UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN FILOSOFÍA
FACULTAD DE FILOSOFÍA Y LETRAS
INSTITUTO DE INVESTIGACIONES FILOSÓFICAS
LÓGICA, FILOSOFÍA DEL LENGUAJE Y DE LA MENTE

ARGUMENTOS GÖDELIANOS: CLASIFICACIÓN Y CONDICIONES DE EVALUACIÓN

TESIS QUE PARA OPTAR POR EL GRADO DE:
MAESTRÍA EN FILOSOFÍA

PRESENTA:
VÍCTOR MANUEL PERALTA DEL RIEGO

NOMBRE DEL TUTOR
DR. JESÚS RAYMUNDO MORADO ESTRADA DEL INSTITUTO DE INVESTIGACIONES FILOSÓFICAS DE LA UNAM

MÉXICO, D. F. , JUNIO, 2016

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Maestría en Filosofía
Lógica, Filosofía del Lenguaje y de la Mente

Argumentos Gödelianos: Clasificación y Condiciones de Evaluación

DEFENSA DE TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN FILOSOFÍA

PRESENTA:
VÍCTOR MANUEL PERALTA DEL RIEGO

TUTOR
DR. JESÚS RAYMUNDO MORADO ESTRADA

LECTORES
DR. AXEL ARTURO BARCELÓ ASPEITIA
DR. MARIO GÓMEZ TORRENTE
DR. CRISTIAN GUTIÉRREZ RAMÍREZ
DR. LUIS ESTRADA GONZÁLEZ

CIUDAD DE MÉXICO, 2018

ARGUMENTOS GÖDELIANOS: CLASIFICACIÓN Y CONDICIONES DE EVALUACIÓN

ÍNDICE

0. Introducción general

En esta tesis voy a defender un tipo de argumentos gödelianos de algunos ataques comunes. Los argumentos gödelianos (AGs, en adelante) tratan de mostrar que, al menos ciertas mentes humanas o ciertos aspectos de la mente humana no son máquinas de Turing o no pueden ser descritas por una máquina de Turing. La tesis es de nivel general, es decir, trata de abarcar todos los AGs, y no sólo algunos. Esta característica hace que sacrifique alguna precisión para conseguir generalidad, dentro de la discusión. Este nivel de generalidad servirá para, en caso de que yo esté en lo correcto, ofrecer algún AG con mejores posibilidades de ser correcto; quizá no alguno de los que propone Penrose, quizá no el que propone Putnam, quizá no el de Bringsjord, pero alguno en el espacio de los AGs posibles. Proponer un argumento gödeliano y defenderlo de las críticas más justas es un trabajo para otro momento. Por ahora me enfocaré en poner algo de orden en la discusión sobre AGs.

Para conseguir estos objetivos ofrezco una clasificación de los AGs con base en la justificación idónea de las premisas que los componen. Los AGs cuyas premisas pueden justificarse sólo por medio de razones, evidencia o teorías que están dentro de los límites de la Turing computabilidad, los llamo AGs Turing testeables, a los AGs cuyas premisas pueden justificarse con recursos superiores a los Turing computables, los llamo AGs no-Turing testeables. Aquí la primera dicotomía de la clasificación. Cabe resaltar que el sentido en el que uso Turing-testeabilidad es una extensión del mismo que puede ser polémica: se admite no sólo evidencia de conducta verbal, sino cualquier tipo de conducta que sea susceptible de ser deslindada por medio de investigación empírica regular usando teorías computables. Una vez explicada la naturaleza de los AGs y su clasificación, hago una caracterización de la noción *computabilidad* y cómo se puede relacionar a la mente humana. Esta relación es de básicamente dos tipos: el computabilismo metafísico y el computabilismo cognitivo. El computabilismo metafísico dice que la mente humana *es* literalmente una máquina de Turing, es decir, un sistema de estados finitos con relaciones de *input* y *output* claramente estblecidas. El computabilismo cognitivo en cambio no dice que la mente humana sea una máquina de Turing, sino que es *completa y correctamente* descriptible por medio de una teoría Turing computable. Así resulta una cantidad de cuatro tipos de AGs: AG Turing testeables, no-Turing testeables, metafísicos y cognitivos. En el último capítulo señalo algunas condiciones de refutación de los AG Turing Testeables, y cómo esa crítica no alcanza a refutar a los AGs no-Turing testeables sin incurrir

en cierta circularidad viciosa. De este modo establezco cómo, atendiendo a detalles internos de la naturaleza de la incompleción gödeliana, el mero conocimiento de que si tuviéramos una teoría consistente de la mente humana, entonces esta teoría sería incompleta de modo que no podríamos conocer el estatus epistémico de ciertas afirmaciones que de hecho sabemos que son ciertas. Defiendo que de aquí que es implausible que esta teoría o los objetos a los que refiere sean incoherentes, lo único que nos quedaría es admitir que estas verdades que de hecho conocemos inclinan la balanza fuertemente hacia la idea de que somos capaces de pensar con recursos intelectuales que superan los límites de la Turing computabilidad. Veamos algunas peculiaridades más de la tesis.

Una advertencia de lectura. Muchos de los términos que uso aquí son traducciones más o menos cognadas de los originales en inglés o alemán que aparecen en los artículos y libros especializados. Las referencias están al fin de cada capítulo de modo que en caso de que esta situación sea un obstáculo serio de comprensión, el lector pudiera revisar los textos originales.

En esta tesis de maestría doy algunas razones a favor de la solidez de un tipo de argumentos Gödelianos (AGs, en adelante). Las razones son de dos tipos: directas e indirectas. Las razones indirectas tienen que ver con elementos de juicio sobre la interpretación de los AGs, la selección de la mejor formulación de un AG, la ponderación de los AG que usan la incompleción de Gödel propiamente, contra aquellos que usan el *halting problem*, con la aplicabilidad de la incompleción de Gödel en agentes epistémicos específicos, y así. Las razones indirectas se vierten en el capítulo 4, principalmente consisten en bloquear un tipo común de objeción que presupone que el finitismo computabilista al respecto de la mente humana es correcto. Deshaciéndonos de este supuesto en el que se basan los mejores críticos de AGs entonces podemos evaluar sin prejuzgar las condiciones de solidez de tales argumentos. Las razones directas son simplemente las premisas de los AGs y las maneras en las que se presentan. Por ejemplo, si las premisas de un AG específico son axiomas la veracidad evidente de las mismas es una defensa tanto de la conclusión del AG, como la fuerza del AG, dado que sea válido el argumento.

Algunos esperarán encontrar en esta tesis un repaso de los textos clásicos de esta materia. Retomar a Lucas y Penrose por decir algunos de los más famosos. Pero escogí no retomar argumentos clásicos en la presente tesis. Las razones son básicamente dos: esteno es un trabajo de interpretación de un autor específico, tanto como de tipos de argumentos en general, y

segundo, las ideas que están en este trabajo pueden aplicar, dependiendo de la discusión, a las propuestas de los principales autores sobre el tema de modo que me tomaría una cantidad de espacio y trabajo más ubicar a tales argumentos en alguna de las clasificaciones, cuando ni siquiera es el objetivo primordial del trabajo.

La motivación filosófica me parece notoria. Si algún AG es sólido, entonces el computabilismo en filosofía de la mente es falso. Dicho de otro modo más coloquial, si algún AG es sólido esto implica que las mentes humanas son fundamentalmente más poderosas que las máquinas de estados finitos en particular so más poderosas que *la máquinas de Turing* (o la Máquina Universal de Turing). Esta consecuencia es filosóficamente interesante dado el auge por el que atraviesan los métodos computacionales de análisis, explicación y modelaje de la mente humana. También se podría concluir sin demasiada resistencia que, de hecho, tenemos acceso a recursos intelectuales distintos de poder superior a los de una máquina de estados finitos, con aplicaciones interesantes para la epistemología. No sobra decir que el propio Gödel murió tratando de sacar provecho de la metodología conocida como fenomenología de Husserl para fundamentar el conocimiento matemático. Penrose en su texto *Why New Physics is Needed to Understand the Mind* afirma que la próxima gran revolución científica no vendrá de la física sino de la psicología. Todos estos elementos justifican plenamente el interés filosófico de indagar en los fundamentos de la psicología, y en particular en los fundamentos teóricos de las teorías psicológicas más exitosas de hoy en día, que, como veremos abajo, tienen un sesgo computacional marcadamente alto. Este es un debate fascinante.

Al respecto de lo anterior haré un comentario personal: probablemente he leído y pensado en casi todo lo que se ha escrito sobre el tema, libros, artículos técnicos y de divulgación, foros de discusión, blogs (sobre todo los más famosos) y temas adyacentes (como son fundamentos de aritmética, de las ciencias computacionales, psicología, inteligencia artificial, biología, física, y demás) y el tono con el que los diversos autores suelen escribir al respecto de esta clase de argumentos es casi siempre tajante y aún así, notablemente polémico. Por ejemplo, Feferman dice en *Gödel, Nagel, Minds and Machines* nos cuenta que Nagel y Gödel tuvieron un desencuentro especialmente fuerte porque Gödel quería tener el derecho a sugerir cambios en el libro de famoso de Nagel y Newman (Gödel's Proof, 1958). Ya que estamos con Feferman, en *Penrose's Gödelian Argument* (1995) lo siguiente:

> Así, ahora Penrose ha caminado una gran distancia en SM [*Shadows of the Mind*] para poner su argumento gödeliano contra toda posible objeción. Debo decir que aunque los teoremas de incompleción de Gödel están entre los teoremas más importantes de la lógica matemática moderna y realmente inspiran preguntas acerca de la naturaleza del pensamiento matemático, y, aunque yo estoy personalmente convencido de la extremada implausibilidad del modelo computacional de la mente, el argumento gödeliano de Penrose no hace nada para incrementar mi convicción, y sospecho que lo mismo será verdad en general para lectores con las mismas inclinaciones. Del otro modo, estoy seguro de que aquellos cuyas simpatías están en la dirección opuesta encontrarán razones para desechar el argumento gödeliano más rápidamente en este fundamento sin tener que pasar por el penoso proceso de elaborar las objeciones. Si estoy en lo correcto este es un esfuerzo destinado predominantemente al fracaso – aún tan laborioso como resulta. Sin embargo, está ahí, y me siento obligado a tratar al menos algunas partes del mismo, especialmente sus más técnicos aspectos.
>
> …
>
> ¿Podemos realmente esperar una teoría reductiva de cualquier tipo al respecto del fenómeno de la cognición humana? Seguro no hay una sola teoría que sirva para "explicar" los miles de aspectos que revisten a este fenómeno. Tal como con muchos otros estudios científicos al respecto de los seres humanos – dentro y fuera – tal empresa seguirá necesitando de la psicología, la psico-física, fisiología (neuro- y otros prefijos), bioquímic, biología molecular, física (macro y micro) y un montón más de asuntos en medio de éstos (incluyendo modelos computacionales de todos los tipos). En mi opinión, "la ausente ciencia de la consciencia" de Penrose es una ilusión.[1]

Cita que muestra un tono emocional que por lo demás, el usualmente brillante y técnico Feferman, podría simplemente dejar de lado y quizá, apuntar a clarificar por qué los teoremas de incompleción de Gödel no pueden demostrar la tesis –muy probablemente verdadera para él

[1] Mi traducción de:

> So now Penrose has gone to great lengths in SM to lay out his Gödelian argument and to try to defend it against all possible objections. I must say that even though I think Gödel's incompleteness theorems are among the most important of modern mathematical logic and raise fundamental questions about the nature of mathematical thought, and even though I am personally convinced of the extreme implausibility of a computational model of the mind, Penrose's Gödelian argument does nothing for me personally to bolster that point of view, and I suspect the same will be true in general of similarly inclined readers. On the other hand, I'm sure that those whose sympathies lie in the opposite direction will find reasons to dismiss the Gödelian argument quickly on one ground or another without wading through its painful elaboration. If I'm right, this is largely a wasted effort – diligent as it is. Nevertheless, it's there, and I feel obliged to address at least parts of it, especially its more technical aspects.
>
> …
>
> Can we really ever expect a completely reductive theory of one sort or another of human cognition? Surely, no one theory will serve to "explain" the myriad aspects of this phenomenon. As with any other scientific study of human beings – inside and out – such an enterprise will continue to need to bring to bear psychology, psycho-physics, physiology (neuro- and otherwise), biochemistry, molecular biology, physics (macro- and micro-) and lots of stuff in between (including computational models of all kinds). In my opinion Penrose's "missing science of consciousness" is a mirage.

por cierto –, de que la mente humana no es computable. En su lista de publicaciones, Feferman sólo tiene un artículo titulado "mind", y es justamente este.

La tesis se divide en cuatro capítulos. El objetivo principal requiere de un objetivo accesorio que consiste en proponer un criterio de clasificación de AGs con base en las razones que justifican o garantiszan la verdad o aceptabilidad de las premisas. El criterio de clasificación se explica y justifica en los primeros capítulos. Finalmente, en el capítulo final, ya habiendo establecido los criterios para la clasificación de AGs, defiendo que un tipo de AGs tiene mejores probabilidades de ser sólido—y por tanto es más plausiblemente sólido –. Este tipo de AG no admite como única evidencia real, la que es conseguible vía Tests de Turing (TT), estricto o ampliados (Total). Este tipo de AG es pues AG no-Turing Testeable. Es importante notar que rechazar el TT como única fuente de evidencia confiable no es lo mismo que rechazar el empirismo, el naturalismo, a las ciencias empíricas o nada semejante. Pero este rechazo puede ser defendido fácilmente como un intento no dogmático de analizar el asunto en cuestión. Quien esté comprometido con que la única evidencia admisible es, por ejemplo, la evidencia conseguible mediante algún TT, está siendo dogmático al respecto, y podría tener una postura racional y consistente, pero prejuiciosa al menos a la hora de juzgar las condiciones de solidez de los AGs.

En la primera parte, capítulo I, *Los Argumentos Gödelianos son*…, expongo la definición de AGs. Básicamente los AG son argumentos que buscan refutar el computabilismo usando como premisa la incompleción Gödeliana de la aritmética. También en ese capítulo expongo una versión resumida y semi-formal de los teoremas de incompleción. Defiendo en ella la razón por la que es importante entender la incompleción de la aritmética como parte definitoria de los AGs, en lugar de usar resultados análogos como el *halting problem*. La razón principal es que la incompleción de Gödel no sólo es la base conceptual del *halting problem*, sino que la incompleción de Gödel nos permite aludir a elementos que no son procesos computacionales de inicio sino específicamente a resultados científicos, i.e., a estados cognitivos un objetivo directo tanto de explicación como de reproduccón técnica de la Inteligencia Artificial y las ciencias cognitivas. En el capítulo I también se explica con detalle la forma de construir la oración G del teorema de incompleción. G informalmente dice que *G no es demostrable*, de modo que si lo fuera, sería demostrable una falsedad (nótese que no he aludido aquí a ningún *loop*, sino meramente a verdad y falsedad), y si no es demostrable, entonces es verdad pero indemostrable. Esta parece una consecuencia de sistemas poco comunes o que describen

situaciones muy abstractas y alejadas de la vida práctica, de modo que en esta sección se explican también las condiciones mínimas que debe tener un sistema cualquiera, para ser considerado como susceptible de un fenómeno de gödelización. Llamo a esta característica *Gödel-susceptibilidad*. Dentro de los sistemas que son *Gödel susceptibles*, hay algunos que no están Gödel limitados, lo que quiere decir que son capaces de enlistar como teorema la oración G, aún sin una demostración. Los sistemas Gödel susceptibles pueden ser humanos trabajando con sistemas formales, sin que ello implique que todos los seres humanos lo somos o lo somos todo el tiempo. También, los sistemas Gödel susceptibles podrían ser computadoras o robots. Explico también que los sistemas Gödel limitados son aquellos sistemas Gödel susceptibles que no son capaces de *reconocer* la verdad de la oración G relevante para ellos particularmente o la oración G en general.

En el capítulo II, *Diferente evidencia, diferentes Argumentos Gödelianos*, planteo que una forma de dividir los AGs con respecto al tipo de elementos de juicio que se consideran suficientes o definitivos para clasificar las condiciones de verdad de la tesis computabilista. Para usar el trabajo excepcional de Alan Turing, en este capítulo establezco la relación entre el Test de Turing (en adelante TT) y la evidencia que puede ser *Turing Testeada*. Hay una extensión natural del TT hacia lo que algunos autores llaman *Total*-TT en el que se admite como TT no sólo diálogo, sino propiamente evidencia como información genética, movimientos corporales, etc. Desde esta concepción, exploro los lazos entre el TT y el computabilismo. Encuentro que hay un lazo estrecho entre ambas nociones, de modo que parece injusto o tendiente a la falsedad, juzgar a los AG, cuya conclusión es que el computabilismo es falso, por medio solamente de un TT. Así, hay AGs evaluables bajo un TT y otros que no.

En el capítulo III, *La tesis computabilista es metafísicamente ambigua*, planteo que el computabilismo se puede interpretar bien de dos modos. El primero modo es epistémico (o metodológico), y el segundo es metafísico propiamente. Las *computadoras*, o mejor, las máquinas de Turing, nos sirven para dos cosas; para modelar (i.e., teorías) o para hacer (i.e., objetos). Así, cuando decimos que algo es computable, podemos estar afirmando que en sí mismo es una computadora, o bien que siendo o no una computadora, puede ser modelado óptimamente por una. En este capítulo abundo sobre estas nociones, y esbozo un par de nociones más: causalidad computacional y causalidad psicológica. Si ésta es igual que la primera, entonces la psicología de las personas es computable. Si las causalidad psicológica

permite más accones o pensamientos de los que podría en principio la causalidad computacional, entonces la psicología de las personas no es computable.

Finalmente, el capítulo IV, *Explicación y justificación de las condiciones de solidez de los Argumentos Gödelianos*, habiendo ya mostrado las condiciones de aplicabilidad sin presuponer en ningún momento la tesis a discusión, *grosso modo*, el computabilismo, en este capítulo muestro que a los AGs que admiten como tribunal de la evidencia sólo evidencia TT, se les refuta fácilmente con críticas semejantes a la de LaForte, Hayes y Ford (1998). Los tipos de AGs que no presuponen un tribunal de la evidencia TT son mejores al menos pragmáticamente : toman en serio una implicación directa de la conclusión de los AGs, a saber, que no todos los fenómenos son computacionales (metafísicamente) o al menos computables (epistémicamente). Es así que dentro de los AGs no evaluables bajo un TT todavía tenemos uno que es de mejor calidad que el otro: el AG metafísico es superior, ya que no implica que algún AG epistémico sea correcto. Es decir, podríamos tener un AG metafísico cuyas premisas no son evaluables por algún TT, sólido que no podamos detectar como metafísicamente sólido o bien a la vez que el AG epistémico resultante sea refutable (es decir, que nos parezcan injustificadas epistémicamente alguna de sus premisas). Al respecto de lo anterior debo decir que esta tesis no explorará las razones por las cuales el realismo (de la mente) tiene más fortalezas racionales que el anti-realismo (de la mente), pero creo que para investigaciones posteriores, la relación entre un AG tipo IV (que sea metafísico pero no admite evidencia de TT) y uno tipo III (que sea epistémico pero no se comprometa con evaluación con TT) puede darse por descartada. Podría parecer justo lo contrario: un argumento cuyos términos no tienen aspiraciones semánticas realistas, es más débil y resiste más objeciones que los argumentos con pretenciones semánticas realistas. Pero no creo que sea así y en el capítulo IV hay algunas razones por las cuales al menos para el caso de evaluar las condiciones de solidez de los AGs tengo razón: podría haber una teoría óptima *dada* que contuviera en sí misma la demostración de su incompleción, a la vez que se demostrara la verdad de su oración *G* relevante y todas las oraciones *G* resultantes como axiomas. Pensando en *agentes racionales* en lugar de *teorías*, un agente racional que sepa que es incompleto y a la vez *se conduzca* como si su oración G fuera verdad, podría no saber que su oración G relevante sea verdad, sino sólo *hacer como si lo supiera*, por ejemplo, ante la pregunta "¿G es verdad?", él contestar "Sí, pero no lo puedo demostrar." O más conservadoramente, "Si lo supiera, sería incoherente", o la secuencia de signos y conducta TTeable que mejor satisficiera nuestra idea de qué es ser un agente racional Gödel-ilimitado. En cambio, *saber G* parece requerir algo superior a meramente repetir una

fórmula o tomarla como una verdad indemostrable dada: existe algo que es comprender cómo y por qué G es verdad, aunque indemostrable.

Casi en cada tema que trato en la tesis hay afirmaciones polémicas y ampliamente criticadas en filosofía. Los autores que han publicado sobre esta clase de argumentos, suelen ser muy tajantes en sus apreciaciones tanto a favor, como en contra del anticomputabilismo. No tengo espacio para profundizar a mi gusto en todo pero estas mismas características hacen de la presente tesis un trabajo del cuál salen naturalmente más. Por ejemplo, la implausibilidad psicológica del axioma modal que expresa *introspección positiva* (S4, es decir, Kp→KKp), la noción de causalidad psicológica, las implicaciones metafísicas de las verdades lógicas, la atribución de Gödel-susceptibilidad a agentes racionales humanos, y así más. Espero que esta tesis sirva, al menos, para dos cosas: la primera es ordenar el debate al ofrecer un criterio de clasificación de AGs que ubique correctamente las críticas a ellos, y permita conocer los alcances de las mismas, la segunda es defender la plausibilidad de los AGs una vez que uno puede pensar sobre el mundo real (natural o como sea que resulte ser) sin tomar como dogma el finitismo o finitismo computabilista. Una vez logrados estos dos objetivos, un objetivo secundario sería que esta tesis consiga luz al respecto de cómo puede construirse a detalle, para trabajo posterio, un AGs que no sólo sea válido, sino que tenga premisas suficientemente verdaderas como para dar un impulso hacia adelante a las ciencias naturales en su objetivo de conocer a la mente humana.

En esta tesis tampoco propongo a detalle un AG~~s~~ definitivo que yo vaya a defender de todo a todo. Ese es el objetivo de trabajo posterior. Pero sí hay lineamientos generales tanto para proponer uno, como para evaluar los AGs en general. Algunas de las críticas son suficientemente fuertes como para considerar terminados todos los AGs que son correctamente atacados, pero hay otros tipos de AGs que no están en estos supuestos y esta idea es una de las más valiosas que emergen de la clasificación que propongo.

