Gödel publicaba las demostraciones de sus dos teoremas de incompletitud, socavando, así, la ilusión de una posible fundamentación de la matemática. Lo que aparentemente parecían fracasos, suscitaron, empero, grandes avances en las ciencias formales, en la medida en que establecieron los confines de los transitable.

El primer teorema de incompletitud de Gödel demostraba que cualquier teoría axiomatizable y consistente de, al menos, la misma complejidad que la aritmética no podría ser completa[2]. Este desenlace situó a los matemáticos en una tesitura amarga: si la aritmética estándar, la aritmética de Peano, es consistente (afirmación no infalible que, sin embargo, casi ningún matemático está dispuesto a rehusar[3]), entonces existen teoremas matemáticos indemostrables. De este modo, la tan deseada correspondencia para la aritmética entre el conjunto de sentencias probables[4] y el conjunto de sentencias verdaderas, correspondencia representada por la propiedad de la completitud, se tornaba imposible, poniendo de manifiesto más que nunca la independencia entre el concepto de "prueba" y el concepto de "verdad". Más allá, pues, del concepto intuitivo de verdad que manejamos en la vida cotidiana, la lógica precisaba de una elaboración exhaustiva de este concepto semántico que nos permitiera establecer desde qué condiciones podemos afirmar que una sentencia es verdadera. En su artículo de 1933 *The Concept of Truth in Formalized Languages*, Alfred Tarski llevaría a cabo esta empresa, siendo el concepto de "verdad para un lenguaje" que allí define el estandarizado desde entonces (con algunas modificaciones) dentro de la teoría de modelos. Uno de los requisitos fundamentales establecidos por Tarski para proporcionar una caracterización satisfactoria del concepto de verdad fue la separación entre lenguaje objeto, que sería aquel para cual queremos encontrar una definición de verdad, y metalenguaje, que equivaldría a aquel en el que dicha definición será formulada. Además, el metalenguaje habría de ser necesariamente más rico en recursos que el lenguaje objeto y contendría a este último como parte. Toda

[2] En su formulación original: "Theorem VI. *For every ω-consistent recursive class κ of FORMULAS there are recursive CLASS SIGNS r such that neither υ Gen r nor Neg(υ Gen r) belongs to Flg(κ) (where υ is the FREE VARIABLE of r)*" (Gödel, 1931, p.607).

[3] El segundo teorema de incompletitud de Gödel limita la posibilidad de encontrar una prueba de consistencia absoluta para la aritmética. Sin embargo, existen intentos notables, como el del matemático Gerhardt Gentzen, cuya prueba usa prácticamente en su totalidad métodos finitistas, a excepción de cierta inducción transfinita hasta el ordinal ε_0.

[4] En este contexto, el término "probable" aplicado a sentencias o teoremas debe ser entendido como "demostrable en un número finito de pasos a partir del uso de axiomas y reglas de inferencia".

búsqueda de definición de la noción de verdad que no cumpliese con este requisito se toparía en algún momento con la aparición de paradojas ligadas a la autorreferencialidad, como la célebre *paradoja del mentiroso*, y es por ello por lo que cualquier tentativa de respuesta a la pregunta acerca de la completitud de un sistema debe ser siempre llevada a cabo dentro del metalenguaje (Tarski, 1969).

Una de las consecuencias directas de todas las consideraciones anteriores queda ejemplificada a partir de la demostración realizada por Tarski de la imposibilidad de definir la noción de verdad aritmética dentro de la propia aritmética, lo que a veces se conoce como *teorema de indefinibilidad*. A través de este artículo, por tanto, trataremos de explorar las relaciones existentes entre el concepto de verdad de Tarski y el primer teorema de incompletitud de Gödel, siendo nuestra hipótesis de partida aquella que afirma que ambas cuestiones son el haz y el envés de la misma moneda: la autorreferencialidad.

2. La aproximación tarskiana al concepto de *verdad*

Nuestro uso habitual del concepto de verdad–y, *a fortiori*, el uso de su contraparte, el concepto de falsedad–consiste en su predicación acerca de oraciones declarativas, las cuales son tomadas como objetos lingüísticos. En este sentido, nuestra noción acerca del significado del término "verdad" concuerda con algunas de las nociones clásicas. Aristóteles, por ejemplo, en su *Metafísica*, proporcionaría la siguiente definición:

> Falso es, en efecto, decir que lo que es, no es, y que lo que no es, es; verdadero, que lo que es, es, y lo que no es, no es. Por consiguiente, quien diga que (algo) es o no es, dirá algo verdadero o dirá algo falso. Sin embargo, ni de lo que es ni de lo que no es puede decirse (indistintamente) que es o que no es (Aristóteles, 1994: 1011, b25).

La anterior definición responde a la noción clásica o semántica de verdad, siendo la semántica aquella parte de la lógica que estudia la relación entre los objetos lingüísticos y lo que es expresado por dichos objetos (Tarski, 1969). En este sentido, la verdad es entendida como "correspondencia" con el estado actual de las cosas: si lo que predicamos acerca de algo queda verificado por el estado real de ese algo, entonces emerge la verdad (*adaequatio rei intellectus*). Sin embargo, para Tarski, la versión aristotélica no será lo suficientemente transparente desde el punto de vista lógico, por lo que pretenderá plantear otra concepción de verdad que reemplace la aristotélica, pero que preserve, empero, sus intenciones iniciales.