Aunque he puesto especial atención en tratar de buscar adeudos históricos de los AGs, no es mi intención principal hacer una historiografía completa en modo alguno. Tampoco es esta una tesis de historia de la filosofía o filosofía de la historia. Un error de naturaleza historiográfica sería de importancia inferior al interés principal de la tesis, que es la clasificación y las condiciones de evaluación de los AGs.

Aunque lo he intentado, no aspiro que las condiciones de evaluación de los AGs de esta tesis sean completas. No veo elementos distintos a los que he manejado aquí, que puedan ser relevantes, aunque esto bien puede ser a causa de la *ceguera de taller*. De modo que aunque aspiro a haber evaluado todos los elementos relevantes para evaluar correctamente los AGs no osaría poner demasiada fuerza en esta situación.

En las conclusiones recojo el fruto que he ido cultivando durante toda la tesis: existen al menos un tipo de AGs que son resistentes a la mayoría de las excelentes críticas que se han hecho a los AGs en general. Ofrezco también en este capítulo algunas apelaciones a la autoridad a favor de la plausibilidad o al menos de la razonabilidad de la solidez de al menos algún AG.

"O bien… la mente humana… sobrepasa infinitamente los poderes de cualquier máquina finita, o, de otra forma, existen problemas diofánticos absolutamente irresolubles." XXV Conferencias Gibbs, Kurt Gödel, 1951. (pp. 310)[2]

1. Los Argumentos Gödelianos son…

1.0 Introducción

El lector encontrará aquí una definición más detallada de lo que consideraré Argumento Gödeliano. Para entenderla, profundizo en una de las premisas de los mismos, es decir, la verdad necesaria de la incompleción de Gödel, ya que con estas nociones podemos ahora entender detalles finos del funcionamiento del los Ags. Por ejemplo, algunas de las críticas más fuertes que sehan lanzado contra la solidez de los AGs tiene indican que para que el teorema de incompleción o cierto teorema de incompleción relativo a un sistema formal específico sea tal necesitamos poder probarlo, para lo que necesitaríamos recursos que no tenemos o que tales o cuales sistemas no poseen. Así, conocer los teoremas de incompleción de Gödel, en especial el primero, es importante para entender qué recursos, incluyendo recursos conceptuales, son necesarios para que cierto sistema sea siquiera susceptible de incompleción gödeliana. Este es un ejemplo. Así, establecemos dos nociones más: que un sistema pueda ser incompleto gödelianamente independientemente de si lo es o no, Gödel-susceptibilidad. Y también que un sistema siendo Gödel-susceptible, no es limitado por la incompleción gödeliana, para los proponentes de AGs, al menos algunas mentes humanas, es Gödel-ilimitación.

En este capítulo expongo la definición de *argumento gödeliano*, algo de historia de los mismos, una propuesta de formalización parcial y resumen de los Argumentos Gödelianos

[2] Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives).

(AGs en adelante). También hago una exposición somera aunque rigurosa de los teoremas de incompleción de Gödel y algunas de las implicaciones que tienen. Con esta información elaboro dos nociones fundamentales: Gödel-susceptibilidad y Gödel-limitación. La primera noción sirve para establecer condiciones de posibilidad de solidez de los AGs, misma que ha sido cuestionada por algunos especialistas. Un agente capaz de aritmética que no es Gödel-susceptible, no puede ser *o bien incompleto o bien incoherente*. Finalmente, una vez que consideramos a los agentes racionales Gödel-susceptibles, aquellos que son capaces de ser coherentes y completos son agentes Gödel-ilimitados. La segunda noción que construí en este capítulo es justamente la noción que es posiblemente vacía, de agentes Gödel-ilimitados. Sostener que un AG es sólido compromete a cualquier persona que lo sostenga con la tesis de que existe al menos un agente Gödel-ilimitado, no potencial o hipotéticamente, sino actualmente.

1.1.1 La objeción matemática de Turing, *Argumentos Gödelianos*

El primer *argumento gödeliano* (en adelante, AG) aparece en el catálogo de objeciones al proyecto de inteligencia artificial que aborda Turing en un artículo celebérrimo, *'Computing machinery and intelligence'* (Turing, 1950) que incluye una lista de posibles objeciones al modelo de inteligencia mecánica artificial de las computadoras digitales. Una de las objeciones es la que Turing llamó *objeción matemática*. Su formulación es ya suficientemente transparente como para considerarla un AG típico:

> El resultado en cuestión [el teorema de Incompleción de Gödel] refiere a un tipo de máquina que es esencialmente una computadora digital con capacidad infinita. [El resultado] establece que hay ciertas cosas que una máquina de este tipo no puede hacer. Si está arreglada para dar respuestas a preguntas como en el juego de imitación [mejor conocido como *Test de Turing*], habrá algunas preguntas a las cuales o bien dará una respuesta incorrecta o bien fallará completamente en dar una respuesta sin importar cuánto tiempo se le dé para que la formule. Puede haber, por supuesto, muchas preguntas de este tipo, y preguntas que no pueden ser contestadas satisfactoriamente por una máquina pero sí por otra. Estamos suponiendo, es claro, en este texto, que las preguntas son del tipo para el que una respuesta 'Sí' o 'No', es apropiada, en lugar de preguntas tales como '¿Qué piensas de Picasso?' Las preguntas en las que sabemos que las máquinas deben fallar son de este tipo, "Considere la máquina especificada de la siguiente forma… ¿Contestará esta máquina 'Sí' a cualquier pregunta?" Los tres puntos están

> para ser reemplazados por una descripción de alguna máquina de manera estándar**…** Cuando la máquina descrita utiliza cierta relación comparativamente simple con la máquina que está siendo interrogada, se puede mostrar que la respuesta es o bien incorrecta o no llega. Este es el resultado matemático: se argumenta que él prueba que las máquinas están sujetas a una cierta discapacidad a la que el intelecto humano no está sujeto.[3] (Turing, 1950, pp. 444-445)

¿Por qué este es un ejemplo de AG? Porque del resultado meta-lógico de incompleción gödeliana se infiere la imposibilidad de que exista alguna *máquina de Turing* (MT) que sea un modelo *apropiado* de la mente humana capaz de lógica y aritmética. Turing responde de manera simple e iluminadora aunque desafortunadamente también, y en mucho debido a su brevedad, con insuficiencia. Ya revisaré con detalle en el capítulo tercero lo que creo son tipos distintos de AGs y que, como se verá aquí, no fueron abordados en el trabajo de Turing (1950).

Una definición inicial de AG es que son razonamientos que sostienen a partir de la incompleción de Gödel o de resultados semejantes que toda *máquina de Turing* (MT en adelante) abstracta o instanciada físicamente habrá algún teorema que no pueda decidir. En caso de que algún agente pudiera decidir si el teorema en cuestión es derivable, ello evidenciaría el uso de un procedimiento que queda fuera de las reglas de programación del agente ~~y~~ coherentes con la lógica de predicados de primer orden (LPO en adelante).

Hay una bifuración importante de los AGs que depende de la evidencia que sea suficiente para atribuir conocimiento matemático. Esta bifurcación de los AGs responde a su vez, elaboraré, a ~~de~~ qué baste para atribuir que alguien o algo *sabe* que cierta fórmula es teorema, lo cual depende en parte de qué evidencia haya a favor de la consistencia de la *aritmética de Peano* (AP en adelante). El término *saber que* ha sido escogido intencionalmente. La idea es mantener una ambigüedad que éste término porta suficientemente bien y consiste en que *saber que* puede

[3] Mi traducción de: “The result in question refers to a type of machine which is essentially a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot do. If it is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course supposing for the present that the questions are of the kind to which an answer 'Yes' or 'No' is appropriate, rather than questions such as 'What do you think of Picasso?' The questions that we know the machines must fail on are of this type, "Consider the machine specified as follows. . . . Will this machine ever answer 'Yes' to any question?" The dots are to be replaced by a description of some machine in a standard form, ... When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject.” (Turing, 1950, pp. 444-445)

ser sustituido por *se ha calculado correctamente que p*, *se ha inferido correctamente que p, tenemos la introspección de que p*, y más.

1.1.2 Forma general de los AGs

La forma general de los AGs es como sigue:

> 1. Si algo es una MT consistente capaz de Aritmética de Peano y Lógica de Predicados de Primer Orden o bien hay al menos una oración Gödel cuya *verdad* no puede *saber*, o bien es inconsistente. (Teoremas de incompleción de Gödel y el hecho de que se puede formular en términos de *máquinas de Turing* también.)
>
> 2. Las mentes humanas (al menos sus capacidades cognitivas) no son más que MTs, **parcial(es) o total(es)**.
>
> 2.1 Las mentes humanas son el mismo tipo de cosa que una MT.
>
> > 2.2 Las mentes humanas son *óptimamente representadas* por alguna teoría dentro de los límites de al menos alguna MT. (En algún nivel o algunos niveles de descripción sea fisicalista o no)
>
> 3. **Algunas** mentes humanas *pueden saber que* las oraciones Gödel son verdad.
>
> **4. Por lo tanto, o bien, ninguna mente humana *puede saber que* las oraciones Gödel son verdad, o bien algunas mentes humanas no son MT, total(es) o parcial(es) y no son representables óptimamente por alguna teoría que sea computable.**

Si alguien sostiene de 1-3, cae en contradicción. La premisa 3 es necesaria para que cualquier AG funcione y buena parte de la literatura sobre AGs se centra en debatir este punto. Hay términos en esta formulación del argumento, que introducen ciertas imprecisiones. Por ejemplo, en la premisa 1, usamos el término '*saber*' en lugar de '*derivar*', porque las derivaciones son una forma de conocimiento, aunque no es universalmente aceptado entre los filósofos que todas formas de sapiencia son derivaciones, y menos aún, derivaciones formales

usando lógica de predicados de primer orden. Estas imprecisiones permiten capturar bien la forma en la que suelen presentarse muchos AGs en las discusiones académicas, y, con la advertencia, no tienen por qué imposibilitar o dificultar innecesariamente la comprensión, la caridad, la validez o incluso la solidez de algunos de ellos. La premisa 3, merece también una advertencia. *Que las mentes humanas puedan saber*, es compatible con el hecho de que haya solo una mente que siendo humana sea capaz de ese conocimiento, aunque también lo es con el hecho de que aunque actualmente no haya alguien que lo tenga, sea asequible, es decir, que las mentes humanas tengan lo mínimo para saberlo, aunque nadie lo haya investigado, comprendido y justificado correctamente, y creído. Quizá no sobra decir que "algunos" no es inconsistente con "todos" o con "todos hasta ahorita." Esta clarificación hasta donde veo, es accesoria al interés principal de los AGs.

Dado que los teoremas de incompleción de la aritmética y la lógica de primer orden (en adelante por simplicidad, teoremas de Gödel) se construyeron en sistemas formales, discutir acerca de las consecuencias no sólo metodológicas (p. ej., si es una derivación correcta, qué implicaciones tiene al respecto de la noción de validez lógica, etc.) sino también metafísicas de los mismos, genera las más contrapuestas intuiciones entre quienes sostienen y critican AGs. Computólogos y filósofos como Selmer Bringsjord para dar un ejemplo, argumentan en contra de la validez de algún AG mientras que en otro lado, proponen la solidez de uno (*cfr*. con Xiao, 2000, y solo 2006, 1992).

No obstante las dificultades, ellas no tienen por qué disuadir, en especial, a los filósofos para tratar los AGs. Alguna defensa férrea de la neutralidad tópica de la lógica formal y sus resultados meta-lógicos podría afectar de entrada los presupuestos generales de la discusión acerca de los AGs. Pero por razones de espacio no defenderé de este tipo de ataques a la importancia filosófica de los AGs. Sólo digo que el que los sistemas formales de lógica o de matemáticas pretendan estar desprovistos de contenido específico no significa que estén completamente desprovistos de alguna implicación fáctica verificable o falseable. "2+2=4" no es lo mismo que "dos manzanas más dos manzanas~~,~~ son cuatro manzanas" pero éstas no son aseveraciones desconectadas del todo.

1.1.3 El objetivo de los *Argumentos Gödelianos*

La caracterización más general de AG es mediante su objetivo. Los AGs son razonamientos que pretenden usando al menos el primer teorema de incompleción de Gödel y una cierta interpretación del mismo además de alguna otra premisa, bastar para probar que la mente humana no es un fenómeno (por completo) modelable computacionalmente[4]. En otras palabras, un AG busca probar que la mente humana no es [*Turing*] *computable*.

1.1.4 Tipos de Argumento Gödeliano

Los AGs cumplen con el requisito de defender que las mentes humanas no son computables, por razones de su composición metafísica o por límites a cómo las conocemos. De esta manera cuando en una discusión alguien menciona que los teoremas de incompleción de Gödel *prueban que* la mente humana no es mecánica o que no es una computadora, está muy probablemente pensando en algún AG. Hasta este momento es compatible con la caracterización de la sección 1.1.2 que podemos tener muchos tipos de AGs, y es propósito de esta tesis primero, ofrecer una manera general de clasificarlos, analizarlos a ellos y a sus consecuencias teóricas, y, para luego poder evaluarlos.

Creo que hay distintos tipos de AGs de acuerdo a distintos criterios que se explicitarán en el transcurso de este trabajo. Trataré de clasificar a los AGs conforme a dos criterios: i) si la verdad de la oración Gödel se trata de establecer empíricamente o, ii) si la verdad de la oración Gödel no se puede establecer empíricamente.

Hay dos grandes *tipos de autoridades* que se han propuesto para decidir qué tanto puede un AG mostrar si la mente humana está fuera de los límites de la Turing computabildiad (*computabilidad*, en adelante). La primera autoridad es la de la observación externa de otras mentes. Si presuponemos que la mente es un fenómeno natural, entonces aceptamos que el conocimiento de los estados mentales es conocimiento de estados cerebrales además de ciertas situaciones externas. La segunda autoridad que admiten los AGs no se compromete con la posición de que el escrutinio desde la tercera persona está dotado de un peso justificador óptimo y defiende la irreducibilidad de, por ejemplo, estados mentales subjetivos a descripciones desde la perspectiva de la tercera persona.

[4] Me doy cuenta de que la expresión es vaga, pero esto es a propósito. Las especificaciones pertinentes serán hechas más adelante en esta tesis.

Así, algunos AGs presuponen claramente que hay al menos en principio un Test de Turing (en adelante TT aunque sea una variación conservadora del mismo) que basta para decidir si hay o no primero, capacidad para la aritmética y la lógica, y si hay, segundo *Gödel susceptibilidad* y en consecuencia, si un agente tiene poderes cognitivos comparables con el modelo capaz de construir una oración Gödel de sí mismos, de su propio modelo, de su tipo de modelo, de un modelo mucho más general que el suyo propio o de otros. "Gödel susceptibilidad" es la característica que tiene un sistema formal abstracto o que describe a sistemas concretos, de que la demostración de su coherencia implica su incompleción por demostraciones como la incompleción de Gödel: posibilidad de aritmética, de representación y auto-representación, demostración, etc. Otros AGs, no presuponen que un TT sea idóneo para determinar si un agente racional supera o no los límites de que impone la incompleción de Gödel.

Bringsjord (1992), por ejemplo, propone un AGs del primer tipo. Este tipo de AGs asegura que aunque hay ciertos objetos particulares como robots o computadoras que *podrían hacer verdaderas* (satisfacer) teorías Turing computables estrictamente sobre la mente humana, muchos de ellos aún podrían fallar en pasar todos los TTs que sí pasarían por lo menos, ciertos humanos. Los TTs que separarían confiablemente a humanos de computadoras. Quizá, por ejemplo, la capacidad mínima de entender relaciones numéricas entre oraciones, sea una de ellas. Estas relaciones no tienen por qué ser formales en el sentido tradicional del término, ya que todo lenguaje natural puede expresar relaciones entre *individuos únicos* (nombres propios, descripciones definidas) u otros que refieren a grupos cuantificables de individuos (nombres comunes, sustantivos de masa, etc.).

Un grupo de AGs representado por el que Gödel propone en *Las Conferencias Gibbs* (1951) no se comprometen con algún criterio siquiera cercano al de los TTs para dirimir la cuestión. La solidez de este tipo de AGs dependen de que *sepamos* que la oración Gödel relativa a los sistemas Gödel susceptibles—incluyendo al *sujeto cognoscente* en cuestión, es verdad sin comprometerse con la tesis de que hay por lo menos una forma conductual TTeable para evaluar si un agente en efecto *sabe que p* o *no sabe que p*. Éste es el espíritu de quien confía más en el conocimiento matemático que en el alcance del conocimiento empírico para determinar las potencialidades de nuestras capacidades cognitivas. Si empíricamente no encontramos alguna conducta (verbal, cerebral, etc., verbal y cerebral, etc.) que nos permita atribuir conocimiento *del teorema de incompleción* a algún agente, existe la posibilidad de que

sepamos que la oración G es verdad, pero *saber que el agente A sabe que la oración G es verdad*, en cualquier combinación, cuando el agente que sabe es él mismo, u otro agente, no sea siquiera posible sin admitir el elemento de duda que arroja la posibilidad de que podríamos ser inconsistentes. Formalizada de manera estándar esta afirmación *parece* una excepción a un axioma epistémico modal, que haría que el operador modal *saber* (L) no sea reflexivo o no sea transitivo. El axioma en cuestión es, donde *x* y *y* son variables de agentes:

$$\text{Sabe}^x(G) \supset \text{Sabe}^y(\text{Sabe}^x(G))$$

Incluyendo en especial el caso en el que

$$x=y$$

tendríamos un contraejemplo intuitivo interesante para la posibilidad del auto-conocimiento total y omnisciencia lógica. Ahora, queda por saber cómo es que esta limitante digamos autoepistémica, afectaría la fortaleza con la que percibimos que son verdaderas las verdades lógicas y aritméticas. Dicho de otra forma, si una de las consecuencias de las limitaciones autoepistémicas implica que no podemos estar seguros de la fortaleza epistémica de *G*, entonces, podría ser que de hecho

$$\text{Sabe}^x(G)$$

Sea incompatible con

G es una verdad necesaria.

O con que

G es una verdad lógica

Por supuesto, estos detalles se discuten a lo largo de esta tesis.

1.2 Requisitos formales del Teorema de Incompleción de Gödel I.

No voy a reproducir la derivación formal de los teoremas de incompleción de Gödel. No es necesario para la comprensión de los AGs. No obstante, a continuación ofrezco un esbozo de la estrategia argumentativa general del teorema con cierto grado de profundidad. Esta es una prueba a lo mucho semi-formal de los teoremas de incompleción de Gödel. (Véase Smullyan, 2001 y Enderton, 2001)

El primer teorema de incompleción de Gödel presenta una paradoja análoga a la del mentiroso. Este tipo de problemas son recurrentes en sistemas dentro de los cuales podemos expresar el propio predicado de verdad.[5]

Dado que la derivación de los teoremas de Gödel es formalmente barroca, fácilmente puede dar la impresión de ser truculenta[6]. Aunque pueden ofrecerse algunas razones de peso para cuestionar la validez del resultado de incompleción gödeliana, en este trabajo presupondremos los teoremas de incompleción de Gödel, además del resultado isomórfico conocido como el *problema de la detención* (o *halting problem*). No voy a tomar el pasado, por razones que espero sean obvias: nos interesa no sólo el aspecto mecánico de los teoremas de incompleción, sino también el epistémico y cognitivo. Un teorema es tanto un producto de un procedimiento mecánico, como un objeto de actitud epistémica, mientras que *el halting problem* se reduce a procedimientos mecánicos, sean estos epistémicos o no.

Una teoría cualquiera es usualmente modelada como un conjunto de oraciones *acerca* del mundo o de cierta parcela de este. Veamos a grandes rasgos en qué consisten los teoremas de incompleción de Gödel. Pensemos en el caso de una teoría científica como lo es la aritmética de Peano. En ella podemos simbolizar a los diferentes números naturales, a las funciones de suma, resta, multiplicación, división y exponenciación, además de dos relaciones, la de igualdad y la de orden (‘menor que’ o ‘<’), algunas conectivas lógicas además de axiomas, y la provisión de reglas de inferencia. Llamemos a esta teoría PA (por el término en inglés, *Peano Arithmetic*). En lo que toca a la exposición de los teoremas incompleción de Gödel, usaré el texto de Enderton (2001) y la traducción de José Alfredo Amor como Enderton (2004, UNAM).

Considérese el lenguaje típico en el que hacemos expresiones de aritmética. Tenemos una lista dentro de la cual están:

1. $2+2 = 4$
2. $4 \times 6 = 24$
3. $0^1 = 0$

[5] Es la tesis que presenta el trabajo de 1951 de Alfred Tarski. Algunas otras contribuciones han aparecido desde entonces.
[6] Como parece que lo interpretó Zermelo en su primera aproximación a él. Para algunos detalles más, véase Dawson (1984).

4. $1 < 182$

5. $1 = 1$

6. $2 = 2$

Y otras fórmulas mucho más largas como:

7. $10{,}000{,}000 \times 902{,}462^{29}$ = [Aquí el resultado correcto]

Considerando las cosas con todo rigor en el tratamiento formal de diversos fenómenos debemos distinguir para empezar entre signo y referente. Esta distinción es además crucial para la derivación de los teoremas de incompleción de Gödel. En las derivaciones más usuales del teorema de Gödel, recurrimos a distinguir entre números, funciones, relaciones monádicas y poliádicas, y sus respectivos signos (numerales, signos de función y predicados de uno o más lugares), de tal forma que, por ejemplo, la igualdad del ejemplo 1, 2+2=4, quedaría:

1'. $P_{=}^{2}(f_{+}^{2}(\mathrm{SS}\mathbf{0}, \mathrm{SS}\mathbf{0}), \mathrm{SSSS}\mathbf{0})$

'$P_{=}^{2}$' está por 'predicado diádico = (es igual a)', 'f_{+}^{2}' está por 'la función diádica + (suma)', 'S' por 'función monádica *sucesor de*' o en otros términos está por '+1', y finalmente '**0**' está por el 0 o *cero*. Tener en mente la distinción anterior es importante a la hora de considerar el concepto de *representabilidad*, es decir, la noción que dice que *cierto predicado P sea representable al interior de alguna teoría cualquiera*.