En primer lugar, es importante señalar que el concepto de verdad es siempre relativo a un lenguaje. Además, si queremos definir este concepto exitosamente, habremos de aparcar el lenguaje natural y considerar tan solo sistemas formales, pues, como tendremos ocasión de poner de manifiesto a continuación, la universalidad del lenguaje natural impide construir una definición de verdad adecuada. Por el momento, consideremos la siguiente afirmación: *"La nieve es blanca" si y solo si la nieve es blanca*. La anterior sentencia es un claro ejemplo de la noción clásica de verdad como correspondencia: la oración *La nieve es blanca* solamente será verdadera si, de hecho, al contemplar la nieve, verificamos que esta es blanca. En realidad, una oración de esta índole tendría la siguiente forma lógica: *"p" es verdadera si y solo si p*. Teniendo presente lo anterior, Tarski (1969) afirmará que la construcción de una noción de verdad general es relativamente fácil. Supongamos, en primer lugar, que disponemos de una parcela del lenguaje natural (en este caso, el castellano) a la que denominamos *L*. Asimismo, *L* presenta unas reglas sintácticas lo suficientemente claras para permitirnos distinguir en todo momento qué es una sentencia en *L* y qué no lo es y, además, el conjunto de sentencias que la conforman, aunque arbitrariamente grande, es finito. Finamente, asumimos que la palabra "verdad" no ocurre en *L* y que el significado de todos los vocablos que determinan su vocabulario está suficientemente claro, de manera que podamos utilizarlos para definir la noción de verdad. Si todos los anteriores requisitos se cumplen, podemos proceder del siguiente modo:

1. Confeccionamos una lista que incluya todas las sentencias construibles en *L*. Supongamos, en pos de simplicidad, que nuestra lista incluye exactamente 10 sentencias, a las que denominamos s_1, s_2, …, s_{10}.

2. Para cada una de estas 10 sentencias, construimos una definición parcial de verdad sustituyéndolas sucesivamente por *p* en ambas partes del esquema *"p" es verdadera si y solo si p.*

3. Finalmente, constituimos la conjunción lógica de todas estas definiciones parciales y le proporcionamos una forma lógica diferente, pero equivalente, que satisfaga los requisitos formales que se imponen a las definiciones en base a las reglas de la lógica, a saber:

Para toda sentencia x de nuestro lenguaje L, x es verdadera si y solo si o bien:

s_1, y x es idéntica a "s_1", o
s_2, y x es idéntica a "s_2",
…,
o, finalmente,
s_{10}, y x es idéntica a "s_{10}".

Finalizado este procedimiento, llegamos a una sentencia final que podrá ser aceptada como una definición general de verdad en este marco: será formalmente correcta e implicará todas las equivalencias de la forma *"p" es verdadera si y solo si p* en las que *p* habría sido reemplazada por alguna sentencia de *L*. No obstante, si queremos llevar a cabo una construcción similar para cualquier lenguaje natural en su totalidad–para lo que aquí nos ocupa nos sirve el castellano–nos encontramos con varios obstáculos. En primer lugar, la gramática del castellano no determina de forma unívoca qué cadenas de caracteres pueden funcionar como sentencias. Por ejemplo, Tarski dirá (1969) que–respecto del inglés–una expresión exclamativa puede funcionar en ciertos contextos como una sentencia y no hacerlo en otros. Por otro lado, en cualquier lenguaje natural el número de sentencias que pueden ser construidas recursivamente en base a la gramática es potencialmente infinito, mientras que la construcción que se ha dado es claramente finita. Sin embargo, la objeción definitiva a que un procedimiento de este tipo pueda ser llevado a término en un lenguaje natural tiene que ver con el hecho de que la palabra "verdad" ocurra significativamente dentro del lenguaje en cuestión. Veamos, pues, qué implicaciones se derivan de esta consideración y por qué está estrechamente relacionada con la *antinomia del mentiroso*.

En primer lugar, y siguiendo a Tarski (1956; 1969), contemplemos cuál es la formulación de la *antinomia del mentiroso* por parte del lógico y filósofo polaco Jan Łukasiewicz. El argumento discurre como sigue:

1. Usemos el símbolo "o" como la abreviatura tipográfica de la expresión *la oración escrita en esta página en color rojo*.

2. Consideremos, ahora, la siguiente oración: *o no es una oración verdadera.*

3. Teniendo en cuenta el significado de "o", podemos establecer empíricamente que "*o no es una oración verdadera" es idéntica a o*.

4. Igualmente, dada la noción de verdad como correspondencia descrita arriba, podemos afirmar algo como *"o no es una oración verdadera" es una oración verdadera si y solo si o no es una oración verdadera.*

5. Finalmente, si unimos las afirmaciones de 3 y 4 tenemos que *o es una oración verdadera si y solo o no es una oración verdadera*, lo que se trata, manifiestamente, de una contradicción.