La noción de *representabilidad* implica entre otras cosas que cada una de las fórmulas de un sistema formal para PA, pueden expresarse dentro de un lenguaje formal. La expresión puede o no preservar características semánticas o sintácticas de las fórmulas representadas, aunque en este caso una representación dentro del sistema (enumeración Gödel) preserva una característica sintáctica fundamental de las fórmulas: su finitud. Ahora**,** hay maneras de conservar la característica sintáctica de la finitud de las fórmulas de PA y ***Lógica de Predicados de Primer Orden*** **estándar** (LPO en adelante). Por ejemplo, "2+2=4", 1', bien puede ser representada como

1''. oo M oo ^ oooo

de nuestra lista anterior, preservando un isomorfismo (que espero sea obvio) o, quizás sólo mediante un número en un listado, digamos, *la fórmula 345 del listado tal*. Más detalles se verán cuando sean pertinentes entre las dos nociones de representación.

Además de poder expresar afirmaciones y términos de PA, agreguemos la posibilidad de expresar afirmaciones de LPO. En ella podremos expresar numerales, variables de objeto, constantes de objeto, y funciones de *n* lugares dentro de un mismo grupo sintáctico, funciones de verdad monádicas y diádicas conocidas, en otro grupo sintáctico, y predicados y relaciones de cualquier aridad, en otro. Además tenemos símbolos auxiliares (paréntesis, por ejemplo). En el lenguaje de LPO podemos expresar fórmulas como la siguiente:

8. $\forall x\,(\mathrm{F}x \rightarrow \mathrm{A}x) \equiv \forall x\,(\mathrm{F}x\ \&\ \mathrm{A}x)$
9. $\forall x\,((\mathrm{H}x\ \&\ \mathrm{A}x)\ \&\ \forall y\,((y = x) \rightarrow (\mathrm{H}y\ \&\ \mathrm{A}y)))$

En el lenguaje de PA+LPO, podemos hacer expresiones que combinan símbolos de los dos sistemas. Por ejemplo:

10. $\forall x((\forall y(x + y = y)) \rightarrow (x = 0))$

Intuitivamente 10 dice 'cero es el *neutro aditivo* y nada más lo es.'

Con estos recursos mínimos y suponiendo alguna versión del axioma de elección, podemos encontrar una función biyectiva PA $\rightarrow \mathbb{N}$[7]. A esta función la llamaremos en general, *aritmetización de la metamatemática* o *enumeración gödel*. La idoneidad de este paso para preservar, por ejemplo, sentidos interesantes de la noción de *prueba*, *verdad matemática* y *conocimiento*, se discutirá en los capítulos subsecuentes. "Cualquier forma suficientemente clara de asignar enteros distintos a fórmulas distintas, **bastará** para nuestros propósitos… [l]o que *es* importante es que de α nosotros podamos encontrar efectivamente #α, y de forma conversa"[8] (Enderton, 2001, p. 184[9]). "α" es para Enderton (2001) una meta-variable de cualquier expresión del lenguaje en el que se expresa PA+LPO.

[7] Las fórmulas de PA incluyen las de LPO, ya que suelen incluir todos los símbolos y funciones lógicas, más reglas de gramaticalidad, y variables.
[8] Mi traducción de: "Any sufficiently straightforward way of assigning distinct integers to formulas would suffice for our purposes... [w]hat *is* important is that from α we can effectively find the number #α, and conversely."

Dado que las derivaciones de todo sistema formal son secuencias de fórmulas, entonces a cada derivación de PA+LPO le corresponde un número Gödel. Una secuencia es un conjunto **que preserva** cierto orden **de sus elementos y cuando sus elementos son enunciados, sus valores de verdad pueden representarse asimismo como objetos de la secuencia del modo siguiente:** *de* **las fórmulas α y β se sigue la verdad de la fórmula γ, es decir (α & β) → γ**. Este orden bien puede excluir el compromiso inverso a saber el de que si uno sustituye α y β por γ y a γ por α y β se sostenga también la misma relación de implicación. La numeración Gödel que veremos preserva el orden mediante la utilización de un sistema ingenioso usando números primos y factorización. Los números primos son fácilmente individualizables en términos de cómo se relaciona aritméticamente con otros números[10]. Para las fórmulas de PA+LPO es crucial el orden. No es lo mismo

11. $(p \rightarrow p) \rightarrow p$

que

12. $p \rightarrow (p \rightarrow p)$

en donde "p" es una constante para enunciado, por ejemplo, "2+2 = 4" o uno cualquiera, otro. El condicional material (→) es sensible al orden tal como exponenciación. Usando números primos y exponenciación podemos pues conseguir la función a la que responde a los objetivos y la llamamos *enumeración Gödel*.

Llamemos *h* a la función que **para** cada expresión α del lenguaje de PA+LPO, de tal forma que **a cada expresión la convierte en un número siguiendo ésta table de tal forma que** $h(V) = 0$ y $h(0) = 2$ **como se ve en la tabla**. **El número Gödel se calcula paso a paso, de la siguiente forma.** En el caso de la expresión ε del lenguaje en el que pretendemos expresar todas las verdades de PA y de PA+LPO, tal que $\varepsilon = (s^0 \ldots s^n)$ su número Gödel #(ε) se obtiene así:

i. $\#(s^0 \ldots s^n) = \langle h(s^0), \ldots, h(s^n) \rangle$

PA+LPO	Número Entero	PA+LPO	Número Entero

[9] En la traducción de José Alfredo Amor publicada por la editorial IIF-UNAM, se lee lo siguiente: "Cualquier forma lo suficientemente explícita de asignar enteros distintos a fórmulas distintas nos servirá para nuestros propósitos… [l]o importante *es* que para cada α podamos decir qué número #α le corresponde, y viceversa."
[10] Por definición, un número primo es aquel número entero positivo que no puede ser dividido (la función inversa a la multiplicación) **sin** remanente entero positivo (de cero en adelante, cualquiera), salvo por dos números: él mismo y la unidad.

0	2	(	3
S	4	)	5
<	6	¬	7
+	8	→	9
x	10	=	11
∀	12		
		x^1	13
		y	15
		x^2	17
		Etc.[11]	

La tupla ordenada resultante de números y habiéndoles sumado un tanto más, constituyen el exponente del listado de números primos empezando con el 2 para s^0 y el primo *n-avo* para la expresión s^n. Veamos un ejemplo.

13. $\forall x^3 \neg(x^3=\mathbf{0})$

14. $h(\forall x^3 \neg(x^3=\mathbf{0}))$ *es* $<h(\forall), h(x^3), h(\neg), h(\ (\ ), h(x^3), h(=), h(0), h(\ )\ )>$

14'. $h(\forall x^3 \neg(x^3=\mathbf{0}))$ *es* <12, 19, 7, 3, 19, 11, 2, 5>

Ahora, en la expresión 13 del lenguaje de PA+LPO que analizamos ocurren ocho signos, incluyendo los que se repiten como la variable x^3, lo que nos indica que serán los ocho primeros números primos los que usaremos como base de los exponentes mismos que son los términos $h(s^0 \ldots s^n)$.

15. $\mathbf{2^{12} \times 3^{19} \times 5^7 \times 7^3 \times 11^{19} \times 13^{11} \times 17^2 \times 19^5}$

Para una cadena ordenada de fórmulas que componen o no una deducción (inferencia válida) en PA+LPO, simplemente encadenamos con multiplicaciones del número Gödel # de la secuencia de fórmulas. A la asignación de número Gödel de una derivación la simbolizamos, en lugar de con #, con *G*, de tal forma que

ii. $G(<\alpha^0, \ldots, \alpha^n>) = <\#\alpha^0 \ldots \#\alpha^n>$

[11] Es importante notar que el condicional material (→) y la negación (¬) son un conjunto de conectivas lógicas suficiente para expresar el resto de ellas. Por simplificación veremos una enumeración Gödel restringiendo la cantidad de funciones de verdad y predicados, los del cuadro.

cuando a un conjunto no ordenado de fórmulas o expresiones Φ cualesquiera, como es también una derivación de LPO o PA+LPO, quedaría como sigue:

iii. #Φ = {#(ε) *tal que* ε *es un elemento de* Φ}

A cada secuencia de fórmulas, que son conjuntos ordenados de expresiones le es asignado un número de Gödel, básicamente con el mismo mecanismo. La forma de enumeración Gödel es recursivamente definible. Lo que implica que existe una forma mediante la cual dotar a toda fórmula de PA+LPO con un número entero positivo a la manera de arriba. Las relaciones ordenadas como la de derivabilidad son representables extensionalmente como relaciones entre números de Gödel y como relaciones aritméticas son recursivamente representables.

Podemos tomar PA+LPO como una estructura u objeto de una interpretación. Una estructura es la función que refiere a un modelo para un sistema formal, es decir, es una interpretación que hace posible asignar valores de verdad a las fórmulas de un sistema formal, en este caso un sistema formal que a su vez se refiere a las inferencias válidas. Tenemos ahora un serie de preguntas interesantes al respecto de PA+LPO. Una por ejemplo, es la de si todas las verdades de PA+LPO son derivables de un conjunto finito de axiomas, o de si algún conjunto de reglas de inferencia bastan para descartar cada afirmación falsa gramaticalmente admisible del lenguaje de PA+LPO sobre tal estructura. Todas estas son preguntas distintas. Los teoremas de Gödel suponen que las reglas de validez se refieren a lenguajes. Una regla de inferencia válida, hace exactamente lo mismo excepto que, si del conjunto de afirmaciones del que partimos es verdadero, entonces las conclusiones no pueden ser falsas.

Así, una derivación válida que va del conjunto de afirmaciones en el lenguaje de PA+LPO, llamémoslo *A*, que nos lleva a una o más fórmulas (conclusiones), llamemos a tal conjunto *B*, lo hemos de simbolizar así**:** $A \vdash B$. Mientras que si *A* hace verdadero a *B*, independientemente de si hay una regla para llegar de *A* a *B*, lo simbolizaremos de la siguiente forma $A \models B$. De "*A*" se infiere "*B*", leemos la primera relación, y "*A*" hace verdad a "*B*", a la segunda relación.

A todas las afirmaciones obtenidas mediante reglas de inferencia válidas a partir de algún conjunto finito *A* de afirmaciones en un lenguaje que expresa PA+LPO la simbolizaremos así Cn(*A*). En el caso de que las consecuencias de PA+LPO sean a partir de cualquier conjunto de afirmaciones, entonces podremos no circunscribirlo: Cn(PA+LPO). En cambio a todas las oraciones hechas verdaderas al asumir PA+LPO, independientemente de si usamos reglas de

inferencia o no, la simbolizamos así Th(PA+LPO). “Cn” está por consecuencias (el último renglón en una derivación), y “Th” está por el término en Inglés “*Theory*” (esto es, *teoría* las verdades deductivamente obtenidas, no necesariamente usando reglas de inferencia).

Una característica que tiene el lenguaje en el que expresamos la estructura PA, es que en él se pueden definir todas las *funciones recursivas*. Una función recursiva no es sino una relación específica entre dos objetos, bien pueden ser abstractos o físicos. Las funciones recursivas establecen esa relación entre por lo menos dos objetos. La expresión de las funciones no obstante tiene un lugar o muchos otros. La función $f^1(x)= x+2$, tal que $f^1(0)=0+2$, relaciona al número natural *cero*, con el número *dos*. Expresada en su forma de “$f^1(x)$”, esta es una función monádica, aunque relaciona a su argumento, con el resultado de seguir su dirección (algoritmo). Una función puede ser completa o parcial. Una función es completa si arroja un valor para cada elemento de un conjunto *A* e incompleta o parcial con respecto a un conjunto *A* si arroja un valor para algunos elementos de *A* pero no todos, en este caso, de números naturales enteros positivos, parcial, de otro modo.

La característica que nos interesa aquí y que tienen las funciones recursivas es su finitud, aunque no tengan límites ni espaciales, ni temporales, ni de energía. Siempre que, en este caso podamos tener, implementando una regla de inferencia válida, todas las oraciones hechas verdad a partir de la PA+LPO = el conjunto Th(PA+LPO) a partir de algún conjunto de axiomas *A*, tal que $A \vdash$ Th(PA+LPO), tendremos entonces que el sistema formal con el que expresamos verdades acerca de PA+LPO es completo. Si además $A \models$ Th(PA+LPO), entonces el sistema formal más los axiomas *A*, constituyen un sistema formal suficiente para modelar toda la aritmética y la lógica clásica de primer orden. Si hubiera un sistema formal tal que pudiéramos obtener Th(PA+LPO) de algún conjunto finito de axiomas, quizá ignoramos ahora cuáles son, entonces PA+LPO sería axiomatizable. Dado que el conjunto de axiomas *A* sea finito, y también así su enumeración Gödel #*A*, #*A* es recursivamente enumerable, y por tanto es representable en el lenguaje de PA+LPO. Si #*A* es recursivamente enumerable, y las reglas de inferencia son válidas y expresables quizá con el lenguaje de LPO, entonces también las reglas de inferencia son recursivamene enumerables. Las reglas de la numeración Gödel que vimos justo arriba, también son recursivamente enumerables. Todo esto, hace a cualquier relación dentro de la estructura PA+LPO, incluyendo la relación de *derivabilidad,* sea

representable en el lenguaje de PA+LPO. Si estas condiciones se sostienen, entonces Th(PA+LPO) sería recursivamente enumerable.[12]

1.2.1 Primer Teorema de Incompleción de Gödel

Llamemos D a la relación de que en PA+LPO n permite derivar m, siendo n y m conjuntos u oraciones con número de Gödel *n* y *m*.

El primer teorema de incompleción de Gödel dice que dado un conjunto de enunciados (axiomas) *A*, tales que sean un subconjunto de Th(PA+LPO), si #(*A*) es recursivo, entonces las conclusiones Cn(*A*) no son una teoría completa. Esto es tanto como decir que no hay un sistema axiomático completo y recursivo a partir del cual generar todas las verdades de PA+LPO.[13] Supóngase que se dá que *A* es un subconjunto de Th(PA+LPO), entonces Cn(*A*) es también un subconjunto de Th(PA+LPO). Si Cn(*A*) es completo con respecto a la estructura PA+LPO, esto es si se da que Cn*A* no sólo es subconjunto de Th(PA+LPO) sino que Cn(*A*) = Th(PA+LPO), entonces #Cn(*A*) sería recursivamente numerable. Si esto fuera así, entonces la función que defina a #Cn(*A*), sería también definible en PA, y por tanto en PA+LPO. Si esto fuera así entonces esa función sobre números naturales, llamémosla *f*, expresada por una fórmula β, al ser negada (tal que ¬β) por el lema de punto fijo (que dice una expresión cualquiera σ es idéntica a $\beta(S^{\#\sigma}0)$, es decir a su número Gödel, tal que ¬β(#β)) diría indirectamente que β no es verdad de sí misma. Así suponiendo que β define recursivamente a Th(PA+LPO), como lo hemos hecho, el lema punto fijo de ¬β, nos dice que:

$$\models^{PA+LPO} [\sigma \leftrightarrow \neg\beta(S^{\#\sigma}0)]$$

Lo que significa que, Th(PA+LPO) es tal solamente si

$$\models^{PA+LPO} \sigma \text{ si y solamente si } |\neq \beta(S^{\#\sigma}0)$$

De tal forma que σ es verdad, pero el número Gödel que le asigna β, o bien falsea a Th(PA+LPO), o bien simplemente σ no pertenece a #Th(PA+LPO). A este nos referiremos aquí como el Teorema de Gödel I. QED.[14]

---

[12] Enderton, traducción de Amor, pp. 334-340. Teorema 34A, Coroloario 34B, Lema del punto fijo.
[13] Enderton, traducción de Amor (pp. 339-340.) Teorema de indefinibilidad de Tarski (1933), y Teorema de Incompletud de Gödel (1933).
[14] Debo la aguda observación al doctor Cristian Gutiérrez que hay demostraciones del teorema de incompleción de Gödel que no echan mano de la noción—semántica – de verdad sino mucho menos comprometidas ontológicamente, como es la noción—sintáctica – de *derivabilidad*. Es importante tomarlo en cuenta debido a que esta posibilidad podría fortalecer distintos AGs al aplicar las Gödel-limitaciones no sólo a agentes racionales,

Hay más detalles técnicos que dejamos de lado al respecto de si podríamos resolver esta cuestión. Por ejemplo aceptar a σ como un axioma que no es elemento de Cn(PA+LPO), pero sí de Th(PA+LPO). Dado que σ es Gödel numerable no hay una numeración Gödel que preserve a la vez coherencia de Th(PA+LPO) o completud de Cn(PA+LPO).

Esta función de mapeo de PA+LPO a PA que es la *numeración Gödel*, logra varios objetivos cruciales para el teorema. Hace corresponder a cada expresión, y por tanto a cada fórmula de PA+LPO uno y sólo un número, mismo que a su vez puede ser descompuesto por factorización, para conocer la fórmula que le corresponde. La numeración Gödel es una función biyectiva de mapeo entre PA y PA+LPO. Contraintuitiva aunque no contradictoriamente,

$$|Cn(PA+LPO)| < |Th(PA)|$$

La razón por la que considero contraintuitiva, incluso quizá paradójica a la afirmación anterior, es, en corto, porque si la noción de *derivabilidad* es recursiva y captura deducción a cabalidad, parecería que cuando menos es forzoso que las verdades derivables en PA+LPO deberían ser si no del mismo tamaño en número que las verdades del sistema, sí más que ellas. Pero no es así: las verdades del sistema superan en número a las fórmulas derivables del sistema. En especial si de cada fórmula se deriva válidamente ella misma.

Como un balance general; tenemos una estructura PA, que no puede ser asida por nuestro sistema de prueba deductiva más poderoso, LPO+PA. Así LPO+PA es incompleto o incoherente, aunque ésta última alternativa no es muy popular, y tenemos a estos teoremas como prueba de incompleción. Un diagnóstico que hago es que el conjunto de las posibles relaciones de números positivos de PA (una teoría muy importante entre otras cosas porque permite definir un conjunto muy importante de funciones recursivas) es más grande que las relaciones sintácticas (también funciones) de LPO, de tal forma que LPO+PA no puede coherente y completamente abarcar a PA.

---

artificiales o no, capaces de aprehender verdades, sino meramente capaces de derivar conclusiones. En este caso, me sirve más el resultado más fuerte, debido a que el mismo no presupone fuertemente que sea Turing-testeable. Más detalles se verán en los capítulos subsiguientes.

1.2.2 Resumen; presentación del teorema de Incompleción de Gödel I en AGs

Nos quedamos con una versión de juguete de la incompleción Gödeliana. Para todo sistema formal S puede modelar PA+LPO, consistencia y completud de S, entonces significa que la oración (Gödel de S) G^S es derivable. Si S es consistente y es completo, entonces G^S. G^S dice que no es derivable. Si G^S es derivable, entonces G^S es falso y así S es inconsistente. De lo contrario, si G^S no es derivable, sería verdad y S sería incompleto. Arriba, nosotros vimos una versión más detallada de este mismo razonamiento haciendo énfasis en una forma legítima de construcción de G^S. Es por eso que en esta tesis no nos quedamos con una versión tan sencilla de la incompleción de Gödel como la presentada en este párrafo porque queremos que se entienda perfectamente cómo funciona la representación del propio predicado de *derivabilidad* dentro del sistema formal usual de PA+LPO.

1.3 Susceptibilidad y limitación Gödel

1.3.1 Susceptibilidad Gödel

¿Qué agentes pueden comprender y evaluar al teorema de Gödel? En otras palabras, ¿qué sistemas pueden derivar o verificar una oración Gödel? En algún sentido lo que nos importa no es saber si algún sistema puede decidir si la oración Gödel es derivable o no, sino además que lo haga justificadamente. Podemos tener un péndulo cuyo movimiento interpretemos como 'sí' o 'no' en tiempo infinito, presentarle los teoremas de LPO y de LPO+PA, uno por uno, de tal forma que cuando lleguemos a la oración Gödel, el péndulo y cierta interpretación puedan decidir la teoremicidad de la oración Gödel, sin embargo, consideramos a esta forma de decisión una forma de decisión que deja más problemas, **tanto metafísicos como epistémicos**, que soluciones.[15]

Tenemos al menos dos tipos interesantes y distintos de sistemas; sistemas teóricos y sistemas físicos. Los sistemas teóricos son sistemas de proposiciones, cuya explicabilidad física es por lo menos discutible. Por sistemas físicos entenderé aquellos sistemas cuya existencia y

[15] Hay algunos intentos seriamente propuestos parecidos a **este**. Véanse las máquinas de Putnam-Gold caracterizadas en Copeland (2002, pp. 466-467).

funcionamiento corroboramos poco controversialmente de forma empírica; sillas, puertas, mares, calculadoras, cerebros y computadoras personales.

La *susceptibilidad Gödel* es la posibilidad de que dentro de un sistema teórico aunque no necesariamente sólo en sistemas teóricos, **se derive** una oración Gödel. En este sentido, la Aritmética de Peano (AP) sola no es un sistema en el cual se pueda generar una oración Gödel, le hace falta una teoría de la prueba además de la posibilidad de autorrepresentar esa misma teoría. Tampoco dentro de una teoría de la prueba como la Lógica de Primer Orden (LPO) sola se puede construir una oración Gödel. Hace falta poder adoptar ciertas definiciones *recursivas* para poder desarrollar una forma de representar el predicado de *prueba* dentro del sistema.

En una ontología consistente con la mecánica clásica[16] es mucho más difícil demostrar que hay susceptibilidad Gödel. Ningún sistema físico en el sentido newtoniano, tiene a su disponibilidad la energía o espacio o materia suficiente como para causar o ser efecto, y así, comprender el infinito, algo que, alegadamente, hacemos ciertos sistemas físicos como somos los seres humanos. Este apunte aunque preciso y bien reputado[17], puede ser al menos provisionalmente suspendido con la intuición de que sabemos cosas acerca de ciertos infinitos de cardinalidad $\aleph_0$. Por ejemplo, la siguiente afirmación es obviamente verdad para una cantidad infinita de casos:

> i. Para cualquier par de números enteros positivos mayores que cero, la suma de ambos siempre es mayor que cualquiera de ellos tomados solos.