Una de las características del lenguaje natural, como hemos indicado anteriormente, es su universalidad. Además de los propios objetos lingüísticos (oraciones y términos) que componen cualquier lenguaje natural, la propia naturaleza de este tipo de lenguajes nos permite disponer también de nombres para estos objetos; por otra parte, también podemos encontrar términos puramente semánticos (como "verdad") que se refieren a las relaciones entre los objetos lingüísticos y lo que es expresado por ellos. Estas propiedades de lenguaje natural nos permiten crear oraciones como la presente en 2, oraciones autorreferenciales que dan lugar a la *antinomia del mentiroso*. Ante la perplejidad que nos pueda producir una paradoja de este tipo, podríamos pensar que un uso tal del lenguaje (un uso en el que se permitan oraciones autorreferenciales) no puede ser tolerado, pero, ciertamente, que cierto fenómeno sea peculiar no es razón suficiente para apartarlo:

> La primera reacción natural ante la paradoja del mentiroso es achacar la contradicción al hecho de que el enunciado hace referencia a sí mismo, y establecer, consecuentemente, como un principio básico del uso correcto del lenguaje, que no puede ser construido un enunciado, o, al menos, ser seriamente discutido, si hace referencia a sí mismo. Sin embargo, el mero hecho de la auto-referencia, aunque parezca extraño, no puede proporcionarnos una explicación satisfactoria de la aparición de nuestra paradoja. De hecho, existen otros casos en los que la auto-referencia es totalmente inofensiva. Tomemos, por ejemplo, la frase:
>
> *El enunciado, hecho en el artículo "Las Paradojas de la Lógica" de Evert W. Beth, sección 5, y que dice "El enunciado…a sí mismo", se refiere a sí mismo.*
>
> Es obvio que este enunciado debe ser considerado como verdadero (Beth, 1975: 15 – 16).

La universalidad del lenguaje natural, por tanto, arruina cualquier aspiración de encontrar un concepto general de verdad. Sin embargo, podemos preguntarnos si una noción así puede ser construida para los lenguajes científicos, en concreto, para los lenguajes formales. Tarski (1956; 1969) afirmará que esta tarea es factible siempre que se cumplan las siguientes condiciones: en primer lugar, el vocabulario al completo de un lenguaje de este tipo debe de estar disponible y las reglas sintácticas para la formación de oraciones a partir de los componentes del vocabulario deben estar formuladas de manera precisa. Asimismo, estas reglas sintácticas deben ser puramente formales y la función y el significado de cada expresión debe depender exclusivamente de su forma lógica. Finalmente, y quizás esta sea la principal contribución de Tarski respecto de este asunto, habrá de realizarse una distinción clara entre lenguaje objeto y metalenguaje. El metalenguaje que usemos para analizar el lenguaje objeto habrá de ser lo suficientemente rico e incluirá al lenguaje objeto como parte, y dispondrá de términos adicionales para nombrar las expresiones del lenguaje objeto y para definir conceptos semánticos como el de "verdad":

> In contrast to natural languages, the formalized languages do not have the universality which was discussed at the end of the preceding section. In particular, most of these languages possess no terms belonging to the theory of language, i.e. no expressions which denote signs and expressions of the same or another language or which describe the structural connexions between them (such expressions I call -for lack of a better term- *structural-descriptive*). For this reason, when we investigate the language of a formalized deductive science, we must always distinguish clearly between the language about which we speak and the language in which we speak, as well as between the science which is the object of our investigation and the science in which the investigation is carried out. The names of the expressions of the first language, and of the relations between them, belong to the second language, called the *metalanguage* (which may contain the first as a part). The description of these expressions, the definition of the complicated concepts, especially of those connected with the construction of a deductive theory (like the concept of consequence, of probable sentence, possibly of true sentence), the determination of the properties of these concepts, is the task of the second theory which we shall call the *metatheory* (Tarski, 1956: 167).

Una vez estipuladas estas condiciones de partida, Tarski tratará de responder a la pregunta fundamental que guía su artículo de 1933: cómo construir una noción adecuada de verdad. Como hemos indicado ya, cualquier definición de verdad es relativa a un lenguaje y, dado que el lenguaje elegido por Tarski es el cálculo de clases, la investigación será llevada a cabo en el lenguaje que podemos llamar *metacálculo de clases*. En primera instancia, podríamos pensar que la siguiente definición valdría para una definición general del concepto de verdad:

> DEFINITION 17. *x is a provable (accepted) sentence or a theorem – in symbols* $x \in Pr$- *if and only if x is a consequence of the set of all axioms* (Tarski, 1933: 182).