Admítase al menos provisionalmente por *mor* del argumento la posibilidad de que un sistema físico de tamaño limitado, como el cerebro humano, es Gödel susceptibilidad. Pienso en que alguna forma de aprehensión de conjuntos infinitos basta para explicar de forma muy clara que podamos *saber* algo acerca de él. Si la forma de aprehensión del infinito es de cierto tipo, entonces eso puede bastar para demostrar que las mentes humanas no son consustanciales con las MTs, y probablemente tampoco óptimamente representables por medio de una teoría Turing computable.

[16] Voy a pasar por alto en este momento, la posibilidad de que la mecánica cuántica refute a una ontología sólo consistente con la mecánica clásica (o newtoniana). Personalmente no considero que sea sostenible ni siquiera a la luz de criterios de interés puramente teórico una ontología no clásica que hagan verdadera a una oración "p & no-p" o bien que presuponga algún infinito actual.
[17] Véase la refutación kripkeana del disposicionalismo en su Wittgenstein On Rule Following (1982).

Los requisitos para que haya susceptibilidad Gödel, deben permitirnos incluir a sistemas físicos (como personas humanas) tanto como sistemas teóricos. La susceptibilidad Gödel tiene, en un nivel, sólo un requisito; la capacidad de comprender o entablar un diálogo, sacar implicaciones o algún predicado que involucre la posibilidad de comprender o fingir que se comprende o domina un lenguaje con ciertas características. Digámosle en adelante simplemente "comprensión de lenguaje" a la mera capacidad de tomar una afirmación y obtener ciertas implicaciones, y quizá actuar de acuerdo a ellas. Abajo aclaramos más en qué consiste este requisito. Una forma de entender la expresión "LPO+PA comprende un lenguaje" es usando una reducción analógica simple; LPO+PA puede entablar un diálogo sobre si cierta afirmación es un teorema o no, y quizá algunos otros tipos de conversaciones, dependiendo de qué tenga como axioma o a partir de qué información use como punto de partida.

1.3.2 Requisitos de la susceptibilidad Gödel en AGs

Hemos establecido arriba cuáles son los requisitos mínimos que debe tener cualquier sistema para ser susceptible Gödel. El requisito debe ser suficientemente flexible como para incluir algo más que teorías, dado que, el fenómeno de la teorización es objeto de explicación física. Esto es lo que pretenden las ciencias cognitivas y la Inteligencia Artificial. Hemos llegado a la idea de que el único requisito que debe cumplir un sistema es que pueda entender un *tipo especial de* lenguaje.

Pasaremos de largo por cuestiones prácticas en torno a qué basta para atribuir a un agente aptitud para un lenguaje. Un agente capaz de lenguaje debe poder conducirse externamente con al menos dos conductas debidamente interpretadas como *sí* y como *no*. Una moneda lanzada al aire además de ciertos supuestos, satisface en condiciones especiales esta descripción y en este momento no molesta dejarla dentro de los sistemas (físicos) lingüísticamente aptos. Algunos candados adicionales se activarán cuando especifiquemos qué clase de lenguaje debe poder ser comprendido por el sistema en cuestión.

Los requisitos que debe tener el lenguaje del que el agente Gödel susceptible son los siguientes:

1. Un lenguaje o un sistema en el que podamos representar cualquiera de sus propios términos o conjunto de términos, oraciones y conjuntos irrestrictos de oraciones, consigo mismo.
2. Las unidades mínimas de expresión deben ser oraciones declarativas o descriptivas i. e., con condiciones de verdad (y no sólo normas, por ejemplo**, sean categóricas o hipotéticas**).
3. Un catálogo básico de símbolos para significar un objeto **singular** (constantes), para significar un predicado (constantes), para significar *al menos algún* objeto (variables parciales de objeto), cuantificadores existencial y universal, negación e implicación material.
4. La manera de expresar extensionalmente una clase de funciones recursivas (por ejemplo, una forma **finita y auto-referenciable** de enlistar a los números naturales).
5. La manera de definir a partir de las funciones recursivas tomadas extensionalmente, otras funciones recursivas y predicados. (Este requisito es análogo del *axioma de elección* pero mucho más débil).

Estas condiciones lingüísticas bastan para poder generar una función que sea extensionalmente igual al predicado 'es demostrable'. La mayor parte de los lenguajes naturales cumplen con estos requisitos, aunque sería interesante evaluar los lenguajes naturales que no contaban expresiones para *números naturales* o *conjunto*, esta inquietud **no es necesaria** para los fines de esta discusión.

Supongamos un caso extremo en el que un sistema físico podría ser susceptible de decidir si la oración Gödel es demostrable o no. Dejando de lado las formas idóneas y los obstáculos prácticos que bastarían para, en efecto saber qué sistemas físicos son capaces de un lenguaje con las características enlistadas de 1 a 4, mencionaré un caso en el que creo que un sistema físico extremadamente parco podría ser tenido como Gödel susceptible.[18]

[18] *La Bocca della Verità*. En Roma hay una piedra labrada con una cara de un hombre barbado. En la boca del hombre barbado, que es un orificio pequeño, cabe la mano de una persona. Dice una leyenda popular que en la *bocca* se esconde alguna criatura mitológica y que las personas que meten la mano al orificio de la *bocca* y dicen una mentira, la criatura les muerde la mano y la arranca. Si uno dice algo verdadero, entonces la piedra deja que saquemos la mano intacta. Si uno hace una inspección empírica exhaustiva del sitio, uno quizá no encuentre rastros de ninguna entidad mitológica, no obstante lo cual, no es concluyente que de acuerdo a cosmologías posibles la criatura mitológica y la *bocca* no sean Gödel susceptibles ***en sus propios términos***. Para que la *bocca de la verità* pueda capturar o dejar libre una mano cualquiera, debe poder *entender* el idioma para poder evaluar si lo que dice quien tiene la mano dentro es verdad o no. Además, claro está, la *bocca de la verità* necesita omnisciencia. **La leyenda no habla de que la *bocca* detecte solamente honestidad, sino verdad, por eso es que**

1.3.3 Limitación Gödel

Una vez explicada la noción de susceptibilidad Gödel sólo nos queda por ver cómo es que G para la que ~~a~~ cada sistema Gödel susceptible es susceptible, separa entre sistemas con y sin *limitación Gödel*. La forma de AG propuesta en la sección 1.2 del capítulo 1 implica tanto que la mente humana es Gödel-susceptible, como que no está limitada por la incompleción gödeliana, esto es, que sabemos que la oración Gödel apropiada es verdad (y si sabemos, pues podemos saber). La limitación Gödel es la imposibilidad de *resolver satisfactoriamente* qué estatus epistémico tiene la oración Gödel respectiva a un sistema formal o físico dado, sea uno cualquiera o el del agente que está *considerando al respecto*. Por supuesto que el estatus epistémico de G oscila entre *es teorema*, *no es teorema*, *es verdad*, *es falso*, *no es ni verdad ni falso*, etc. La resolución o decisión que nos interesa responde a dos grandes **grupos de** autoridades epistémicas; evidencia empírica y evidencia no-empírica.

Todo AG requiere para su solidez de que al menos ciertas mentes humanas sean Gödel susceptibles. Así, quienes atacan la posibilidad de susceptibilidad Gödel de todas las mentes humanas atacan la solidez de los AGs. Esta estrategia[19] bordea peligrosamente la postura de que las mentes humanas tenemos una certidumbre espuria al hacer aritmética o al razonar de acuerdo a principios lógicos como los de LPO en especial si comparamos entre diversos tipos de conocimiento. Y aunque, por ejemplo, la certeza de que 2+2 es 4 sea espuria, uno debería poder mostrar el caso fehacientemente. Aunque poner en duda que hay al menos algunas mentes humanas Gödel susceptibles, no necesariamente ponemos en duda la fortaleza de todas nuestras aptitudes aritméticas y lógicas, sí transmitiría algo de la duda a algunas de ellas.

En principio el valor y certidumbre que nuestros conocimientos matemáticos pueden ofrecer son explicados de dos formas. La primera forma es de corte subjetivista echando mano de la

---

necesita omnisciencia. En el caso de las verdades o falsedades de la lógica y la aritmética, la *bocca della verità* no requiere omnisciencia en general sino un tipo de omnisciencia más restringida: lógica. ¿Cómo es que podríamos constatar empíricamente si la piedra de la verdad maneja un lenguaje con las características mínimas para que haya Gödel-susceptibilidad? Fácilmente. Uno construye una serie de **tautologías**, tal como lo haría en un examen de lógica, y en lugar de que haya derivaciones en un papel o subrayado de respuestas correctas, pues la mano debiera quedar libre en cada cuestión a contestarse afirmativamente. También puede uno construir una serie de contradicciones y ver que quienes metieran la mano la perdieran. No importa para esta tesis la realización práctica de este examen y que tengamos un voluntario valiente dispuesto a perder una mano por la ciencia.)

[19] Véase por ejemplo, Putnam (*Minds and Machines*, 1979) y LaForte, Hayes y Ford (1998), y otros.

intuición, estados epistémicos incondicionados autoevidentes, una forma de *estado subjetivo*. La segunda forma admite verificabilidad o falseabilidad empírica aunque sea en principio.

Comprender una aseveración, aislada o sistemática, una teoría o un teorema, parece requerir que seamos capaces de extraer cierta información del mundo incluyendo proferimientos lingüísticos y, a veces, que nos comportemos de alguna manera especial dados ciertos supuestos. Por ejemplo

i. Pásame un vaso de agua
ii. El cielo es azul
iii. El ejemplo iii contiene una oración falsa

son adecuadamente comprendidos si se entiende que en i hay una orden de hacer tal o cual cosa, y para ii que hay una descripción del cielo, misma que se corrobora o falsea, buscando y consiguiendo una cierta experiencia sensorial. Pero con los teoremas o meta-teoremas **de lógica tenemos ciertas dificultades en comparación. Parece que** no hay algo así como buscar una cierta experiencia sensible como en el caso de ii para cerciorarnos de tales tesis salvo claro que hablemos de la lectura o escritura de su derivación o que alguien nos la explique, o algo parecido. Quizá iii contiene encapsulada alguna regla complicada con un hecho vacío o formal, pero tampoco es del todo claro entonces qué hacer con ella, cómo se entiende, qué implica y qué puede implicarla. Después de todo creemos que afirmaciones semejantes a iii, sean teoremas o meta-teoremas, son descripciones acerca de sistemas formales, no reglas de formación ni siquiera reglas de inferencia.

Si suponemos por ejemplo que iii **expresa una proposición,** una tesis susceptible de ser verdadera o falsa, entonces puede ser evaluada adecuadamente usando evidencia **a favor de ella, razones a favor de ella, incluso en casos límite, la ausencia de evidencia en contrario**. La evidencia empírica no es la única evidencia de la que podemos echar mano.

Además de que un agente del que evaluaremos limitación Gödel puede ser para casos extremos una piedra como la *bocca della verità*, es necesario entender qué implicaciones empíricas y no empíricas tiene el conocimiento meta-lógico, de tal forma que podamos entender qué significa para agentes limitados a respuestas sí-no, responder ‘sí’, ‘no’ o silencio o falla de detención, *halting problem*. Después de todo, todos los agentes pueden descomponerse, el preguntador impacientarse esperando una respuesta, o, quizá, habernos topado con una **afirmación cuyo valor de verdad sea** imposible **darse** en términos “sí”-“no”.

Algunos defensores de algún tipo de AG sostienen que, si el agente a cuestionar fuese un *espíritu digital*[20] *arrancando manos detrás de la* bocca della verità, preguntándole una oración semejante al primer teorema de incompleción de Gödel, nosotros podríamos sacar la mano a salvo en un tiempo anterior a infinito. Una vez transcurridos infinitos segundos el espíritu digital podría resolver, quizá, la cuestión y decidir dejarnos sacar la mano a salvo.

Un agente Gödel susceptible debe además comportarse de cierta forma para que sepamos que *no está Gödel limitado*. Por ejemplo, si se le pregunta que si G es verdad o no, pues que responda que sí, o **presentando evidencia a su favor**. Si se le preguntara si una cierta oración G circunscrita al software conforme al cual está él mismo programado, igualmente debe poder responder correctamente para que sepamos que no está Gödel limitado.

Exactamente qué comportamiento es lo que conocemos como *saber* cuál es el estatus específico de G para un sistema formal, exactamente cuál sea la solución correcta y qué características cognitivas hagan falta, será tratado en los capítulos subsiguientes.

Conclusiones de capítulo 1

Para que un argumento sea un *Argumento Gödeliano* es necesario que éste tenga como premisas algunas afirmaciones dentro de las cuales está una premisa que para ser verdad, requiere básicamente de que la incompleción de Gödel en general, o para sistemas concretos en particular, sea verdad.

No todo sistema (formal o no) podría ser Gödel incompleto. Por ejemplo, un sistema formal que tiene un máximo de teoremas posible más pequeño que el infinito más pequeño, no es Gödel-susceptible. Para que un sistema pueda ser Gödel incompleto se necesita: a) ser capaz de producir los teoremas de la Aritmética de Peano o algo isomórifico, b) ser capaz de interdefinir términos (para lo que técnicamente se conoce como *numeración de Gödel*), c) ser capaz de

[20] Nadie sabe exactamente si los espíritus de los que nos hablan las leyendas populares como la de la *bocca della verità* son o no algún tipo de sustancia espiritual digital, pero podemos incluirlos con base en las siguientes considraciones; entienden el lenguaje y pueden reaccionar en consecuencia, por lo menos algunas reacciones son empíricamente medibles con supuestos bastante terrenales que usamos incluso para personas vivas; el fantasma puso atención, quiere responder y quiere responder bien, puede responder, o sea arrancar la mano, mover el puntero de la Ouija.

deducción o un mecanismo que refleje deducción. Una vez que un sistema (formal o no) puede ser incompleto al estilo Gödel,, lo llamaremos Gödel-susceptible.

Dentro del conjunto de todos los sistemas Gödel susceptibles, hay algunos que son incompletos y al menos en principio hay algunos que pueden ser completos.

Cuando un sistema es Gödel susceptible, pero no es Gödel limitado de modo relevante es cuando ese sistema tiene a su disposición recursos cognitivos que superan *de algún modo* los recursos que se requieren para la Gödel susceptibilidad. En qué medida y cómo es que esto sucede, es objeto de la discusión alrededor de las condiciones de solidez de los AGs.

> For my own part, I think that if one were looking for a single phrase to capture the stage to which philosophy has progressed, 'the study of evidence' would be a better choice than 'the study of language'.
>
> — A.J. Ayer, Philosophy in the Twentieth Century

> Of Hume we may say not merely that he was not in practice a metaphysician, but that he explicitly rejected metaphysics. We find the strongest evidence of this in the passage with which he concludes his Enquiry Concerning Human Understanding. 'If', he says, 'we take in our hand any volume; of divinity, or school metaphysics, for instance; let us ask, Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames. For it can contain nothing but sophistry and illusion.'
>
> — Ayer, A.J., Language, Truth, and Logic, 1946.

Se lee en Philosophische Untersuchungen de Wittgenstein, que "§580. *Ein ›innerer Vorgang‹* ***bedarf*** *äußerer Kriterien.*" ("Un 'proceso interno' **necesita** criterios externos", con mis negritas.) Pero ¿es esto cierto? El verbo *bedürfen* se usa en alemán también para expresar el mismo tipo de necesidad de los trámites y demandas jurídicas, y ¿qué necesidad hay más débil que la de los trámites y exigencias si es comparada con la necesidad de las leyes aritméticas o de la física, por ejemplo? Aún así, cuando los *requerimientos* de los trámites son idénticos a los *requerimientos* de la causalidad física o las leyes aritméticas, el trámite adquiere en esa medida, la fortaleza modal de esas leyes.

2. Tribunales de la evidencia diferentes, diferentes *Argumentos Gödelianos*

En este capítulo ofrecemos una forma de distinguir AGs con base en los compromisos epistémicos que tienen las premisas. En corto si la verdad de las premisas sólo pueden determinarse en principio por evidencia Turing computable, parece que de entrada la conclusión de los AGs no podría ser verdad. De este modo es que se vuelve necesario entender bien qué clases de evidencia estamos dispuestos a aceptar para establecer la posible verdad de las premisas. El espíritu de la distinción de este capítulo es básicamente tratar de deshacer una ambigüedad importante. El asunto de qué significa cada término y en qué consiste referir, proposición, etc., en parte determina si algún enunciado se ha de considerar verdadero o no. De modo que si la única evidencia o razones admisibles para éste efecto son sólo las computacionales, pues todo enunciado que para ser verdad requiera de un mundo no computable pues será falseado en principio por razones de lenguaje o de la teoría de modelo y no de cómo es de hecho el mundo. Otro asunto es que sepamos qué clase de evidencia admitimos para decidir si alguna premisa específica es cierta o no, y otra es de hecho conseguir esa evidencia y decidir una u otra postura. Una vez que se consiguiera la evidencia adecuada o la justificación adecuada, podríamos saber si las premisas son ciertas o no, y en esa medida, si algún AG es válido, entonces sabríamos que el computabilismo es falso.

Una vez explicado y discutido lo anterior, dedicaremos las partes finales del capítulo a entender cómo afecta nuestro tribunal de evidencia predilecto al tipo de AG que aceptamos, y el tipo de críticas que le hacemos o que puede recibir. Un tribunal de la evidencia sea computacional o Turing testeable, haría que creamos que incluso ningún cerebro humano puede ser siquiera Gödel-susceptible. Analizamos pues a la luz de nuestro tribunal de evidencia favorito, los efectos que ellos tienen en los AGs que se nos puedan presentar por la vía

2.0 Introducción; evidencia para la susceptibilidad Gödel y la limitación Gödel

En el capítulo anterior revisamos lo que define a un AG. Un componente esencial de tales argumentos es la supuesta atribución de *limitación Gödel* a ciertos agentes racionales (incluídos aquí por brevedad también *sistemas formales* y *objetos abstractos* como *máquinas*

de Turing). De la manera en la que hacemos tal atribución es de la manera en la que determinamos *qué es que un agente racional* sepa que *la oración Gödel es verdad*. Agrupamos aquí dos grandes tipos de evidencia como evidencia *ad hoc* para justificar tales tesis. El primero es evidencia conductual, con ciertas características asequibles intersubjetivamente, mientras que el otro grupo es el de la evidencia subjetiva que no es asequible intersubjetivamente.

En este capítulo se analizan una serie de nociones con varios objetivos. La primera es la de *Test de Turing*, los alcances y variantes que son útiles para los fines del presente trabajo. El *Test de Turing* es una prueba básicamente conductista para determinar características tan relativamente ambiguas y escurridizas como la de *femineidad*, y *pensamiento*, y sin muchos problemas, otras semejantes.

Una de las características más importantes, no obstante es el tipo de evidencia al que el *Test de Turing* se limita y es la evidencia pública (intersubjetiva u objetiva) y empírica. Un tipo de evidencia diferente para movernos a creer en alguna proposición es por tanto la evidencia *privada* (la llamo subjetiva) y que argüiblemente no es empíricamente accesible. Una división más que es prima con respecto al origen de la evidencia (empírico y público, contra privado y no necesariamente empírico), puede ser el de interno y externo. El comercio entre lo interno y lo externo, es materia de complicaciones y descomprensiones que afectan la fortaleza y solidez, por supuesto, de los AGs. En este capítulo buscamos deshacernos de ellas.

Un objetivo de este trabajo, también importante, es el de esbozar el tema de la *mejor justificación* o *justificación adecuada* para creencias meta-lógicas, aritméticas y lógicas. Hay *argumentos gödelianos*, como el de Putnam (*Reflexive Reflections*, 1985) que hacen énfasis no en la capacidad de una computadora para, de forma gruesa, imprimir una derivación de *su propio* teorema de incompleción de Gödel o *su propia* oración Gödel, sino en la veracidad de la misma y en el hecho de que nosotros tenemos acceso a la comprensión de ella. Justamente el criterio de evidencia externa, misma que entiendo como empíricamente accesible y pública, e interna, misma que entiendo como subjetiva y probablemente no empíricamente accesible y quizá también, privada.

Debe notarse aquí el siguiente compromiso de este trabajo de tesis: al separar entre verdades por razón de evidencia de distinta naturaleza, estoy tomando provisionalmente a la evidencia

como algo muy parecido a suposiciones o razones. Un anti-realista podría decir que no hay evidencia alguna de la existencia de tal o cual objeto o situación, pero que *por la cultura*, *la sociedad*, *la historia*, *la idiosincrasia*, *los fetiches*, *la cosmovisión*, etc., de tal o cual sociedad o persona, pues se acepta que existe un objeto tal o que ese objeto tiene cierta característica. El anti-realista tiene que aceptar que o bien el universo es realmente de manera tal que cierto grupo de cosas o situaciones no existen, lo que es contradictorio, o bien tiene que admitir que *consideramos como verdaderas ciertas tesis por razones distintas a la realidad de ellas*. Estas razones aquí hacen las veces de evidencia. Quizá con trabajo más puntual podamos descubrir que esta clase de posturas filosóficas o bien son en el fondo realistas, o bien simplemente admiten evidencia subjetiva de calidad pobre (indirecta o irrelevante) como evidencia para admitir la verdad de cierta tesis. Por esto es que he tomado como una división más o menos exhaustiva entre evidencia finitaria y empírica, no finitaria y empírica, por un lado y finitaria y no empírica, y no finitaria y no empírica por otro lado. Al primer lado le llamo *Turing testeable*, y al otro lado subjetivamente testeable o no *Turing testeable*. Dicho de otra forma; no hay anti-realistas absolutos, todos los anti-realistas tienen que admitir por lo menos razones como elementos de juicio para el efecto de conseguir (o perder) creencias, y así debe entenderse "evidencia" en adelante, y "evidencia TTeable" y "evidencia no-TT", como subtipos.[21]

2.1 Test de Turing, límites, alcances, variantes y tergiversaciones

Normalmente, un niño en la escuela básica tiene que hacer algunas sumas para que la maestra **determine** si ya es capaz de sumar. No basta con el dicho del niño de que sabe sumar, tiene además, que mostrar suficiencia en el tema de la suma. En casos típicos no le pedimos al niño que haga una semblanza fenomenológica de cómo se siente entender la suma. Sin embargo, un niño con una tabla de sumas puede, en teoría, hacer bien su prueba, sin tener idea de cómo se suma. El *Test de Turing* (TT, en adelante) o como lo conoció Alan Turing (1950), *juego de imitación* (*the immitation game*, en palabras de Turing), es una forma de medir si algún agente posee o no cierta capacidad intelectual en general, si tal agente sabe aritmética y lógica, por ejemplo. El test tiene ciertas características las cuales han sido criticadas de vez en cuando por los filósofos, como su compromiso conductista, digamos, conductismo funcionalista. Veamos

[21] Para más detalles sobre esta reformulación de la disputa entre realistas y anti-realistas, en general, sirve apelar a "*Realism*" de Dummett, publicado en Synthese en 1982, y de ahí, a una plétora de filósofos incursionando en esta manera de reacomodar el debate.

para el tema de los AGs, la *Gödel susceptibilidad*, y la *Gödel limitación*, la herramienta del TT, variantes y alternativas.