Esta definición presenta la ventaja de ser puramente estructural, en la medida en que su construcción depende exclusivamente de los axiomas del metalenguaje. Sin embargo, Tarski se percata de que esta caracterización no se corresponde con la idea intuitiva de "sentencia verdadera", pues en el dominio de las sentencias probables, el principio del tercio excluso no es válido, mientras que nuestra noción cotidiana de verdad precisa de la validez de tal principio. Es decir, se puede dar el caso de que sea imposible probar tanto una sentencia como su negación[5]. Si el dominio de las sentencias verdaderas no es coextensivo con el dominio de las sentencias probables–algo que sabemos que es cierto (por el primer teorema de incompletitud de Gödel) para cualquier teoría axiomatizable de, al menos, la misma complejidad que la aritmética (criterio que cumple el cálculo de clases)–, entonces nuestro concepto de verdad deberá cubrir también aquellas sentencias que, aún siendo verdaderas, no pueden ser probadas.

La solución de Tarski pasará por aproximarse a la verdad desde un punto de vista semántico y no estrictamente formal. Para ello, el lógico polaco introduce su célebre *Convención T*, por la cual establece qué condiciones ha de cumplir una definición adecuada de verdad. Denote el símbolo "S" la clase de las sentencias, en general, y el símbolo "Tr" la clase de las sentencias verdaderas:

> CONVENTION T: *A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an adequate definition of truth if it has the following consequences:*
>
> *(*α*) all sentences which are obtained from the expression '*$x \in Tr$ *if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expressoin which forms the translation of this sentence in the metalanguage;*
>
> *(*β*) the sentence 'for any x, if* $x \in Tr$ *then* $x \in S$ *(in other worlds, 'Tr* $\subseteq S$*')* (Tarski, 1933:188).

En realidad, la *Convención T* no es sino una formulación precisa del concepto semántico de verdad al que hemos tenido ocasión de referirnos: (α) la oración *La nieve es blanca* es verdadera si y solo si, de hecho, la nieve es blanca[6] y (β) toda oración verdadera es una

[5] Para ejemplos concretos, véase el artículo original de Tarski.
[6] Podemos observar cómo el significado (o traducción, si se quiere) de la oración "La nieve es blanca"–esto es, las condiciones materiales que verifican que la nieve sea, de hecho, blanca–se expresa en el metalenguaje.

oración[7]. Para un lenguaje que contenga un número de sentencias finitamente enumerable, una construcción del concepto de verdad similar a la explicitada en la tercera página de este artículo satisfaría, sin mayor problema, las condiciones de la *Convención T*. No obstante, para un lenguaje con un número infinito de sentencias, un procedimiento de este tipo, por razones obvias, no podrá ser llevado a cabo, por lo que la definición de verdad que proponga Tarski para este tipo de lenguajes será subsidiaria de la noción de "satisfacibilidad".

La gran virtud de la satisfacibilidad es que su definición admite recursividad y, en este sentido, se trata de una definición composicional. Una definición de satisfacibilidad podría ser la siguiente (Hodges, 2018): "*a" satisface la fórmula "F" si y solo si, al considerar cada una de las ocurrencias libres de las variables en "F" como nombres del objeto asignado a ellas por "a","F" se torna verdadera*. En el caso de que "F" fuera una fórmula compleja, por ejemplo, bastaría con considerar qué asignaciones satisfacen cada uno de los componentes de "F"[8]. Finalmente, la deducción de la definición de verdad a partir de la noción de satisfacción es relativamente sencilla: si una fórmula "F" no contiene ocurrencias libres de sus variables, entonces, o bien todas las asignaciones la satisfacen–por lo que "F" sería verdadera–o ninguna lo hace–por lo que "F" sería falsa–.

Grosso modo, se ha esbozado la caracterización que hace Tarski del concepto de verdad y las dificultades con las que se encuentra. Consideramos, no obstante, que el abordaje de un último resultado es pertinente para preludiar la sección ulterior: nos referimos al *teorema de indefinibilidad*. Podemos enunciar este teorema tal y como sigue (Boolos, G.S., Burgess, J.P & Jeffrey, R.C, 2002: 223): *el conjunto de los números de Gödel para sentencias del lenguaje de la aritmética que son correctas–o verdaderas en la interpretación estándar–no es aritméticamente definible.* Su prueba es relativamente sencilla si, primero, introducimos el siguiente lema: *Sea T una teoría consistente que extiende* ***Q*** *(la aritmética de Robinson). Entonces, el conjunto de los números de Gödel asignados a los teoremas de T no es definible en T.* Sea la prueba de este lema:

> *Proof*: Let *T* be an extension of ***Q***. Suppose $\theta(y)$ defines the set Θ of Gödel numbers of sentences in *T*. By the diagonal lemma there is a sentence *G* such that

[7] Como el propio Tarski advierte, la condición β no es esencial. En la medida en que nuestro metalenguaje dispone del símbolo "Tr" y este satisface la condición α, podemos definir un nuevo símbolo, "Tr'" que satisfaga la condición β: "Tr'" representará la intersección entre "Tr" y "S" (Tarski, 1933).

[8] Esta noción de satisfacibilidad es la empleada, por poner ejemplo, en la demostración del *Lema de Verdad* en la prueba de completitud para la lógica de primer orden de L. Henkin.

$$\vdash_T G \leftrightarrow \sim\theta(\ulcorner G \urcorner).$$

In other words, letting g be the Gödel number of *G*, and g its Gödel numeral, we have

$$\vdash_T G \leftrightarrow \sim\theta(\mathbf{g}).$$

Then *G* is a theorem of *T*. For if we asume *G* is not a theorem of *T*, then *g* is not in Θ, and since θ(y) defines Θ, we have $\vdash_T G \leftrightarrow \sim\theta(\mathbf{g})$; but then since $\vdash_T G \leftrightarrow \sim\theta(\mathbf{g})$, we have $\vdash_T$ *G* and *G* is a theorem of *T* after all. But since *G* is a theorem of *T*, *g* is in Θ, and so we have $\vdash_T \theta(\mathbf{g})$; but then, since $\vdash_T G \leftrightarrow \sim\theta(\mathbf{g})$, we have, $\vdash_T G$, and *T* is inconsistent (Boolos, G.S., Burgess, J.P & Jeffrey, R.C, 2002: 223).

Básicamente, la prueba que hemos visto para el anterior lema hace uso de la diagonalización para encontrar una fórmula autorreferencial que nos conducirá a contradicción. Si suponemos que la extensión de ***Q*** a la que hemos llamado *T* es la aritmética (lo cual es posible al ser *la aritmética* una extensión consistente de ***Q***) tenemos que la indefinibilidad a la que nos referíamos no es otra que la indefinibilidad en la aritmética, obteniendo así el teorema de Tarski. Este resultado es, en realidad, una consecuencia inmediata para la aritmética de Peano del primer teorema de incompletitud de Gödel, siendo tan solo un caso particular en el que el concepto de verdad está explícitamente involucrado. Tendremos ocasión de volver a este teorema al final de este este artículo, pero, con ánimo de obtener una comprensión más completa del mismo, abordemos antes algunas cuestiones relacionadas con el primer teorema de incompletitud de Gödel.

3. Aritmetizar la metamatemática: el primer teorema de incompletitud de Gödel

Hemos presentado en la sección anterior algunos de los grandes avances llevados a cabo por Tarski respecto a la semántica de los sistemas formales, pero retrocedamos ahora algunos años atrás. En 1930, el matemático y mentor de Kurt Gödel, Hans Hahn, presentaría en la Academia de las Ciencias de Viena un adelanto de los resultados alcanzados por su discípulo respecto a la consistencia y completitud de ciertos sistemas axiomáticos recursivamente enumerables, resultados que serían publicados en detalle por el propio Gödel en el artículo de 1931 *On formally Undecidable Propositions of Principia Mathematica and Related Systems*. La conclusión a la que arriba Gödel, como sabemos, afirma que la consistencia y la completitud de cualquier teoría axiomatizable de la misma complejidad que la aritmética no pueden ser probadas simultáneamente. El procedimiento a través de cual Gödel consigue demostrar la anterior afirmación, además de rebosar ingenio, está estrechamente relacionado con los temas que hemos tratado hasta el

momento. Él mismo reconoce en la primera sección de su artículo (Gödel, 1931) que su razonamiento es similar al seguido en la *antinomia del mentiroso* o en la *paradoja de Richard*[9]; si en el caso de la *antinomia del mentiroso* llegábamos a un enunciado autorreferencial que predicaba su propia falsedad, Gödel tratará de probar que en cualquier teoría de los rasgos anteriormente mencionados podemos encontrar un enunciado similar -a partir de entonces denominado enunciado "G"- que afirme su propia indemostrabilidad. Si el enunciado "G" fuese falso, entonces -por su propio significado- podríamos demostrarlo, lo que significaría que nuestro conjunto de axiomas permite probar enunciados falsos y, por ende, no es consistente. Si queremos, en cambio, preservar la consistencia, entonces "G" habrá de ser verdadero, teniendo, así, que es indemostrable:

> By the construction of A,
>
> (1) A means that A is unprovable
>
> Let us asume, as we hope is the case, that formulas which express false propositions are unprovable in the system, i.e.
>
> (2) false formulas are unprovable
>
> Now the formula A cannot be false, because by (1) that would mean that it is not unprovable, contradicting (2). But A can be true, provided it is unprovable. Indeed this must be the case. For assuming that A is provable, by (1) A is false, and hence by (2) unprovable. By (intuitive) reductio ad absurdum, this gives that A is unprovable, whereupon by (1) A is also true. Thus the system is incomplete in the sense that it fails to afford a proof of every formula which is true under the interpretation (if (2) is so, or if at least the particular formula A is unprovable if false).
>
> The negation ¬A of the formula is also unprovable. For A is true; hence ¬A is false; and by (2), ¬A is unprovable. So the system is incomplete also in the simple sense defined metamathematically in the last section (if (2) is so, or if at least the particular formulas A y ¬A are each unprovable is false) (Kleene, 1974: 205).