Algunos filósofos podrían defender la tesis de que no hay una manera completamente **confiable** de saber si, por ejemplo, una computadora puede derivar una oración Gödel específica a una teoría LPO+PA con un predicado de *demostrabilidad* general o específico de sus particulares procesos de inferencia y demostración. Quizá el ejemplo siguiente ayude un poco: las personas sabemos que nuestros sentidos nos pueden engañar, en general, pero saber que uno mismo tiene tal o cual instancia de sordera, debilidad visual o sesgo racional, es un asunto hasta cierto punto distinto. Otros filósofos pueden decir que hay algunos modos legítimos para que los teoremas de Incompleción de Gödel nos permitan diferenciar entre mentes humanas y computadoras a razón de diferenciar un humano talentoso para razonar, y una máquina de Turing abstracta, con un conjunto de axiomas suficientemente genérico. Las diferentes combinaciones se han presentado por ejemplo en la distinción entre *los propios contenidos mentales* (subjetividad) vs. *los contenidos mentales de terceros* (intersubjetividad), *los contenidos mentales propios* (subjetividad) vs. *los contenidos mentales de un género al que pertenece cierto agente* (intersubjetividad potencial, o incluso objetividad), y así. En este capítulo tocamos este asunto.

Según el texto, ahora clásico de Alan Turing, *Computing Machinery and Intelligence* de 1950, la pregunta acerca de si una máquina piensa o no, es demasiado ambigua para recibir un tratamiento significativo. No obstante para Turing (1950) podemos clarificar la cuestión mediante un *juego de imitación*. Si dos agentes dialogando uno de los cuales es una computadora, parecen indiscernibles a un interrogador que atiende sólo a los proferimientos no manuscritos de la conversación que tiene con ellos, entonces la capacidad intelectual de las máquinas es indiscernible de la de los seres humanos. Así, por imitación conductual, sabemos que las máquinas pueden *pensar*. En el juego de imitación, el interrogador puede preguntar cualquier cosa a cualquiera de los dos agentes. Turing nos da una tercia de ejemplos de lo que puede contener las tarjetas con tipografía uniforme que se intercambian entre el interrogador y los agentes interrogados:

> *"P: Por favor escríbeme un soneto con el tema de Porth Bridge.*
> *R: No cuentes conmigo para esta. Yo nunca podría escribir poesía.*
> *P: Suma 34957 a 70764.*
> *R: (Pausa de cerca de 30 segundos y entonces ofrece respuesta) 105621.*

P: ¿Juegas ajedrez?

R: Sí."[22]

Con base en las respuestas dadas al interrogador, él, al final concluye si alguno de los dos agentes, por ejemplo, era una máquina incapaz de platicar o si carece de alguna otra capacidad intelectual o cualidad mental general, como pensar, y más específica, como si algún agente es capaz de aritmética o de lógica formal, por ejemplo. Un test para probar la capacidad para entender o hacer aritmética y lógica es lo que entenderé por TT en este trabajo en adelante si se usa sin más.

Una variación del TT la entenderé como un *test* igualmente conductista aunque o bien, cambia la pregunta que se resolverá con él, por ejemplo, de "¿puedes pensar una computadora?" a "¿puede bailar una computadora?" o "¿puede declamar poesía una computadora?". La forma de conductismo a la que se circunscribe el TT como originalmente lo propuso Turing, no es incompatible con escalamientos hacia ningún tipo de funcionalismo; la caja negra del conductismo clásico es solamente la tabla de transiciones de la MT y ésta a su vez se puede especificar hasta llegar a umbrales de voltajes de cualquier intensidad. Sin embargo, una variante del TT es un *test* cuya evidencia busca responder a la misma pregunta pero con cambios en la forma y objetivo de realización, por ejemplo, admitiendo más de un interrogador, circunscribiendo el tema de la pregunta, la manera de realizarlo, por ejemplo ya no mediante tarjetas para el diálogo, sino mediante una cámara web, o mediante rayones en un pizarrón. Veamos ahora las particularidades del tema de los TTs para los AGs.

Hemos visto ya el problema al respecto de qué es *susceptibilidad Gödel* y qué cosas pueden tener esa propiedad. El comportamiento lingüístico es empíricamente mesurable. El comportamiento lingüístico tiene esencialmente un componente físico, pero hay ciertos contenidos semánticos (referentes de conceptos, instancias aceptables de uso, etc.) cognitivamente importantes que son, algunos filósofos han argüido, necesaria o contingentemente privados, lo que implica que no son empíricamente mesurables, no al menos directamente. Un caso típico de esta clase de referentes son los estados mentales de dolor y también la información obtenida de la experiencia de ver colores.

[22] Mi traducción de: "Q : Please write me a sonnet on the subject of the Porth Bridge. / A : Count me out on this one. I never could write poetry. / Q : Add 34957 to 70764 . / A : (Pause about 30 seconds and then give as answer) 105621./ Q : Do you play chess ?/ A : Yes." Turing (1950, p. 35).

Para algunos filósofos y corrientes filosóficas al menos algunos estados mentales privados jugaron un rol epistémico justificador mucho muy ambicioso. El caso de Descartes es paradigmático. Descartes responde al problema de si el conocimiento es siquiera posible con la certeza del *Cogito cartesiano*, que era garantizadamente verdadero debido al acceso que tiene *el meditador* a ciertos estados mentales privados. *El meditador cartesiano* no es sino el agente racional que sigue en primera persona el curso intelectual de las Meditaciones Metafísicas y del Discurso del Método de Descartes.

Otra manera de glosar el problema de si es imposible que hagamos atribución de Gödel susceptibilidad por medio de evidencia accesible a la tercera persona (intersubjetiva, o objetividad pública) es el de si el pensamiento es condición necesaria para que haya capacidades lingüísticas y si el pensamiento es sólo de acceso privado, entonces la evidencia empírica podría no ser la mejor forma para saber qué agentes son o no Gödel-susceptibles . Si algún tipo de evidencia empírica indirecta—como proferimientos particulares, sirven para *explicitar el contenido del pensamiento*, nos dejaría todavía con la duda legítima de si el pensamiento adecuado explica la conducta lingüística correcta; el problema de la suma y la *tasuma*, del Kripkenstein y otros semejantes. El problema pues de si el pensamiento es, por decir así, metafísicamente anterior que la capacidad lingüística, si son simultáneos o si es al revés, es un problema que determinaría si la evidencia empírica es el mejor tribunal para saber si hay algunas mentes humanas que superan o no los límites de la Turing-computabilidad al no estar Gödel limitados. En un tipo de casos límite, no necesitamos saber si *alguna mente humana* en general, sino si al menos la propia mente (de la que se tiene acceso subjetivo) supera ese límite[23].

Un objeto que podemos conocer empíricamente que es además Gödel-susceptible puede ser una persona humana y otro quizás una computadora avanzada. En principio cualquier objeto que sea empíricamente cognoscible y que tenga comportamiento lingüístico sin importar nada más, como vimos secciones arriba, es tanto *Turing testeable* como Gödel susceptible. Para saber si algo es Gödel susceptible, se admiten dos grandes grupos de evidencia: empírica finitista y no empírica finista.

[23] Parece que hacer investigación cualitativa a partir de la subjetividad escaparía, por definición de toda evaluación con afán de objetividad. Deberíamos resistir esta conclusión, aunque el tipo de problemas epistemológicos que impondría, son importantes. No hay ninguna razón por la que tengamos que excluir la posibilidad de que haya una ciencia sistemática del estudio de la subjetividad. Algún ejemplo filosófico serio viene de fenomenólogos como Husserl.

El primer grupo de evidencia es la del comportamiento lingüístico revisable empíricamente, y el segundo grupo no necesariamente, el segundo grupo es de cierta evidencia mental accesible a la primera persona si es que la distinción entre sentidos e introspección agotan todas las fuentes de información nueva a disposición de los sujetos cognoscentes. En la medida en la que los dos grupos de evidencia son independientes uno del otro y dependiendo de cómo se construyan y cómo sea nuestra teoría ontológica favorita, el segundo grupo de evidencia suele presentarse como desconectado o no relacionado con la evidencia de tercera persona finitista o empíricamente evaluable. Este es otra instancia del problema de qué tan diferentes son ontológicamente el pensamiento y la materia, cómo se relacionan entre sí y cómo es que conocemos esas relaciones. Llamaré 'subjetivo', un poco desprolijamente, a todo supuesto hecho que no se conozca de forma alguna, idóneamente incluso, por medio de la información sensorial finita A la información que no puede ser adquirida directamente por los sentidos, sea cual sea la razón, la llamaré *datos subjetivos*, contrastándolos con los *datos de los sentidos* (*sense data*). Quizá la evidencia por la que una tautología genuina es verdad, es un tipo de evidencia, digamos, subjetiva en la medida en la que sea independiente de la evidencia empírica finitista. Los datos subjetivos hacen posible lógicamente que haya un desacuerdo entre una persona y un grupo de expertos junto con el resto de la población y que quien tenga la razón sea el grupo numéricamente más desaventajado o con credenciales epistémicas menos importantes. La desventaja puede ser no sólo por el número de personas, la calidad epistémica de ellas (si son expertos o no), las teorías científicas que tengan a su alcance, e incluso la información que tengan. Esta es una forma de expresar una idea simple: una persona puede sentir dolor verdaderamente sin importar si ello contradice alguna teoría respaldada empíricamente que indicara lo contrario en ese caso. (Para este efecto se citan casos de dolor en miembros del cuerpo después de haber sido amputados, por ejemplo.)

Así, para saber si algo es empíricamente Gödel susceptible necesitamos evidencia de que un agente específico tiene capacidad lingüística en general y aritmética en particular. Con esto garantizamos que un agente puede, por lo menos en principio, describir e inferir. Para saber si algo es lingüística o aritméticamente competente tenemos una prueba clásica en el campo de investigación en Inteligencia Artificial es la conocida como *Test de Turing* (TT). Es una prueba que Alan Turing reformuló para aplicarla en estos casos en su artículo de 1951 y que planteó entre otras cosas para evaluar la inteligencia de computadoras a través de las habilidades dialécticas de una máquina en un contexto irrestricto de preguntas y comentarios. Es por lo

anterior que el TT que propuso originalmente Turing es una prueba demasiado exigente para estimar que un cierto agente es Gödel susceptible. La razón es que no necesitamos un agente, artificial o no, que tenga habilidades de conversación (manejo lingüístico) especialmente desarrolladas como para mimetizarse con las humanas de forma irrestricta; en las artes, leyes, ciencias, cultura general, temas de sentido común, habilidades de razonamiento no deductivo, etc. Para establecer Gödel susceptibilidad en agentes racionales sólo necesitamos saber que ellos tienen un mínimo dominio de ciertas operaciones aritméticas y la utilización de ciertos predicados como el de igualdad, el de desigualdad, y la posibilidad de expresar internamente el predicado de derivabilidad, estipulaciones mínimas para realizar gödelización.

Es importante sobre todo la posibilidad de expresar el propio predicado de verdad para los agentes que sean *Gödel susceptibles* porque para poder siquiera escribir los teoremas de incompleción de Gödel es necesario podernos referir a las fórmulas del propio sistema, y al valor de verdad de las mismas. El valor de verdad de las fórmulas de, por ejemplo un sistema formal suficiente para aritmética y lógica de predicados de primer orden, es condición de posibilidad para entender la noción de derivabilidad que es la que modela principalmente la lógica de predicados de primer orden. Si no pudiéramos referir a la verdad de las fórmulas del sistema, perderíamos la posibilidad de expresar la noción de *derivación* y la de *prueba*,.

Para ser Gödel-susceptible quizá no es necesario poseer habilidades lingüísticas exactamente iguales a las humanas , y quizá tampoco es cierto que todo agente con habilidades lingüísticas humanas, i.e. cualquier conversador humano, sea Gödel-susceptible en el siguiente sentido. Cabe la posibilidad de que haya una persona capaz de comunicarse usando un lenguaje natural humano, pero que por cierta incapacidad o falta de oportunidad, no desarrollara habilidades mínimas ni de reflexión (hablar de sí mismo, solamente o de sí mismo a través de hacer teorías generales que lo incluyan a él mismo) o de inferencia consciente y explícita, o no-consciente e implícita. Muchas personas sostienen que los estados mentales de consciencia en realidad son ilusiones ya que el cerebro procesa la información que nos parece categóricamente presente. Si este proceso es, por lo menos en parte, consustancial al proceso psicológico de inferir, entonces no es necesario que el proceso cerebral por el que inferimos no-conscientemente *ejecute siempre inferencias explícitas* sino que la limitación o no-limitación Gödel podría, en potencial al menos, alcanzar a humanos, p.ej., que no razonan y sólo perciben con ese aparato cerebral que *decide* implícita y no-conscientemente. Nótese que no usé el término ‘inconsciente’ por la posible confusión a que nos arriesga el uso freudiano del mismo.

Putnam, Maudlin (1995), Kirk (1986) y Kripke (1982), ofrecen distintas razones al respecto de que ningún objeto físico empíricamente determinado tiene una competencia aritmética o lógica suficientemente grande y buena como la tiene por ejemplo la *Aritmética de Peano* o cierto aritmetizador humano idealizado. Las calculadoras, computadoras y humanos, concretos o idealizados, tienen capacidades limitadas. Las calculadoras tienen sólo un espacio limitado en sus pantallas de tal forma que no pueden realizar operaciones con, típicamente, resultados de más de nueve dígitos, las computadoras tienen además de una limitación como la de las calculadoras, la limitación que tenemos los humanos, a saber, que el transcurso del tiempo termina por poner un límite a las operaciones aritméticas que podemos de hecho, realizar.

La objeción anterior, aunque interesante de suyo, no es una objeción que ataña directa e irrefutablemente a los *argumentos gödelianos* dado que no nos interesa saber si los humanos pueden de hecho probar todos los teoremas de PA, sino que probando los teoremas de incompleción de Gödel, los humanos, limitados como pueden estar para PA, son intelectualmente más poderosos que cualquier máquina para computar en el asunto especial de conocer su propia consistencia de tal forma que puedan acceder completamente a toda la fortaleza de las verdades, no sólo aritméticas, sino inferenciales. Si los seres humanos no muriésemos, si el universo no se enfriara, si no desistiéramos por aburrición y algunas otras cláusulas similares podríamos teóricamente enlistar todos los números naturales, los teoremas de PA y los teoremas de un sistema de Lógica de Primer Orden y Aritmética, incluso aquellos que refieren a la propia consistencia por vía de diagonalización o por medio de gödelización.

La cantidad y calidad de la evidencia empírica que basta para justificar la tesis de que un agente posee la capacidad de realizar cómputos aritméticos iguales al menos a una cierta clase de funciones definibles en la *Aritmética de Peano* puede estar sometida a discusión. El interés en este tema está principalmente en la Filosofía del Lenguaje y la Filosofía de las Matemáticas, además de la Epistemología, claro. Por esta razón, es que para el presente trabajo, tengo cierto escepticismo en que la discusión anterior sobre la atribución de capacidad aritmética sea una objeción parcialmente interesante en contra de la atribución empírica de Gödel-susceptibilidad a agentes lingüísticos.

El criterio de Gödel-susceptibilidad es suficientemente débil como para dar cabida –i.e., para no prejuzgar- a diferentes respuestas en debates en filosofía del lenguaje (al respecto de

capacidades lingüísticas, aritméticas o no) y en Filosofía de las Matemáticas. Divido así al tipo de evidencia a favor del establecimiento de que un agente A o un tipo de agentes A es o son Gödel susceptibles. La primera evidencia es perfectamente aceptable para el naturalismo filosófico y el intuicionismo en matemáticas. El segundo tipo de evidencia, muy cercano al que proveerían las intuiciones racionales, un tipo de evidencia que no es necesariamente evaluable por medio de la conducta, skinneareana o funcionalista, usando algún *test de Turing*, es al que se acogen algunas posiciones epistemológicas y filosóficas. Veremos más sobre este asunto abajo.

2.2 Evidencia empírica, Test de Turing y Conductismo

¿Cómo sabemos si alguien puede hablar castellano? ¿Cómo sabemos si alguien puede sumar? ¿Cómo sabríamos si alguien puede hacer meta-lógica? Todas estas preguntas pueden recibir una respuesta usando lo que conocemos como *test de Turing* (TT en adelante). A esta primera forma de decidir la verdad o falsedad de las afirmaciones como las anteriores es el primer grupo de evidencia posible que revisamos en esta tesis. Aunque el TT clásico usa solamente evidencia lingüística, el test de Turing total (TTT, cfr. Harnad, 1992 y Bringsjord, Noel y Caporale, 2000) puede modificarse para incluir otro grupo de evidencia que no es lingüística. En este capítulo me referiré a evidencia TT, como evidencia Turing-testeable ya sea por el TT clásico o alguna extensión total o parcial del mismo (el TTT).

El TT consiste básicamente en una prueba para discernir empíricamente de forma sistemática y controlada entre diferentes aptitudes que un cierto agente puede tener, cuidando que la conclusión no se vea afectada por ciertos prejuicios que pueden tener quienes se encargan de juzgarlo. El TT como lo propuso Turing (1950) es una forma para determinar si una máquina puede o no ser tan inteligente como un ser humano. La prueba es interesante entre otras cosas porque es una prueba que controla ciertas variables que pueden sesgar a los seres humanos encargados de juzgar el asunto, y porque es una prueba que controla el factor de los prejuicios. Esto último equivale a decir que el TT es una prueba que no presupone que una persona tiene una aptitud mental alta o baja sólo en función de que satisface de entrada la condición de pertenencia a una especie o un subconjunto de una especie, o una forma de hablar, cierto léxico o ciertas gesticulaciones. El TT está diseñado para que dos agentes estén frente a un interrogador, quien juzgará de entre los agentes el que es humano que posee cierta

característica digamos, psicológica o intelectual, por medio de acceso cuidado a la información que le pueda proveer al interrogador. Dentro de los agentes que compiten tiene que haber un miembro de control, un humano, y una máquina en la prueba original.

Si después de un cierto tiempo de, por ejemplo, conversación, el juez o jueces humanos no disciernen a la máquina del ser humano tomando en cuenta sus habilidades para conversar por ejemplo, entonces, diría Turing, no hay diferencia entre las capacidades intelectuales de una máquina y una persona humana. Si las capacidades lingüísticas son el mejor vehículo o escaparate de la inteligencia o si son siquiera un ámbito en el cual encontrar inteligencia o no, sería discutir un punto nimio para este asunto de las capacidades intelectuales de las máquinas. El tema interesante es que el TT se ha propuesto generalizado (ver Bringsjord, Noel & Caporale, 2000) para poder encontrar o adaptar un mejor test de Turing para buscar y encontrar inteligencia. Si se mueve como pato, hace como pato, **y podemos agregar,** si aletea como pato, tiene plumas patiformes, se enferma como pato, se reproduce como pato, tiene información genética de pato, …, es pato.

Supongamos que la inteligencia se mide en la forma en que un humano puede pintar con óleo ciertos temas, o representar ciertos contenidos conceptuales o lo que sea de tal modo que nuestra teoría de inteligencia nos dijera que *en* las pinturas se ve quién posee inteligencia y quién no, sustituyendo a las habilidades lingüísticas. El esquema general del TT podría **juzgar** teorías de la inteligencia incluso así o más **excéntricas**. Debemos pues, poder construir una máquina para que compitiendo al pintar un óleo, ella tuviera una oportunidad de *llenar el ojo* del juez o jueces que deciden quién tiene y quién no tiene inteligencia. Igual si la prueba se libra jugando ajedrez, descubriendo teoremas de matemáticas o encontrando rutas óptimas para la entrega de correspondencia; lo que sea, debe haber igualdad de oportunidades para la máquina.

El tema de quiénes pueden ser los interrogadores en un TT, en este caso, quién puede juzgar razonablemente si hay Gödel susceptibilidad, habilidades aritméticas, lógicas y demás, tampoco es un problema irremontable para el esquema general del TT. Uno puede poner como interrogador a un experto en psicología contemporánea, a un filósofo experto en racionalidad, a un lógico, a un matemático, o a un jurado popular. Incluso, el TT podría incluir a un comité de expertos y otras personalidades. Podemos incluso imaginar un juez ideal e infalible. Este no es problema para el esquema general de los TTs.

El punto interesante aquí es el de si hay o no uno o más conjuntos de evidencia empírica que basten para sustentar la tesis de que algún agente, natural o artificial, o una mezcla, es de hecho inteligente, o creativo, o empático, o capaz de aritmética, lógica y meta-lógica, o lo que se nos ocurra. La evidencia empírica que cuenta para el esquema general del TT es aquella que es susceptible de *medición científica ortodoxa*, i.e., que es, típicamente, de acceso público y que se reconstruye adecuadamente utilizando oraciones formuladas en la tercera persona en la medida en que el resultado y las consideraciones no dependan de la voluntad o capricho de un individuo o un grupo de individuos que pueden tener intereses particulares en tergiversar o preservar la teoría. Kant llamó a esta disposición de ánimo investigador crítica (*Kritik*). El contraste entre evidencia subjetiva y evidencia empírica de tercera persona, se verá más claro después de las secciones siguientes en las que hago una caracterización de la noción de evidencia subjetiva (o datos subjetivos).

Debe quedar claro que, la variación del interrogador particular no es el punto importante, sino la posibilidad de que cualquiera que tenga la disposición de revisar la evidencia puede decidir en un TT si la máquina es o no poseedora de cierta capacidad mental o intelectual, general (domina una teoría) o concreta (resuelve un ejercicio, prueba un teorema, responde una pregunta específica, etc.). Podemos tener personas que sean expertas en dos temas distintos, y podríamos pedirle a esta persona que con la evidencia empírica que se le pase, lingüística o no, juzgue la posesión de dos aptitudes. No porque variemos a la persona, estamos variando el tema de evaluación, ya que cualquiera en las condiciones de *expertiese* adecuadas, debe poder ver la posesión o ausencia de tal capacidad. Pensemos en que juzgamos si algún programa de computadora es hablante del castellano. Tenemos a nuestros dos agentes, A y B. Uno es humano y el otro una computadora que intentará convencernos que habla español. A preguntas como "Buenos días", los agentes debieran responder algo relevante y en castellano. Por ejemplo, responder "what?" haría a cualquiera, A o B, perder puntos en esta competencia, algo que la computadora no desea, pues desea tener más puntos de habla de castellano que el otro. El hablante humano desea tener más puntos de habla de castellano, para *confundir* al juez, así que conversa en su mejor castellano, obviamente. Habiendo más de 350 millones nativos del castellano, de ahí tenemos varios juzgadores competentes de la posesión de esa habilidad. El dictamen no depende de nadie en particular, ya que la mayoría de los hablantes del castellano deben poder darse cuenta; la evidencia aunque tuviera elementos subjetivos, es de naturaleza fácilmente inter-subjetivable. Intersubjetivo quiere decir: "si cambiamos a la persona apta para

juzgar, no cambiará el resultado" u "otros igualmente competentes llegarán a la misma conclusión." Así, el cambio de juez particular no implica cambio de tema, ni de TT.

Para los fines de esta tesis, debemos entender TT como un esquema de prueba con una gran plasticidad, es decir, el tipo variabilidad que permite conseguir el resultado adecuado. Una de estas variaciones es que podemos modificar el predicado que buscamos atribuir por ejemplo 'ser humanamente inteligente', 'tener capacidad de hacer aritmética', 'tener capacidad de hacer lógica', 'ser hablante competente del castellano', etc. Para ciertos predicados tendremos expertos muy bien calificados, para otros no, para ciertos predicados requeriremos comités de expertos, para otros cálculos muy precisos, buena parte de este trabajo se hace en metodología de la investigación, en el marco conceptual adoptado.

En el caso de los *argumentos gödelianos* es especialmente importante el tema de si hay alguna evidencia empírica suficientemente aceptable como para atribuir a uno mismo o a otros la posesión actual de capacidades aritméticas, lógicas y meta-lógicas, y la posesión de capacidades cognitivas matemáticas. De la misma forma también son relevantes para esta discusión otras características y qué tan adecuado es atribuirlas vía TT. Puesto de esta forma, queremos saber si en nuestro universo histórico 'poseedor de mente humana' es coextensional con 'máquina de Turing programada de las formas X, Y y Z' en donde 'X, Y y Z' es cualquier programa de computadora (tabla de transición, función o software incluso) que pueda ser implementado por alguna *máquina de Turing* o que pueda *ser* una *máquina de Turing* concreta. Para los AGs es especialmente importante la relación que hay entre el predicado 'poseer una mente' adecuadamente atribuido con las capacidades para aritmética, lógica, meta-lógica y el lenguaje, y si éstas son empíricamente evaluables o no.

Al respecto de la atribución correcta de aptitud Aritmética y lógica por medio de evidencia empírica juzgable por medio de TTs sobre los usos del lenguaje, entre otros Dummett (1978, pp. 186-201) toma precisamente a la incompleción de Gödel como una razón a favor del platonismo en matemáticas (aunque hace una lectura cauta[24]). Si el platonismo en Filosofía de

[24] Por ejemplo dice Dummett (1978), con mi traducción, en p. 187:

> "Aquellos que aceptan esta perspectiva de la materia en cuestión [a saber que el teorema de Gödel afecta a todo sistema formal que contenga PA] de forma pronta sacan la conclusión de que la expresión 'número natural' es un contra-ejemplo a la tesis de que el *significado* de una expresión se explica en términos de su *uso*."

O bien dice, p. 189:

las Matemáticas es cierto, entonces parece que el uso de las palabras no es relevante para entender su significado más allá que entender cuál es el referente de las mismas (*Bedeutung* o *denotation*), o además del referente, el sentido (*Sinn* o *connotation*) de las palabras. Esta clase de relaciones de implicación que, si son filosóficas o no tan filosóficas, serán tenidas en suspenso en esta tesis. Si hay algún tipo de preeminencia metodológica y necesitamos entender mejor el fenómeno del significado antes de poder entender el fenómeno de la verdad y de prueba matemática, no queda del todo claro. Muchos proyectos de investigación filosófica han partido, no sólo de tesis completas (no de investigar palabras y sus significados,por separado) sino de juicios de identidad como leyes lógicas, afirmaciones analíticas, a priori y necesarias.

Así, un tipo general de atribución de posesión actual de capacidades aritméticas, lógicas o meta-lógicas, y lingüísticas la hacemos vía *algún* TT si es que todos estos predicados son o bien puramente empíricos y finitistas (identidad mente-cerebro-cierta máquina de Turing) o bien accesibles empírica y finitariamente con un grado bueno de confianza, aunque ellos mismos no sean MTs o *Turing computables*. Si la atribución de una característica cualquiera se puede hacer idóneamente vía TT, diremos que tal característica *es atribuible vía TT*.

2.3 Evidencia subjetiva que no puede ser Turing testeada

Es importante notar que el TT fue diseñado para aportar evidencia sobre la presencia de estados internos, subjetivos o mentales, no específicamente sobre la veracidad o corrección de aquello de lo que el estado mental en cuestión versa. Dados intereses metodológicos sobre ambos elementos, tanto los estados mentales como el *aboutness* de los estados mentales, podemos postular y evaluar tanto la existencia meramente mental o subjetiva de un objeto, por ejemplo, unicornios, personajes de ficción, alucinaciones, etc., o bien las características e interrelaciones de distintos estados o capacidades mentales, por ejemplo, cómo distinguimos el miedo de la ansiedad, en uno mismo o en otros, etc.

"… [E]n el contexto actual, [el de decidir si el uso se reporta en términos explicativos, descriptivos o normativos,] no nos interesan estos complejos problemas; el asunto para nosotros es solamente qué tanta luz arroja el teorema de Gödel sobre el significado de 'número natural' en la medida en la que entender el significado involucra entender la aplicación del predicado 'verdad' a los enunciados aritméticos."

Esta clase de evidencias son difíciles de caracterizar. En parte, dependen de caracterizar adecuadamente a lo que sería el aparato receptor y procesador de la evidencia subjetiva. Normalmente esa función la realiza *la consciencia* o a veces nombrada como 'introspección' o 'intuición'. Intentar una caracterización unívoca de la misma, parece ser un trabajo filosóficamente polémico como para presuponer alguna cosa al respecto de ella.

En esta sección daremos algunos ejemplos de evidencia subjetiva. Dicho de otra forma, en la siguiente sección vamos a caracterizar los datos subjetivos y en esa medida oraciones que sólo pueden ser verdaderos en caso de la tenencia de estos datos. (Para no decir "existencia positiva de estos datos" y arriesgar la confusión de explicaciones no-subjetivas por motivo de que ciertas nociones o tipos de datos subjetivos sean intencionales o, dicho en otra palabra, representacionales.)

Dicho de una forma un poco más esquemática; vamos a entender por la evidencia que no es *Turing-testeable*, aquella que, además de ser decisiva para un sujeto racional y epistémicamente apto, es infinita en cantidad por su naturaleza o bien no es empíricamente medible ni directamente ni indirectamente. Podemos llamar a la evidencia empírica directa como *sense data* (*datos de los sentidos*). Cuando *veo con mis ojos una manzana frente a mí*, estoy teniendo evidencia empírica (*veo con mis ojos*) directa (el verbo está conjugado en presente indicativo) de cierto hecho empíricamente asequible (*que hay una manzana frente a mí* y las manzanas son objetos visibles, además de, saboreables, olfateables, audibles y tocables). La evidencia empírica indirecta la tenemos por una experiencia empírica directa de evidencia que apunta indirectamente a cierto hecho, o bien por medio evidencia indirecta sobre un hecho empírico. Cuando *veo que hay una columna de humo*, tengo un dato empírico directo, a saber, una experiencia visual de humo, y esta evidencia empírica es evidencia empírica indirecta de que hay fuego en la fuente de la columna de humo que veo directamente. Cuando *me dice Juan que antes de los 80s México era un país próspero*, tengo evidencia indirecta de cierto hecho de la economía, aunque el testimonio de Juan lo recibo directamente por los oídos, el testimonio no es en la misma medida en la que el humo es signo de fuego, signo de la prosperidad de México antes de los años 80. Esta es evidencia empírica indirecta, que experimentó alguien más y no yo mismo suponiendo que mi reporte subjetivo es completamente correcto para empezar.

Así, a la evidencia que no es TTeable la tomo como evidencia subjetiva *simpliciter*. Hay un caso que considerar aquí. Por ejemplo, hay evidencia subjetiva directa que puede también no ser evidencia empírica. El fenómeno de la alucinación es un caso claro de esto; subjetivamente una persona tiene lo que cree que es una experiencia directa de un monstruo. No hay tal monstruo, su evidencia es sólo subjetiva; ¿cómo se puede Turing Testear la existencia de tal monstruo usando sólo al sujeto que afirma que lo percibió? Simplemente no se puede. A lo más a lo que podemos aspirar testear es la honestidad o genuinidad del fenómeno de la alucinación. Pero es un hecho distinto el de que *hay un monstruo debajo de la cama* a *Fulano alucina un monstruo debajo de la cama*. El primero puede ser falso sin que lo sea el segundo hecho. ¿Es TT el primero? Al parecer sí lo es, pero de forma indirecta. Sometemos al sujeto alucinador a cuestionarios, vemos si en su torrente sanguíneo hay signos de estrés correlativos al grado de miedo que le inspira una criatura monstruosa en la intimidad de su hogar, probamos la galvanización de su piel, etc. Toda esta es evidencia empírica indirecta, y sólo en esta medida es confiable.

Podemos defender la posibilidad de casos en los que la evidencia subjetiva no sea TT de ninguna forma suficientemente confiable, y que no obstante, el hecho *intuido* o *introspectado* sea genuino. El antecedente del *cogito cartesiano* es un candidato para esto; *pienso*. Podemos encontrar correlaciones cerebrales entre cuando las personas dicen estar pensando, y la activación de ciertas zonas. Pero si alguien quisiera evidencia más confiable, es posible que no exista una manera de inferir con grado alto de confianza de que *se mueven tales y cuales flujos electrónicos en el cerebro, o manos y papel, dejando tales garabatos, o boca, laringe, aire y se oyen tales sonidos*, por lo tanto *hay un pensamiento genuino allí ocurriendo*. Maudlin (1995) y Kirk (1986) reconocen abiertamente esta posibilidad. Por ejemplo, Kirk (*idem*, pp. 449-450) dice:

> Supongamos que Alf [el excelente lógico] escribe '2 + 2 = 4' en un pedazo de papel. Los movimientos que hace al escribir, las manchas resultantes, y cualquier cosa que involucre su entendimiento ortodoxo de la marca que escribió, son, claro, modelables mediante el sistema formal en cuestión (el que supuestamente es instanciado por el universo completo, o por algún pedazo del mismo tal que incluya a Alf). Ciertamente es natural tomar el patrón '2 + 2 = 4' y asumir que *ese* debe ser un 'teorema' del sistema (estrictamente, el análogo físico de un teorema). Pero consideremos, si estamos pensando en cierta región del mundo como la instanciación de un sistema formal tal que la región del mundo implementa justamente el sistema formal, tal como lo requiere la posición mecanicista, entonces el sistema debe hacerse responsable de alguna manera de todos los estados y eventos en esa región del

> mundo. El sistema no hará bien su trabajo si apenas formaliza con algunos pedazos de la región del mundo que formaliza [lo que en esta tesis llamamos representación óptima] … Porque el patrón ['2 + 2 = 4'] será un archipiélago de tinta negra en una pieza de papel blanco – cuando ambas cosas, las marcas y el papel, son partes iguales del mismo estado del mundo en su totalidad. … Así, ya que esta clase de esquemas está a disposición del mecanicista Laplaceano, parece que no pueden ser forzados a conceder que las oraciones matemáticas sean teoremas (¡sic!) de un sistema formal que representa adecuadamente ya sea a Alf, a Alf además de su entorno inmediato, o a Alf junto con el resto del universo.[25]

Así, la evidencia completamente capturada por un sistema formal que representa bien al mundo o a porciones relevantes (Alf y sus alrededores, p. ej.), es completa solamente si no se escapa nada relevante para considerar que hay no sólo manchas sobre papel, sino que se entienden como un teorema. ¿Qué clase de evidencia podría escapar? Quizá tenemos una descripción exhaustiva de todos los eventos tal como accedería a ellos un físico, pero aún así nos faltaría aquella evidencia que nos convence de que las manchas en el papel son un teorema. Mi hipótesis aquí es que si hay un buen candidato para jugar ese rol, ese candidato son los *datos subjetivos*, propios, de otros, del resto, de una comunidad actual o potencial, o la combinación de subjetividades que quieran.

Ahora, ¿cómo podría ser el caso de que tengamos *datos subjetivos* que nos permitieran saber sobre el infinito? ¿Hay datos subjetivos sobre el infinito? Podemos conceder con que sí los hay. Una noción cualitativa, opuesta a la cuantitativa, es un elemento común de la poesía y la literatura.[26]

En *Auguries of Innocence*, William Blake propone una suerte de caracterización de una perspectiva:

---

[25] Mi traducción de: "Suppose Alf [the excelent logician] writes '2 + 2 = 4' on a piece of paper. The movements he makes in doing this, the resulting marks, and whatever is involved in his understanding it in the usual way, are of course all assumed to be somehow modeled by the formal system in question (the one supposedly instantiated by the universe as a whole, or by some chunk of it which includes Alf). Of course it is natural to seize on this nice pattern '2 + 2 = 4' and assume *it* must be a 'theorem' of the system (strictly, the physical analogue of a theorem). But consider, if we are thinking of a certain region of the world as the instantiation of a formal system which does justice to it as mechanism requires, then the system must somehow embrace the totality of states and events in that region of the world. The system will not do its job if it deals only with scattered bits and pieces of it. … For the pattern ['2 + 2 = 4'] will be an archipelago of black ink on a piece of white paper – when both marks and paper are equally parts of the same total state of the world. … So, since that sort of scheme is available to Laplacean mechanists, it seems they cannot be forced to concede that mathematical sentences will be theorems (sic!) of a formal system which adequately represents either Alf, or Alf plus his immediate surroundings, or Alf plus the rest of the universe."

[26] Por ejemplo, Dummett (1978, p. 190) lo dice abiertamente: podría ser que algo no sea analizable, y en esa medida, que sólo podamos reconocerlo cada quien, en sí mismo y que reconocerlo en otros siempre depende de "fe ciega". Nin siquiera es algo susceptible de cuantificación, sino una especie de estado mental subjetivo, que podría o no ser correcto.

To see the world in a grain of sand,
And a heaven in a wild flower;
Hold infinity in the palm of your hand,
And eternity in a hour.

No hay en este verso dato alguno que nos permita saber si hay alguna diferencia entre 'infinito' y 'eternidad', o si hay distintos tamaños de infinitos, o cómo se relacionan los infinitos entre sí, o cómo funciona exactamente la representación de un infinito con respecto a lo que representa. Quizá en la comprensión de ciertos poemas encontremos el candidato más aceptablemente popular de *evidencia subjetiva de algo infinito*. Cabe mencionar que nada hay que impida que, el aparato cognitivo que se usara para comprender este verso y quizá su belleza y fortaleza, sea usado por las personas al entender afirmaciones matemáticas, evaluar su veracidad o falsedad, y sus demostraciones. La poesía y las ciencias formales, dentro de las que están la Aritmética, comparten algunas características.[27]

2.3.1 Caracterización teórica de la fenomenología en relación al platonismo matemático

El segundo grupo de evidencia, el que llamamos *datos subjetivos*, es bastante más difícil de caracterizar. Un juicio con contenido proposicional cuya evidencia idónea es de primera persona puede tener la forma siguiente:

i. [Yo] [soy consciente de que] *p*.

Para los fines de este trabajo *i*) no es una forma que deba ser aceptada como un análisis exhaustivo y correcto de la forma de los juicios cuya verdad depende de evidencia subjetiva ("juicios subjetivos" en adelante). Quizá ni siquiera sea un análisis, que no sea conceptual, es decir que no es susceptible de ser un análisis funcionalista o *Turing mecánico*. Por ejemplo, en el caso del método cartesiano, en particular el *cogito*, la justificación es de primera persona. Un filósofo inspirado por el cartesianismo, quizá usaría la forma siguiente para caracterizar los juicios cuya evidencia es de primera persona:

[27] Timothy Williamson dijo en una entrevista al New York Times que la lógica y la poesía se parecen mucho en su apego a la expresión precisa y elegante, y a la forma en la que podemos aprender de una forma poética o lógica de decir cosas. (XXXX)

ii. [Yo] [pienso que] *p*.

En este caso pensar no sólo es la habilidad de manipular mentalmente expresiones lingüísticas o realización mental de cálculos, sino incluso también el acceso a estados típicamente subjetivos como el de creer que se siente un dolor o que se ve un color rojo. En este caso, *ii* variando *p* por *pienso* es lo mismo que

ii'. Pienso.

De modo que todo análisis susceptble de TT trataría la verdad de esta afirmación meramente como tautología, o como una contingencia verdadera. El punto aquí es que es una forma de acceder a una verdad de forma subjetiva.

Habría otras formas de sistematización de lo que podríamos llamar una metodología subjetivista. Creo que éste es el caso de todos los tipos de *fenomenología* de la filosofía germana. En especial, la fenomenología husserliana suscribiría, dejando de lado, por ahora, aspectos importantes, que la forma de los juicios subjetivos es más cercana a

iii. [Me] [parece que] *p*.

Para ciertos naturalistas fenomenistas, como Quine, quizá la mejor forma de capturar el contenido de un juicio subjetivo sea la siguiente:

iv. [Me] [siento fuertemente inclinado/tengo una fuerte inclinación a asentir a] *p*.

Para este caso puede uno discutir si la evidencia de primera persona es, en sentido metafísico, evidencia necesariamente de primera persona o si tenemos alguna forma, aunque sea en principio, de acceso a ella por métodos indirectos de tercera persona y así, tenemos alguna traducción fiel de juicios subjetivos a juicios complejos de tercera persona. Esta es una discusión filosófica y científica. El tema de cuál es la manera más correcta para capturar la forma de los juicios evaluables (absolutamente necesaria o contingentemente necesaria) por medio de evidencia subjetiva, no debiera tampoco desviarnos demasiado de la discusión acerca

del establecimiento de la Gödel-susceptibilidad, la Gödel no limitación, y la veracidad de la oración *G*.

¿Cómo pueden los datos subjetivos ser la mejor justificación de la atribución de Gödel-susceptibilidad de un agente que sea no exclusivamente uno mismo? ¿Qué tipo de evidencia de primera persona prueba que un agente tiene o no ciertas habilidades lingüísticas y/o aritméticas? Esta pregunta debe ser dividida en dos. Primero, uno sabe la mayor parte del tiempo que es lingüísticamente capaz o aritméticamente capaz, o incapaz en su caso, sin necesidad de recurrir a la opinión de un tercero o de un medio empírico de prueba o sin necesidad de estar haciendo operaciones aritméticas. Pero también podemos imaginar contextos en los que la evidencia de primera persona baste para determinar que una persona distinta a uno mismo, tiene las características de ser capaz de hablar y de realizar ciertas operaciones aritméticas. Los datos subjetivos pueden tenerse de manera intersubjetiva y de forma indirecta, empíricamente. Es incluso lógicamente posible que por cierta comunicación no física, o física pero nada tradicional, como la telepatía o alguna conexión neurológica directa entre individuos, tengamos acceso a otras subjetividades *qua* subjetividades.

Así, a *i)*, *ii*), *iii)* y *iv)* les podemos variar lo que está entre corchetes, por todo el horizonte de pronombres personales, y quizá otros como el dual (‘tú y yo’). Parece más natural inclinar la balanza de los datos subjetivos no sólo al referente del ‘yo’, sino si hemos de extenderlo, al referente del pronombre ‘nosotros’. Después de todo, los juicios de tercera persona parecen fácilmente perder la referencia a la subjetividad, pero esta es una ilusión quizá motivada por la idea común en filosofía de la ciencia de que la forma de escribir y comunicar resultados científicos es evitando toda subjetividad; usar oraciones que no estén en la primera persona de ninguno de los números, sino sólo en tercera persona de ambos números; él/ella/eso y ellos/ellas.

También podemos variar la expresión en la segunda ocurrencia de los corchetes a cualquiera de las cuatro formas sugeridas en *i)*, *ii*), *iii)* y *iv)*. Muchas de ellas parecen incluir características típicas del conocimiento de primera persona, como son la incorregibilidad, la disminución del compromiso epistémico o de la aceptación generalizada.

Llamo subjetivos a los juicios cuya forma es alguna de las anteriores, es decir, de *i* a *iv*. No llamaré subjetivos a aquellos juicios que son verdaderos en simplemente virtud de que alguien

los crea. Un ejemplo de esta clase de juicios subjetivos que se hacen verdad por el sólo *acto psicológico* de que sean creídos, es: 'Me parece que algo me parece'. No es que esta clase de juicios no puedan ser verdaderos **sólo o idóneamente** por evidencia subjetiva, sino que además parecen tautológicamente ciertos. Así dejo abierta la puerta a que tengamos juicios subjetivos que puedan estar subjetivamente mal justificados y subjetivamente corregidos. Si aceptamos que *tener un dolor* es un buen candidato a un juicio subjetivo que sólo es verdad por medio de evidencia subjetiva directa, aún así podemos imaginar una situación en la que una persona nace con un defecto neurológico que le causa un dolor. El dolor podría estar de manera que quien tiene el dolor en cierto grado, lo pierde de vista por acostumbrarse a él, y que cuando se le pregunta a esta persona si le duele algo, ella responderá que no. Una vez identificado y sanado el nervio defectuoso, pues la disminución de una molestia hace patente a esta persona que tenía un dolor pero *creía que no le dolía*. Lo mismo puede pasar con la auto-atribución de felicidad, de estrés, de pobreza, de debilidad o fortaleza, etc.