El resultado de Gödel es sorprendente en la medida en que, en primera instancia, lo natural es pensar que toda cuestión matemática que pueda ser expresada dentro de un sistema puede ser, al mismo tiempo, dilucidada en tal sistema. Recordemos que habíamos dicho con Tarski que un concepto semántico como el de verdad siempre ha de ser definido en

[9] Sea *S* el conjunto que consta de todos los enunciados que contienen un número finito de palabras y que definen un número real *r(0 < r < 1)* y sea *T* el conjunto de todos los números reales definidos por enunciados de *S*. Consideremos, ahora, el siguiente enunciado: "el número de Richard, ρ, es el número real que resulta de aplicar el procedimiento de la diagonal de Cantor al conjunto enumerable infinito *T* de los números reales *r* que satisfacen la condición $0 < r < 1$ y que pueden ser definidos mediante enunciados que contienen un número finito de palabras". Si, como dice el enunciado, ρ es el resultado de aplicar diagonalización al conjunto *T*, entonces -por la propia naturaleza del método de la diagonal- ρ no puede estar incluido en *T*; sin embargo, el enunciado anterior es una definición correcta del número de Richard y, por tanto, deberíamos concluir que ρ pertenece a *T*, llegando así a una contradicción (Cfr. Beth, 1975:17 – 18).

un lenguaje jerárquicamente superior a aquel para el cual queramos definir dicho concepto; pues bien, el logro de Gödel consiste en mostrar mediante la construcción de la prueba para el primer teorema de incompletitud que el metalenguaje para un sistema formal que pretenda formalizar la matemática–la metamatemática–es susceptible de ser reconstruido en términos de ese propio sistema. Teniendo en mente lo que decíamos a propósito de la universalidad del lenguaje natural, podríamos decir ahora que lo que le ocurre a este tipo de sistemas es que su propia pretensión de universalidad actúa en detrimento de ellos mismos.

El sistema para el que Gödel demuestra su teorema–al que, a partir de ahora, denominaremos *S*–está constituido por los axiomas de Peano más la lógica de los *Principia Mathematica* (PM) de Whitehead y Russell, así como por el axioma de elección (Gödel, 1931), pero vale igualmente para cualquier sistema de características similares. Como hemos advertido, la demostración del teorema requiere de la representación del metalenguaje en términos de la aritmética, por lo que esta será la primera tarea a realizar. En primer lugar, se establece una correspondencia uno a uno entre los signos primitivos de *S* y ciertos números naturales (en este caso, los números primos hasta 13) del siguiente modo:

"0" …1
"f" …3
"~"…5
"v"…7
"Π"…9
"("…11
")"…13

Igualmente, a cada secuencia de signos primitivos que dé lugar a una fórmula bien formada se le asigna un número único, así como a las pruebas (que, al fin y al cabo, son secuencias finitas de símbolos) llevadas a cabo dentro del sistema. A este proceso se lo conoce como "numeración de Gödel" y ha influido considerablemente en disciplinas como la criptografía. Sea, ahora, el predicado *A(a,b)*, donde *a* es el número de Gödel de una fórmula ("A_a(a)") y *b* el número de Gödel de la prueba de la fórmula "A_a(**a**)", y sea el predicado *B(a,c)*, donde *a* es el número de Gödel de una fórmula ("A_a(a)") y *c* el

número de Gödel de la prueba de la fórmula "$\neg A_a(\mathbf{a})$"[10], consideremos el siguiente lema (Kleene,1974): "Existe una numeración de Gödel para los objetos formales de nuestra teoría de manera que los predicados *A(a,b)* y *B(a,c)* son expresables aritméticamente en *S*". Tomemos ahora a "A(a,b)" y "B(a,c)" como las fórmulas ya aritmetizadas que expresan los predicados *A(a,b)* y *B(a,c)*. A continuación, consideremos la fórmula "$\forall b \neg A(a,b)$" que contiene solo la variable libre *a* y llamemos ρ a su número de Gödel; entonces -siguiendo el razonamiento anterior- podemos llamar a esta fórmula "$A_\rho(a)$". El próximo paso consiste en utilizar el lema de diagonalización[11] para obtener la fórmula "$\forall b \neg A(\boldsymbol{\rho},b)$" -a la que llamaremos "$A_\rho(\boldsymbol{\rho})$"-, la cual no contiene ninguna variable libre. Podemos decir que la fórmula "$A_\rho(\boldsymbol{\rho})$" expresa que "$A_\rho(\boldsymbol{\rho})$" no es demostrable en *S*, es decir, es la fórmula "G" que estábamos buscando y que afirma su propia indemostrabilidad.

Tenemos, pues, que la falsedad de la fórmula "G" no puede ser probada en base a la presunción de consistencia de nuestro sistema. No obstante, si queremos demostrar que "¬G" tampoco puede ser probada, necesitamos estipular una condición adicional para *S*, condición que Gödel denominará ω-consistencia. Examinemos la propia definición de Gödel de la ω-consistencia:

> Here a system is said to be ω-consistent if, for no property F(x) of natural numbers,
>
> $$F(1), F(2), \ldots, F(n), \ldots \text{ ad infinitum}$$
>
> as well as
>
> $$\overline{(Ex)F(x)}$$
>
> are probable. (There are extensions of the system S that, while consistent, are no ω-consistent) (Gödel, 1931:596).

Obviamente, todo sistema ω-consistente es también consistente, pero la afirmación inversa no es cierta[12]. Teniendo en cuenta lo anterior, ya estaríamos en condiciones de

[10] Se ha elegido convencionalmente utilizar tipografía en cursiva para escribir las fórmulas del lenguaje objeto, las comillas para enmarcar las fórmulas del metalenguaje y la tipografía en negrita para señalar los números de Gödel asociados a las pruebas de fórmulas.

[11] Podemos enunciar el *lema de diagonalización* como sigue: *Sea A(x) una fórmula arbitraria del lenguaje de S (el cual contiene a* ***Q****, es decir, a la aritmética de Robinson) con una sola variable libre. Entonces, una sentencia D puede ser mecánicamente construida de tal modo que* $F \vdash D \leftrightarrow A(\ulcorner D \urcorner)$. A menudo, este lema es llamado el *lema de autorreferencialidad* o el *lema del punto fijo* (Cfr. Raatikainen, 2020).

[12] En 1936, J. B. Rosser consiguió extender el resultado de Gödel para sistemas solamente consistentes (y no ω-consistentes).

probar la indemostrabilidad de "¬G"; observemos la presentación de la prueba por parte de Kleene:

> THEOREM 28. *If the number theoretic formal system is (simply) consistent, then not* ⊢"$A_\rho(\boldsymbol{\rho})$"; *and if the system is ω-consistent, then not* ⊢"$\neg A_\rho(\boldsymbol{\rho})$". *Thus, if the system is ω-consistent, then it is (simply) incomplete, with "*$A_\rho(\boldsymbol{\rho})$*" as an example of an undecidable formula.*
>
> PROOF that, if the system is not consistent, then not ⊢ "$A_\rho(\boldsymbol{\rho})$". Suppose (for intuitive reductio ad absurdum) that ⊢ "$A_\rho(\boldsymbol{\rho})$", i.e. suppose that "$A_\rho(\boldsymbol{\rho})$" is provable. Then, there is a proof of it; let the Gödel number of this proof be κ. Then, *A(ρ,κ)* is true. Hence, since "A(a,b)" was introduced under the lemma as a formula which numeralwise expresses *A(a,b),* ⊢ "A(**ρ,κ**)". By ∃-introd., ⊢ "∃b A(**ρ,**b)". Thence, by *83a[13], ⊢¬∀b¬A(ρ,b). But this is ⊢ "$\neg A_\rho(\boldsymbol{\rho})$". This, with our assumption that ⊢"$A_\rho(\boldsymbol{\rho})$" contradicts the hypothesis that the system is consistent. Therefore, by reductio ad absurdum, not ⊢ "$A_\rho(\boldsymbol{\rho})$", as was to be shown. (We could also have contradicted the consistency by using ∀-elim. to infer ⊢ "¬A(**ρ,κ**)" from "$A_\rho(\boldsymbol{\rho})$".)
> PROOF that, if the system is ω-consistent (and hence also consistent), then not ⊢ "$\neg A_\rho(\boldsymbol{\rho})$". By the consistency and the first part of the theorem, "$A_\rho(\boldsymbol{\rho})$" is not provable. Hence each of the natural numbers 0, 1, 2, … is not the Gödel number of a proof of "$A_\rho(\boldsymbol{\rho})$"; i.e. *A(ρ,0)*, *A(ρ,1)*, *A(ρ,2)*, … are all false. Hence, since "A(a,b)" numeralwise expresses *A(a,b)*, ⊢ "¬A(**ρ**, 0)", ⊢ "¬A(**ρ**, 1)", ⊢ "¬A(**ρ**, 2)", … By the ω-consistency , then not ⊢ "¬∀b¬A(**ρ**, b)". But it is not "$\neg A_\rho(\boldsymbol{\rho})$", which was to be shown (Kleeene, 1974:207 – 208):

Evidentemente, la prueba de Gödel es más extensa y reviste mayor complejidad, pero, para los que aquí nos interesa, un esquema como el anterior parece adecuado. Podríamos decir, quizás coloquialmente, que lo que Gödel ha demostrado es que la *antinomia del mentiroso* puede ser reconstruida en el lenguaje de la aritmética, lo que implica que siempre van a existir en la matemática enunciados verdaderos que no pueden ser probados, aunque en la práctica real de la disciplina nunca nos encontremos con ellos. Podríamos pensar, sin embargo, que una posible solución para la indecidibilidad de "G" pasa por crear una nueva teoría aritmética axiomatizable que añadiese "G" como axioma, i.e.: $S \cup \{\text{“}G\text{”}\}$. Sin embargo, a esta nueva teoría también se lo podría aplicar el teorema de Gödel, de modo que encontraríamos una nueva fórmula "G_1" indecidible en ella. Por supuesto, este procedimiento se podría iterar hasta el infinito, obteniendo, así, nuevas –$S \cup \{\text{“}G\text{”}\} \cup \{\text{“}G_1\text{”}\}$, $S \cup \{\text{“}G\text{”}\} \cup \{\text{“}G_1\text{”}\} \cup \{\text{“}G_2\text{”}\}$, etc.–, todas ellas extensiones axiomatizables de *S* y cada una mejor que la anterior.