Los juicios subjetivos pueden ser verdaderos en principio o bien contingentemente. Una subdivisión más; a los juicios cuya justificación idónea es la evidencia empírica y que esta situación es explícita en algún lugar de la cadena justificativa relevante y mínimamente necesaria, los llamaré juicios empíricos. No importa que un mismo juicio sea bien soportado en evidencia subjetiva y TTeable. La posibilidad de que ambas evidencias sean coherentes e incoherentes entre sí, y cuál tenga más peso, es un compromiso que no tomaré aquí por lo menos de manera presupuesta. Hasta aquí sólo introduzco la terminología y la forma en la que la entenderé. Si *p* es un juicio empírico entonces es un juicio verdadero o falso por evidencia empírica, y si *p* es un juicio subjetivo, entonces *p* es verdadero por evidencia de primera persona. Si coincide la evidencia empírica y la evidencia subjetiva, cuánto, cómo y por qué lo hagan, no es problema que necesite muchos más detalles por el momento. Basta con entender la distinción.

Otra distinción conceptual que resulta útil es; *a priori* y *a posteriori*. Veamos un ejemplo de un juicio subjetivo falseado *a posteriori* como un ejemplo ilustrativo. Imaginemos que una persona despierta de una operación en la que uno de sus miembros, el brazo derecho, le fue amputado. La persona despierta y antes de percatarse de que fue amputado el brazo identifica lo que él diría que es un dolor de brazo derecho. La forma del juicio que esta persona daría es: 'Siento-pienso-creo-asentiría-a que tengo un dolor en el brazo derecho'. Al notar por inspección ocular que no podría tener un dolor en el brazo derecho, porque el brazo no está allí,

entonces él bien puede tener las dos creencias, por un lado, que siente un dolor en el miembro amputado, y por otro, que no puede tener un dolor en el miembro amputado porque ese miembro no está allí y la localización del dolor indica un problema cerebral. Un juicio subjetivo verificado a posteriori es un caso más sencillo de acuñar y quizá más común. La gente solemos atender a nuestras intuiciones, asegurando, por ejemplo, *Fulano* da mala espina. De una forma un poco más clara dirían algo así como *tengo la sensación de que Fulano es malo o traerá cosas malas*. Días después, Fulano golpea a alguien y entonces la sensación se confirma sobre bases independientes a las subjetivas y propias. ¿Esta es una confirmación a priorística de la fuente de nuestros datos subjetivos? Parece que claramente no es así.

Una serie más de divisiones sobre las que cabe tener alguna idea son distinciones cuasi-clásicas entre verdades necesarias y contingentes, analíticas y sintéticas. Son verdades analíticas aquellas verdades en virtud del significado de los términos que quedaron explícitamente a salvo del ataque quineano en su texto *Dos Dogmas del Empirismo* (Quine, 1951), es decir, las oraciones analíticas cuya verdad no depende de ninguna forma en el fenómeno de la sinonimia. En ese mismo artículo Quine atrajo la atención a la posibilidad de que el fenómeno de sinonimia es afectado por situaciones históricas y sociales (contingentes) de los usos del lenguaje, de tal manera que, por ejemplo, es contingente que *la edad del Presidente de México en 2013 es 47*. Sabemos que variando términos co-referenciales en una oración, salvamos la verdad de la oración. Así en la oración anterior tenemos que *47 es 47* sin depender de si ocupa la presidencia alguien de menos o más años que 47. Las verdades *a priori* las tomo como verdades cuya justificación es independiente de la experiencia empírica, y a posteriori son aquellas verdades cuya justificación idónea es crucialmente de evidencia empírica en al menos algún eslabón de la cadena de justificación. Las verdades necesarias no las agota el conjunto de verdades analíticas, ni al de verdades *apriorísticas*. Sólo para hacer explícito que creo que hay una parcela de traslape, por lo demás muy considerada ya en la literatura filosófica contemporánea, entre verdades necesarias y verdades a posteriori. Así, tenemos que los juicios subjetivos no son necesariamente *apriorísticos* especialmente pero no exclusivamente en el caso de que tengamos un buen análisis empírico de los términos que constituyen la forma o las formas de los juicios subjetivos. Este no es un problema cerrado. El análisis empírico de los elementos constitutivos (lo que está entre corchetes en los ejemplos de *i*) a *iv)*, arriba) de los juicios subjetivos puede ser adecuado sólo en principio sin perjuicio de la utilidad contingente que tienen. Lo anterior no obsta para que considere el valor conceptual de la distinción entre juicios subjetivos y juicios verdaderos *a priori*.

Estas distinciones tienen una utilidad obvia para reconstruir el debate acerca de si los juicios de la aritmética por ejemplo, son, en caso de serlo, verdaderos necesariaaprioríistica, analíticamente y, en especial, si lo son subjetivamente o no. Lo anterior incluso en los casos en los que uno de los términos de la ecuación es contingente ("la edad del Presidente de México", por poner un ejemplo). De esta forma, para Gödel, los números naturales son entidades existentes a las que tenemos acceso por medio de alguna facultad subjetiva como la intuición (Wang, 1996 y Tieszen, 1984, 1998, 2006), así, para Gödel también, los juicios verdaderos de la Aritmética son verdades de acceso subjetivo, *a priori*, no necesariamente inconsistentes con la evidencia empírica y son necesariamente verdaderos. Para este efecto, pueden verse los trabajos ya citados de Wang y Tieszen, además de fuentes enciclopédicas confiables.

Veamos entonces, que un juicio aritmético particular o conjuntos completos de ellos, pueden recibir su justificación de maneras complejas, atendiendo a las diferentes categorías que he caracterizado ya antes. La justificación típicamente aceptada para juicios aritméticos es su teoremidad, es decir, que aparezca en algún eslabón de una cadena inferencial formal algorítmica regida únicamente por reglas de la lógica deductiva, reglas gramaticales y definiciones de términos lógicos y no lógicos. Este tipo de justificación pretende ser analítica, necesaria y apriorística, y tampoco es incoherente, no al menos obviamente, con otros tipos de justificaciones fenomenológicas. Las pruebas formales de que "2+2 = 4" tienen la estructura de una inferencia sobre una serie de establecimiento de identidades, por ejemplo que "2 = 1+1, así (1+1) + (1+1) = 1+1+1+1, y 1+1+1+1 = 4, por lo que 2+2=4 es 1+1+1+1=1+1+1+1, o 2+2=2+2, o también, 4=4." No es la única forma de dar alguna evidencia (o prueba) a favor de que 2+2=4, ya que también podríamos apelar a una cierta intuición, no necesariamente formal, y quizá ni siquiera necesariamente lingüística. Este tipo de imágenes o impresiones en la intuición las podemos llamar consistentemente con la literatura filosófica en fenomenología, evidencia apodíctica, un tipo de evidencia subjetiva del que echa mano sobre todo el sistema de la fenomenología husserliana. (véase Tieszen, 1984.)

La evidencia subjetiva tiene el problema de que para operaciones aritméticas con muchos numerales y/o de muchas cifras, puede dejar de ser útil, así, los cálculos formales son más estables para este tipo de operaciones complejas, barrocas o largas. No queda del todo excluida la posibilidad de que alguien o de acuerdo a alguna técnica, uno pueda tener evidencia

subjetiva suficiente a favor de operaciones aritméticas largas y complejas, tal como la tenemos a favor de las más simples, pero ese es un caso a elaborar mejor.

Así, para conectar los temas que tenemos aquí, podemos determinar que un agente es Gödel-susceptible apelando a su conducta externa empíricamente mesurable, pero también podemos hacerlo apelando a evidencia subjetiva que puede o no ser consistente con la evidencia empírica sobre la conducta externa. Es trabajo de la epistemología de las matemáticas encontrar qué relaciones hay entre estas dos formas de determinar si un agente, sea uno mismo o no, tiene competencia matemática genuina y también determinar si un método justifica o no de forma adecuada nuestro conocimiento matemático independientemente de la posesión de competencias matemáticas. Aquí yo sólo quisiera rescatar la posibilidad de sostener con alguna razón que la evidencia conductual empíricamente mesurable en la cantidad que sea no baste o no sea adecuada para determinar Gödel susceptibilidad, y que tengamos la necesidad de una teoría alternativa basada no en evidencia conductual empíricamente obtenida, sino, por ejemplo en cierta intuición o fenomenología, si es que buscamos preservar la superioridad del conocimiento matemático por sobre el conocimiento empírico puestos en la balanza de la certidumbre.

La fenomenología es una metodología filosófica sistematizada por Husserl que Gödel sugirió como la mejor epistemología matemática, en especial a la luz de que el formalismo, sea logicista o no, parece ser insuficiente para realizar ese trabajo.

Es posible que haya otro tipo de metodologías subjetivistas a la caracterizada aquí, o incluso fuentes o colecciones particulares de datos subjetivos, distintas a la caracterizada arriba. Esta metodología subjetivista puede servir a distintos propósitos. El primero, es el de determinar la veracidad o falsedad de tesis particulares. El segundo propósito es el de ayudar a comparar entre la fortaleza (modal) de verdades de distinto tipo.

Una prueba fenomenológica probablemente dependa de algo que no puede presentarse en texto; ¿cómo recuperar una visión particular en un texto? ¿Cómo fotografiar un pensamiento tal como aparece a quien lo posee, en especial si es un pensamiento sin contenido puramente empírico como es la veracidad de un teorema aritmético? La evidencia fenomenológica es siempre indirecta, a menos que sea la que tenemos nosotros mismos. También puede circunscribirse a oraciones atómicas no condicionadas (*"soy consciente de que hay dos manos*

frente a mí", *"estoy fuertemente dispuesto a asentir a que hay una computadora frente a mis ojos* ahora mismo*"*, *"me parece que estoy pensando"*, etc.). Los teoremas son afirmaciones cuya verdad depende de premisas, a menos que sean axiomas. Muchas de las afirmaciones que capturan datos subjetivos parecen cumplir el rol de presupuestos o axiomas, de forma incondicionada, categórica incluso, y dentro de ellas, las que son verdad en virtud de los datos subjetivos que las respaldan, lo pueden ser de forma incondicionada.

2.4 Dos formas de determinar *susceptibilidad Gödel*

Tenemos tres supuestos agentes detrás de una pared y sólo a través de una rendija recibimos notas en donde están las respuestas a interacciones verbales que nosotros les podemos hacer y el número del agente que la escribió. Supongamos que queremos saber si los agentes detrás de las paredes saben sumar. Preguntamos cosas como *¿cuánto es dos más tres?, ¿qué número es el resultado de la suma de cuarenta y cinco y quince?*, etc. Si en el primer caso recibimos notas con un cinco en ellas, con un sesenta en las otras notas, y así por alguna cantidad finita de casos, lo más probable es que diríamos que los agentes detrás saben cómo sumar. Atravesamos la pared, y nos encontramos con que había un niño de cinco años, una calculadora con un dispositivo especial para entendernos oralmente y también tenemos a un egipcio de la corte de Akhenatón con un ábaco y un diccionario del castellano. El egipcio y el niño son ambos, competentes sumadores en los términos de este TT.

Uno de los problemas con este criterio es el que Kripke (1982, pp. 35-37) ha esbozado al respecto en una célebre nota al pie contra el disposicionalismo. Este asunto no ha pasado inadvertido para otros filósofos ya citados aquí: Dummett (1978), Kirk (1986) y Maudlin (1995) para dar sólo unos nombres. El tema es que la evidencia empírica (en este caso las preguntas y sus respectivas respuestas correctas según la regla convencional para suma) nunca es suficiente para saber si los tres agentes están usando de hecho la regla (convencional) para sumar y no una regla diferente que hiciera que para la cantidad de casos que les suministramos como examen, se comporten convencionalmente, pero para otros casos nuevos, estos agentes evidencien el uso de una regla diferente de la que llamaríamos suma. Dicho en otras palabras, es un problema serio el de qué tanto podemos confiar en la inferencia que nos lleva de ciertas instancias de comportamiento lingüístico empíricamente evaluado, a la generalización de que tal o cual agente posee el *concepto* o *significado* adecuado o si sólo está contestando por suerte

o memoria, o así. En general este es un problema que tienen los filósofos que suscriben una posición naturalista incluso en filosofía de las matemáticas. No importa si la posición naturalista es disposicionalista o no, basta con que sea Turing-computabilista, tanto metodológicamente como metafísicamente. Nóten la semejanza del tipo de naturalismo y el Turing-computabilismo, con el tipo de tesis computabilista de los AGs; hay un computabilismo que afirma que la naturaleza la podemos describir (para predecir o explicar) por medio de *computadoras abstractas* (teorías formales Turing-computables) o bien que la naturaleza es una Turing-computadora particular, que por tanto, puede al menos en principio ser descrita con una teoría Turing-computable. Si el significado es un fenómeno natural y mental más, sea individual o grupal, histórico o no, psicológico o meramente neuronal, etc., se debe responder, a mi juicio, no por medio de dogmas de investigación que nos arriesgan a sesgos de evidencia y otros tipos de defectos intelectuales, sino que debemos poder ponerlos a prueba y, en caso de que haya suficientemente buenas razones, pues perder tal presupuesto en la investigación. De entrada ubicar el criterio de la evidencia Turing-testeable como la única forma de determinar no sólo si algo es verdadero o no, sino si existe o no, nos forzaría a aceptar solamente aquello que es coherente con tal presupuesto, aunque sea de hecho el único criterio que preserve veracidad, el objetivo esencial de la investigación científica. Un lujo que no se puede dar la investigación filosófica, es justamente ese.

Una alternativa a este aparente laberinto es aceptar la posibilidad de que tenemos acceso (que yo llamo subjetivo, aunque otros puedan llamar al menos en parte 'interno', 'inmediato', 'de primera persona') que ha sido defendida explícitamente por gente de alto peso intelectual como Gödel y más recientemente por ejemplo, por Richard Tieszen (especialmente 1984, 1998, o bien en 2005 reaparecen algunos de los textos, y finalmente en 2011). Querríamos decir que podemos tener evidencia interna (intuición racional) que garantiza que estamos implementando la función suma como tradicionalmente la conocemos, pero ello no tiene porqué garantizar solamente que nosotros mismos examinándonos internamente entendemos y podemos usar la regla convencional de la suma, sin ayuda previa o sin buscar poder usar esa función matemática. Este tipo de evidencia a favor de la verdad de ecuaciones aritméticas de primer o de segundo orden, puede ser suficiente o consistente con la tesis de que hay otros agentes con acceso subjetivo a evidencia que apoyara ecuaciones aritméticas. Para decirlo de una manera un tanto cuanto fácil de confundir si es tomada fuera de contexto: para platonistas fenomenológicos hay evidencia subjetiva (datos subjetivos) que son, finalmente, tan o más

objetivos que muchos datos empíricos, y que no son privados en el sentido de que no pueda "replicarse" si se siguen ciertos pasos o se dirige bien a la consciencia.

El tema de si tenemos evidencia decisiva, mixta, subjetiva, objetiva o intersubjetiva para atribuir estados subjetivos a otros agentes diferentes de nosotros mismos es complicado. Nos interesa porque una teoría de la mente debe poder explicar el fenómeno no sólo de una mente, en este caso la propia, sino de todas las mentes. Así parecería que la estrategia subjetivista para atribuir la posesión de conocimiento aritmético nos restringiría en tal cantidad a nuestros propios estados mentales que sería imposible que hiciéramos pasar como un buen argumento gödeliano uno basado en la atribución de Gödel-susceptibilidad apelando solamente a la evidencia subjetiva de un sujeto: con mi intuición veo que soy Gödel susceptible, que además entiendo a la incompleción de Gödel, que conozco mi propia constitución, mi propia oración Gödel y veo que ella es verdad, y por todo esto, infiero (puede ser que lo intuya o introspecte también) que no estoy Gödel limitado.

Sabemos pues que una teoría de las capacidades mentales en general, y las cognoscitivas en particular, debería poder explicar también el caso del conocimiento aritmético, de verdades particulares quizá infinitas en número, verdades en soledad o conjuntos totales de ellas, y si obtener, producir, comprender o ser consciente de este conocimiento no es reducible a cálculos o computaciones, pues debería poder explicarse esto también. Así, sea por la vía de la evidencia subjetiva o de la evidencia TTeable, la atribución de Gödel susceptibilidad es uno de los presupuestos de los argumentos gödelianos que han sido atacados.

Tenemos frente a nosotros mismos una nota con unos garabatos que se parecen bastante a "2+2=4". Sabemos que el garabato '2' refiere al número dos, que el garabato '4' refiere al número cuatro, que el garabato '+' refiere a la función suma y finalmente que las rayas paralelas '=' refieren a la igualdad. Sabemos que ahí hay una oración que es susceptible de tener un valor de verdad y que el que tiene, es verdadero. 'Referir al mismo objeto' es otro nombre para la igualdad, y sabemos que '4' es otro nombre para '2+2' porque refieren al mismo objeto, práctica o uso, sea inferencialmente (análisis) o, *por mor* de la noción de datos subjetivos, subjetivamente. Quizá el conocimiento aritmético es hasta cierto punto divisible, y que un experto tenga una cantidad mayor de conocimiento que un novato, al respecto de lo que es el número dos, p.ej. Esto no es tan poco común en general y hay casos para el asunto

particular de la Aritmética.[28] Otras expresiones además de las aritméticas y lógicas formales, pueden recibir el mismo tratamiento. Saber si Clark Kent es Superman[29], es decir si ambos términos son co-referenciales, es una clase de conocimiento que tiene un lector experto de las historia de Kal-El, y uno que no es experto o que no es un experto completamente sabio, pues podría ignorar la veracidad de esta igualdad. Las disputas entre expertos, y entre expertos y novatos, muchas veces se pueden reducir a que los no-expertos desconocen ciertas igualdades, aunque tengan buen conocimiento del objeto en cuestión. Hemos tratado de caracterizar así a la evidencia subjetiva para apoyar la tesis de que podemos realizar operaciones aritméticas, que éstas son correctas y que, en ese sentido, sabemos aritmética y lógica. Así, cualquier otra clase de conocimientos podemos igualmente tenerlos apelando no a inferencias, lectura o memoria, sino a la intuición. Vemos que '2+2' es '4' y que equipararlos por medio de la igualdad aritmética es una manera de expresar ese contenido mental. Otros conceptos que *podemos ver* que se cumplen también son: '2+2' *se parece a* '4', '2+2' *es la misma cantidad que '4*', etc. Por lo menos un sujeto tiene conocimiento de aritmética y que puede inferir y definir, es decir, que sabe lógica *simpliciter*, por lo menos un sujeto es Gödel-susceptible.

No son necesarios para los fines de este trabajo más detalles de cómo funcionaría una epistemología fenomenológica de las matemáticas, sólo basta con notar el contraste entre estas formas de atribuir competencia aritmética y lógica, y competidoras netamente empiristas vía, normalmente, algún TT. Es fundamental, no obstante, notar que la diferencia mínima que debe haber entre la naturaleza ontológica o metafísica de *los datos subjetivos* (fenómenos no TTeables necesariamente) o de los datos de *los sentidos Turing testeables* es de una forma que el requisito necesario de los primeros es tal que no deben ser TTeables, por no ser Turing computacionales. Si resulta que todos los casos de *datos subjetivos* que puedan tener sentido, son Turing computables en cierto nivel ontológico o metafísico, pues entonces deberían contar como evidencia TTeable. Esta no es sólo una decisión terminológica o una estipulación ya que la evidencia empírica y finita, es mucho más fácil de caracterizar que la que no lo es, como se vió a en este capítulo. No obstante, un buen candidato a satisfacer la caracterización de evidencia no-Turing Testeable son los que llamo *datos subjetivos.*

[28] Quien quiera conocer un caso de esta clase de debates está en asunto de si la igualdad 1 = 0.9… (cero punto nueve periódico) es verdad o no.

[29] En la historia ficticia de *Superman*, el nombre que tiene este superhéroe en su vida cotidiana es "Clark Kent" pero pocos lo saben. Está la complicación extra de que el nombre que le dieron sus padres biológicos es "Kal-El".

Conclusiones del capítulo II

Hay muchas formas de establecer la verdad, la falsedad o la plausibilidad de los AGs. Una forma es mediante Tests de Turing. Las premisas del tipo general de AG del capítulo 1 tienen condiciones de verificación distintas. Algunas de estas formas de verificación son consistentes con el Test de Turing (TT) o con extensiones empiricistas y computacionales del Test de Turing; formas extendidas del TT, *total Turing Test* de Harnad u otras versiones. Llamo a este TT extendido TT*. Los TT se pueden modificar y extender de manera ilimitada y atendiendo el fenómeno que debe juzgarse. La evidencia que no es TTeable, puede ser indirecta (poco confiablemente) TTable. A los TT modificados de forma conservadora, los llamo TT*. Los TT* todavía presuponen, para ser verdad, computabilismo. Hay otro tipo de fenómenos que, al menos lógicamente posibles, pueden presentarse y no ser idóneamente (digamos, necesariamente) establecidos por medio de TT* alguno.

Un TT* es extendido si se varía el TT original ya sea por el tipo de evidencia que juzga el juez, si se varía por el tipo de juez, o el tipo de teoría que usa el juez para establecer si el agente tiene o no cierta característica. Algunas de estas extensiones, siempre que conserven cierta estructura del TT, son no sólo computacionales sino que son la mejor oportunidad que tiene el computabilismo en general de ser cierto, suponiendo que cierta teoría computacional no sea la adecuada para cierta característica de interés. Los elementos de la estructura que exploro en este capítulo son el empirismo y el funcionalismo para establecer incluso si son buenos acercamientos al establecimiento de estados mentales subjetivos.

Para que las premisas puedan ser ciertas, fue importante abarcar las consecuencias que en la noción de Gödel-susceptibilidad tienen los dos tipos de evidencia que caracterizamos antes. ¿Qué significa para un defensor del TT* decir que las mentes humanas no son objetos consistentes, de modo que no son Gödel-susceptibles? ¿O qué significa decir que las mentes humanas, dado que son consistentes no saben si hay verdades de la aritmética o de la LPO absolutamente inaccesibles o indemostrables? Algunos de los AGs que se han presentado han tenido deficiencias en establecer algunas implicaciones que tiene la Gödel-susceptibilidad, es decir, basicamente la capacidad para hacer lógica, aritmética, establecer igualdades y definiciones, razonar, o finalmente, que todo esto se haga de forma coherente en algún sentido interesante.