4. Consideraciones finales

[13] *83a. ⊢∃xA(x) ⊃ ¬∀x ¬A(x) (Kleene, 1974:163).

No deja de ser asombroso que una paradoja que, como mínimo, se remonta al siglo IV a.C. haya inspirado o, incluso, propiciado algunos de los más grandes resultados matemáticos de la historia. Las antinomias, desde nuestro punto de vista, lejos de suponer un óbice para el pensamiento, forman parte inextricable del mismo y puede considerarse que actúan, a menudo, como revulsivo del ingenio. Desde esta perspectiva, la búsqueda de paradojas puede ser considerada, quizás, algo deseable, en la medida en que, en muchas ocasiones, son las únicas que pueden proporcionarnos una base firme sobre la que edificar el pensamiento.

En esta línea pueden ser interpretados los resultados de Tarski y Gödel. El primer teorema de incompletitud nos enseña que en cualquier sistema consistente con la suficiente capacidad expresiva como para poder hablar acerca de sí mismo, podemos encontrar una sentencia indemostrable. En realidad, la demostración de Gödel es puramente sintáctica y tiene que ver más con el concepto de decidibilidad que con el de verdad, ya que en ningún momento se invoca explícitamente esta última noción. En este sentido, la conclusión del primer teorema de incompletitud no afirma la existencia de verdades indemostrables, sino de sentencias indecidibles, y será Tarski quien, partiendo de las bases sentadas por Gödel, enriquezca los resultados del matemático austriaco al tratar de definir una noción semántica de verdad aplicable a los lenguajes formales[14]. Aunque es cierto que habitualmente se presentan los resultados de Gödel como relativos a la verdad de ciertas sentencias, es importante tener en cuenta la apreciación que hacemos, ya no solo por otorgar a cada autor el reconocimiento que se merece, sino también porque conceptos como "verdad" y "prueba" o "cálculo" y "semántica", aunque íntimamente vinculados entre sí, han de ser diferenciados.

Finalmente, nos parece interesante recalcar cómo la potencia de un determinado lenguaje formal, o su capacidad recursiva, influye decisivamente en su decidibilidad. Los lenguajes formales, siguiendo la metáfora que hace Frege en el prólogo a su *Conceptografía* (1879), no podrán nunca sustituir al lenguaje natural en su versatilidad y capacidad expresiva–como un microscopio no puede sustituir al ojo humano–pero sirven para analizar detalladamente cuestiones que de otro modo no podrían ser tratadas. En realidad, tanto lo demostrado por el primer teorema de incompletitud como por el teorema de indefinibilidad para la aritmética es análogamente aplicable al lenguaje natural. Cabría preguntarse, entonces, si la aspiración a la universalidad y a cubrir la mayor extensión de

[14] Desde esta perspectiva, el vínculo de los aportes de Tarski con la *antinomia del mentiroso* sería más fuerte que el vínculo con los teoremas de incompletitud de Gödel.

pensamiento posible de estos lenguajes no los acerca demasiado al lenguaje natural y, por ende, no los hace prescindibles.

5. Bibliografía

Aristóteles. (1994). *Metafísica.* Traducción de Tomás Calvo Martínez. Madrid: Editorial Gredos.

Beth, E. V. (1975). *Las paradojas de la lógica*. Presentación, versión al castellano y notas por Juan Manuel Lorente. Valencia: Departamento de Lógica y Filosofía de la Ciencia, Universidad de Valencia.

Boolos, G.S., Burgess, J.P & Jeffrey, R.C. (2002). *Computability and Logic*. New York, NY: Cambridge University Press.

Frege, G. (1879/1967). 'Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought'. In Heijenoort, v. J. (Ed.). *From Frege to Gödel.* Cambridge, MA: Harvard University Press.

Gödel, K. ([1931] 1967). "On formally Undecidable Propositions of Principia Mathemática and Related Systems". En van Heijenoort, J. (Ed.), *From Frege to Gödel. A Source Book in Mathematical Logic, 1879 – 1931*. Cambridge, MA: Harvard University Press, pp. 596 – 617.

Hodges, W. (2018). "Tarski's Truth Definitions". En Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Otoño 2018 Ed.). Recuperado de https://plato.stanford.edu/entries/tarski-truth/

Kleene, S.C. (1974). *Introduction to Metamathematics*. New York, NY: American Elsevier Publishing Company, INC.

Raatikainen, P. (2020). "Gödel's Incompleteness Theorems". En Zalta, E.N. (Ed)., *The Stanford Encyclopedia of Philosophy* (Verano 2020 Ed.). Recuperado de https://plato.stanford.edu/entries/goedel-incompleteness/#AriForLan

Tarski, A. [1933] (1956). "The Concept of Truth in Formalized Languages". En *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Translated by J.H. Woodger. Oxford: Oxford Clarendon Press, pp. 152 – 278.

Tarski, A. (1969). "Truth and Proof". *Scientific American*, 220(6): 63 – 77.