La evidencia o metodología que no es computacional, la llamo en este capítulo, *evidencia subjetiva*, y es *en general*, de primera persona. (Si ha de servir como guía, Dummett, 1978, (p. 190) lo establece así:

> Es verdad que algunas veces podemos ser forzados a afirmar [que algo es] inanalizable: considerando todo lo que podemos saber, puede no haber un análisis de lo gracioso. Pero esto es tolerable sólo cuando el asunto en cuestión es algo cuya presencia nosotros podemos reconocer. Sabemos lo que es reconocer algo como gracioso, y estamos de acuerdo en un grado suficiente sobre este asunto. **Un significado, que no se reduce a cómo las personas usan esa palabra, al cual yo le adjunto una palabra, de otra manera, es algo que sólo yo puedo reconocer en mí mismo: no puedo reconocerla en ti, y tampoco puedo saber cómo la reconoces tú mismo en ti.** Podría de hecho tomar por garantizado que, al decirte ciertas cosas a ti, puedo inducirte a que le adjuntes el mismo significado que yo le atribuyo a una palabra, **pero no puedo tener evidencia de que mi hipótesis es correcta; debo creer por fe ciega.** (Mi traducción y mis negritas).

Discutí en este capítulo qué tipo de formas de comprobación puede haber de esta clase de estados subjetivos, siempre que no sean computacionales en sí mismos, puedan indirectamente explorarse computacionalmente y qué efectos tendrían para la veracidad o falsedad de las premisas del AG general. Encontré que el análisis computacional de un estado no computacional en sí mismo serviría para que si se ataca la conclusividad del análisis computacional, no se atacaría necesariamente la verdad de la premisa que no es analizable computacionalmente. (Estos ataques darían pie a una falacia de *irrelevancia*; hombre de paja).

A esta más bien misteriosa y vaga caracterización de evidencia no TT, o no TT*, la llamo evidencia subjetiva. Se discute hasta qué grado se necesita evaluar siempre desde el punto de vista personal (primera persona), o si puede ser juzgada por otros individuos (tercera persona o segunda persona), en especial la Gödel-susceptibilidad. Una objeción común es que no tenemos aptitudes coherentes para la lógica o la aritmética. Esto implicaría que, por ejemplo, como especie humana podemos equivocarnos en establecer que, en general, 2+2 es 4, lo cual parece altamente implausible. No es implausible que nos equivoquemos con las cuentas de vez en vez, pero que siempre nos equivoquemos sí parece muy implausible. También parece una objeción sensata a los AGs en general la que tiene que ver con que los seres humanos somos irracionales. Pero la irracionalidad podría ser sólo subjetiva, en el sentido de que alguien

afirma o cree en algún grado dos afirmaciones contradictorias, pero no obviamente. En cuyo caso, no decimos que el cerebro o la mente de esta persona sea contradictoria *per se*, sino sólo que tiene creencias contradictorias. En este caso sucedería un fenómeno no muy conveniente para los computabilistas, tampoco: la ciencia de la mente humana sería trivial, o bien ninguna mente humana existe, o cualquier mente humana se puede describir con cualquier teoría, ya que ninguna teoría lo hace. Sólo podríamos hacer historia, pero nada de predicciones sistemáticas.

Vimos a lo largo del capítulo que quienes tienen como dogma metodológico el TT o el TT extendido, están prejuzgando la cuestión de modo que ningún AG les parecerá sólido pero por razones dogmáticas. El vicio racional en el que pueden incurrir quienes se apegan dogmática a los TT o a los TT* para criticar la posibilidad de sustanciar la verdad de alguna de las premisas del AG, es el de la *petición de principio* y las críticas serían todas orientadas a mostrar que AGs singulares padecen de algún defecto lógico. Muchas críticas a alguna propuesta de AG caen en este problema los más significativos de los cuales los veremos en el capítulo 4, mencionados. Vimos las ventajas de tomar el TT extendido como un símil con el presupuesto computabilista generalizado. De este modo, si se propone una revisión del computabilismo, siempre que sea reducible a un TT extendido, sabremos que tomado como dogma no sirve en principio para refutar la solidez de los AGs.

3. La tesis computabilista es metafísicamente ambigua

> A los recién ingresados a la carrera de Matemáticas en la Facultad de Ciencias de la UNAM les platican como si fuera un secreto que después en sus carreras van a necesitar una *máquina de Turing* que como son difíciles de conseguir, pues que mejor las vayan buscando ya. La broma consiste en que la gente de hecho va a la Plaza de la Computación a preguntar si tienen *Máquinas de Turing* en sus inventarios. Al no encontrarlas en esa plaza los estudiantes se estresan. Habría sido un buen detalle que alguno de los vendedores le proporcionara copia del artículo de Alan Turing de 1950. Luego de leer el artículo de Turing, ¿cómo puede el estudiante estar seguro de haber encontrado la *máquina de Turing*?

3.0 Introducción

En este capítulo clasificamos los tipos de computabilismo que hay en un nivel muy general y hasta cierto grado, poco ortodoxo. El criterio para admitir que una teoría es computabilista es que indique que el fenómeno al que refiere la teoría sea descrito correctamente mediante relaciones incluso abstractas, desconocidas o hasta cierto grado, postuladas. Esta clasificación es útil porque tiene implicaciones y aspiraciones tanto metodológicas como metafísicas distintas, y, también su corrección se evalúa de forma distinta en uno y en otro caso. Esta discusión versa sobre la premisa 2 del esquema general del AG del capítulo uno de la presente tesis. La clasificación que hago aquí es dicotómica: computabilismo cognitivo y computabilismo metafísico. El computabilismo metafísico dice, en general, que la mente humana es literalmente de la misma naturaleza que una *máquina de Turing*. El computabilismo cognitivo lo entenderé aquí como todo computabilismo (o interpretación de una teoría computabilista) como si sólo describiera computacionalmente un fenómeno que podría no serlo en sí mismo. Esta clasificación cumple el propósito de ordenar los tipos de AGs que se han presentado y que se pueden presentar, así como las distintas formas de refutarlas. No tiene las mismas implicaciones que la mente no es –literalmente- una computadora, a que la mente humana no sea describible como una computadora.

Distintas teorías de la mente, como la teoría de la identidad, distintas formas de funcionalismo, conductismo, etc., encajarían de modos distintos en estas dos categorías. El criterio que uso

para clasificar a las teorías como computabilistas cognitivas o no, es simple. La teoría es computable si tiene un diccionario, reglas o principios de inferencia. Si la teoría aspira hacer una descripción literal del fenómeno y es computacional, entonces es metafísica. Si la teoría aspira a hacer una descripción no literal, vaga o relativamente ambigua, entonces es cognitiva. De este modo, prácticamente todas las teorías de la mente humana, si son computacionales, o bien tienen aspiraciones metafísicas o sólo cognitivas.

Por ejemplo, el computabilismo cognitivo admite con facilidad cierta realizabilidad múltiple, y simplemente sostiene que lo que podemos conocer de los fenómenos es descriptible como una relación computacional que no tiene por qué describir hechos en sí mismos computacionales. Esta tesis puede tomarse como un dogma en el sentido quineano del término: sólo se admite como ciencia de la mente (computacional) si el fenómeno referido es finito y mecanizable, todo lo demás indicaría falta de rigor en los conceptos, términos mal entendidos o simplemente pseudo-ciencia. Así, tomado como dogma quineano o principio, evaluamos su corrección no como si describiera la realidad, sino como si nos conduce a ella, si la conserva, si la comprime, si nos permite manejarla mejor, etc. Es decir, qué ventajas teóricas tiene. Una crítica a un principio metodológico de esta naturaleza (dado que es dogmática) se hace más bien en desde una perspectiva normativa. En un caso extremo, podríamos tener entidades que, en sí mismas consideradas, no sean fenómenos computacionales sino fenómenos correctamente descriptibles como computacionales: un jugador de ajedrez en su cabeza puede estar teniendo un tren de pensamientos incomputables, aunque el resultado final de sus movimientos en un tablero sea completamente computacional. En cambio el computabilismo metafísico no tiene esta diferencia ya que en efecto, para ser verdad requiere que los objetos que describe sean objetos sólo computacionales. En este capítulo trazamos justo esta distinción. Además, diferenciamos los computabilismos de metafísico y cognitivo, también en débiles y fuertes. El computabilismo cognitivo fuerte diría que sean como sean los fenómenos descritos, siempre se comportan dentro de los límites de la descriptibilidad computacional. El computabilismo cognitivo débil dice que los objetos descritos podrían comportarse de manera no computacional, si fueran fenómenos no computacionales metafísicamente. El computabilismo metafísico fuerte dice que los objetos son sólo y en todo, objetos computacionales. El computabilismo metafísico débil dice que los objetos son objetos computacionales o se comportan como computacionales, aunque podrían comportarse fuera de estos límites.

Consideremos el debate acerca de la naturaleza, por ejemplo, de los números naturales: son objetos abstractos, i.e., no empíricamente discernibles, no al menos directamente (quizá apenas TTeables: vemos cómo algún agente cognoscitivo apto para matemáticas habla de los números), o bien son objetos concretos, quizá complejos como son las prácticas lingüísticas, más culturales, y sociales, y cerebrales. En esta tesis la perspectiva no es, por razones antes aducidas (evitar circularidad o el vicio racional de la *petitio principii*, en los AGs) –que dependiendo de la naturaleza de los conceptos sepamos qué evidencia esperar para saber qué son y dónde están. Al revés, dependiendo de la evidencia que tenemos de su existencia *ex post facto*, sabemos qué naturaleza tienen. Lo mismo aplica para los AGs: no porque no encontremos los números naturales y nuestro conocimiento de ellos en alguna parte del reino de lo computacional, eso quiere decir que los números naturales y nuestro conocimiento de ellos no exista, sino que dependiendo de nuestro conocimiento sobre números naturales podemos inferir qué naturaleza tienen.

Los objetos TTeables pueden ser o bien modelados por MTs, o bien pueden ellos mismos ser MT. Los objetos que no son TTeables, pues pueden ser total o parcialmente modelables por MTs o, no serlo de plano. Si las clases "totalmente modelable por MT", "parcialmente no modelable por MT" o "completamente no modelable por MT" son vacías o no, es un asunto que está sólo parcialmente dentro de los límites de esta tesis. Los AGs exitosos presuponen que al menos un tipo de objetos (los agentes Gödel susceptibles) poblaría al menos alguna de estas tres clases.

En este capítulo desambiguaremos el término "Turing-computable" al dividirlo en dos sentidos: *i*) el sentido metafísico, y *ii*) el sentido epistémico. El primer sentido equipara las MTs con los objetos físicos y a las relaciones de causalidad, con cierto tipo de funciones *Turing computables*. El segundo sentido equipara las MTs con conceptos que pueden representar objetos por lo menos parcialmente modelables, si la representación mediante MT es al menos parcialmente buena: las teorías que son *Turing computables* y que describen bien a cierta región de la realidad, nos permiten decir de tal región de la realidad que es al menos parcial y cognitivamente computable. Hay problemas muy detallados al respecto de cómo y qué inferir a partir de la tesis de que tal fenómeno es al menos parcialmente representable por una teoría Turing computable, o no, qué elementos la componen, cuáles son los elementos atómicos con los cuales se construyen esas teorías, a qué refieren, etc. Pero como estos son asuntos de detalle, me concentraré en sólo los que considero necesarios o aclaradores. Así,

tenemos al menos cuatro situaciones: un objeto descrito es una MT o no lo es, y la teoría que describe a ese objeto es Turing Computable o no lo es. Cada una de las combinatorias es parcialmente cierta o totalmente cierta.

Esta forma general de entender el predicado *Turing computable* nos dará claridad para acomodar no sólo teorías de la mente particulares, sino incluso teorías futuras. La aspiración es que esta distinción sea exhaustiva, aunque por supuesto podría no serlo. Veremos abajo un problema específico a ese respecto.

Es un problema especialmente interesante que, si usamos como marco conceptual uno que nos impone usar como descripción mínima de la realidad al menos dos términos unidos mediante una función supuesta—conocida o desconocida, cómo podríamos describir computacionalmente estados únicos. Si siempre que hablemos de la mente humana, hemos de tener al menos dos términos

<input, output>

como parte de la definición de la función. Si el *output* es el mismo que el *input* para todo el rango de la función, entonces estamos frente a un punto fijo (que además es la relación de identidad). Hasta qué grado esto califique como computar y más interesantemente, hasta qué grado esto califique como Turing-computar, es un justo el punto de esta tesis. Pero para muchos casos, un estado mental simple y atómico de cosas puede ser referido mediante funciones computacionales de punto fijo.

Quizá un ejemplo ayude a ver la diferencia. Una persona puede representar el cero con el siguiente numeral: **0**. Pero si aplicamos la función *suma cero* (+0), una vez al numeral, pues el resultado es exactamente el mismo, cero, aunque se escriba así: 0+0. Si aplicamos la función '+0' dos veces al numeral 0, el resultado no cambia, pero la expresión sí: 0+0+0. Podemos aplicar de manera infinita la función *suma cero* al numeral 0, y tendremos el mismo resultado. 0 (y cualquier otro número) es un punto fijo de la aplicación recursiva de la función '+0'.

No todas las funciones tienen puntos fijos, pero si hemos de representar computacionalmente un objeto que no cambia en al menos algún aspecto, una manera de hacerlo es mediante la relación de identidad. Ya que la identidad es una función relacional reflexiva y transitiva

(además de simétrica), si el estado mental en el que, subjetivamente nos percatamos de la oración G relevante de un sujeto cognoscente es un estado Turing-computable, por lo menos mediante un punto fijo como identidad, la falsedad de alguna de las premisas de esta los AGs haría de tal estado uno no reflexivo. No hay un objeto computacional en cualquier sentido que viole de algún modo instancias del principio de identidad. De modo que un estado mental atómico y único, es para efectos de esta tesis un caso de punto fijo. Algunas de las razones las vimos arriba: sección "*1.1.4 Tipos de Argumento Gödeliano*", de la presente tesis. Así, si tomamos al operador modal epistémico "saber que" como parte esencial de la mejor forma de describir un estado computacional cognitivo, al mantener uniforme la noción de conocimiento, tenemos que:

$$Sabe^{A}(\mathrm{G})$$

y así, ese estado es tal que

$$\sim(Sabe^{A}(\mathrm{G}) \rightarrow Sabe^{A}(Sabe^{A}(\mathrm{G})))$$

La fórmula de arriba es equivalente a la negación de un axioma de lógica epistémica que se llama S4. El axioma dice:

$$Sabe^{A}(\mathrm{G}) \rightarrow Sabe^{A}(Sabe^{A}(\mathrm{G}))$$

Y aunque el término *saber* aquí es intencionalmente vago, puede especificarse fácilmente como una instancia más precisa:

C. Creencia de que *G*
V. *G*
D. Demostrabilidad de que *G*

$$\sim[\mathrm{Dem}^{\mathrm{A}}(\mathrm{G}) \rightarrow \mathrm{Dem}^{\mathrm{A}}(\mathrm{Dem}^{\mathrm{A}}(\mathrm{G}))]$$[30]

[30] Esta implicación de la incompleción de Gödel está en McCall (1999, p. 529).

Donde “Dem” es un operador modal que representa la *demostrabilidad* del sistema PA+LPO. La discusión sobre Gödel-susceptibilidad se puede ver grandemente enriquecida si se centra sobre la *aplicabilidad* de S4 a agentes racionales, además de las razones por las que sería aplicable o no. Por ejemplo, una fuente posible de inaplicabilidad tiene que ver con la preteórica diferencia de contenido que hay entre

$Sabe^{A}$(p)

y

$Sabe^{A}(Sabe^{A}$(p))

La interpretación intuitiva e informal de S4 ha sido defendida no sólo formalmente sino filosóficamente en Hintikka (1962, pp. 40-59 y 103-125) aunque no ha pasado sin críticas (Williamson, *Knowledge and Its Limits*, 2002, es un ejemplo). Hay por lo menos una relación conceptual en la que parece obvio que *conocimiento* implica *conocimiento de conocimiento*, sin que metafísicamente sea ni necesariamente infinitista, ni contradictorio con posturas finitistas sobre la mente humana. Las relaciones conceptuales que describe S4 pueden darse independientemente de las relaciones, digamos, empíricas aunque no falsificables por ser éstas de una manera en la que *un único hecho empírico* implica una infinidad de verdades lógicas. No es que haya infinidad de verdades empíricas (un estado epistémico), sino que son infinidad de verdades lógicas dada la verdad de al menos *una proposición* empírica.

Puede no obstante usarse la incompleción de Gödel como un caso a nivel del conocimiento matemático y sistemas formales a favor de la posibilidad de que D*p* y ~D.D*p* (No es demostrable que es demostrable que *p*). Pero esto sería confundir la idea de que hay una sola forma de *demostración,* sea formal y sistemática, científica (finitaria y Turing-computable) o no. Es una implicación de las ideas que hay en esta tesis que no carece de sentido del todo, incluyendo a la incompleción de Gödel misma, la idea de que tengamos fuente de conocimiento matemático que no son computacionales. En especial nos interesa aquí que la teoría de la mente sea capaz de explicar no sólo cómo tenemos conocimiento matemático, sino que sea capaz de preservar la certeza de sus *ítems* ciertos. Así, no queda claro que D.D. sea instanciado de forma no polémica por la incompleción de Gödel.

Pero vamos a ir con más detalle en las subsecciones siguientes.

3.1.1 Computabilismo metafísicamente fuerte y computabilismo metafísicamente débil

La metafísica se ocupa de determinar si la realidad o la naturaleza es exclusivamente material, espiritual, ilusoria, si es un conjunto de procesos infinitos, etcétera.

Así, una cosa es decir que metafísicamente la mente es finita y otra cosa es decir que parece infinita, otra cosa es decir que la materia es finita, y otra cosa es decir que la materia es mecánica, etcétera. Las tesis:

a. La mente es infinita
b. El cerebro es finito
c. La mente es una computadora

son demasiado ambiguas para sernos útiles. Aclararemos en lo que sigue, algunas que aunque no son inconsistentes con a-c, son mucho menos imprecisas. La distinción que hemos hecho antes acerca del *computabilismo* [ver capítulo 1], quedan aún imbuidas de una ambigüedad estorbosa que vamos a deshacer ahora.

Generalmente no ayuda mucho invocar la tesis Church-Turing para aclarar el asunto de qué es que algo sea computable o no. Algunas de las complicaciones, como la de qué es que un objeto *sea una computadora* o qué es que un objeto *implemente una función computable*[31] no se resuelven apelando a la tesis Church-Turing. Que un proceso de pensamiento, de digestión o de desplazamiento en el espacio sea temporalmente finito no quiere decir que es de hecho un proceso finito en aspectos no temporales. Avanzar del lugar 0 al lugar 1, a un metro de distancia, aunque es una tarea que realizamos todo el tiempo y que pensamos que es finitamente realizable, puede implicar la realización de un proceso infinito. Entre 0 y 1 hay $\aleph_0$ números racionales. Cuando menos podemos dar una reconstrucción teórica de una tarea espacio-temporalmente finita, mediante un conjunto de descripciones numéricamente infinito.

[31] Véase Bringsjord et al (2006, pp. 2), Potgieter (2004, pp. 2-3), Stannett (2003, 115-117), Maudlin (1989), Chalmers (1993), Cotogno (2003, 181-184).

De tal manera que, un sistema físico puede de hecho ser finito en el sentido de lo que puede hacer a lo largo del tiempo, a lo largo del espacio y a lo largo del tiempo y del espacio. También es el caso que un sistema físico puede *ser modelo* o *hacer verdadera* a una teoría que contenga una cantidad finita de oraciones y sus conexiones lógicas. Las otras dos posibles combinaciones son igualmente coherentes. Un sistema físico infinito en el sentido de lo que puede hacer a lo largo del tiempo, a lo largo del espacio y en ambos, puede ser descrito *correctamente* al menos por una teoría infinita, tal como que un sistema físico finito puede *ser modelo* de alguna teoría de magnitud ω. Trivialmente, por vacuidad, incluso un universo vacío satisface alguna teoría de magnitud ω. Esta teoría a la que todo modelo satisface es la teoría que contiene, por ejemplo, verdades lógicas de la lógica proposicional. Pero descontando el conjunto de teorías de verdades lógicas, obtenidas por medio de funciones de verdad y clausura deductiva, un sistema físico finito o infinito puede satisfacer una teoría no cerrada bajo deducción que sea finita o infinita de tamaño $\aleph_0$.

En este trabajo supondremos que los sistemas físicos son en principio objeto de conocimiento empírico total. Las observaciones empíricas pueden a su vez ser objeto de investigación empírica.

Aunque al menos cierto sistema de estados infinitos, llamémoslo s^{infinito}, sea descrito correctamente por un mínimo de una cantidad infinita $\aleph_0$ de oraciones o $\aleph_0$ constituyentes de oraciones, ello no quiere decir que no haya una teoría con una cantidad finita de oraciones que no sea también satisfecha por el sistema s^{infinito}. Si las oraciones que describen a s^{infinito} son, por ejemplo, axiomatizables, esta es una manera de reducir la cantidad infinita de oraciones verdaderas de s^{infinito} a una cantidad finita de oraciones y ciertas reglas de inferencia. Un sistema axiomatizable como este tiene su infinitud no en la extensión de enunciados sino en las reglas de sustitución, constantes, y demás.

En este capítulo estamos precisando lo que se entenderá por *AG metafísico* y AG cognitivo. El primero es expresado por la tesis 2.1 de la formulación general de los AG, en el capítulo 1 de esta tesis. El segundo es expresado por la adicion a la tesis 2.2 de dicha formulación, pero no necesariamente el 2.1. La solidez del AG metafísico implica la falsedad de 2.1, es decir, que al menos alguna mente humana no Gödel limitada no es consustancialmente idénticas a alguna MT o a la MUT. La validez del *AG metafísico* implicaría que hay una prueba *a priorística* a

favor de la existencia de un objeto no computable, suponiendo que demostración de la consistencia de la aritmética sea a priorística. La solidez del AG cognitivo en cambio sólo implicaría que una teoría adecuada al respecto de lo que es la mente humana es cuando menos no Turing computable.

Un tema común en las discusiones metafísicas es si algo es real o no. Por ser real, se entiende la existencia de un objeto, el que sea, no dependa exclusivamente en la voluntad de nadie, y que el objeto está allí y es lo que es, independientemente de quién lo conciba o cómo lo conciba. Un ejemplo de algo cuya existencia no creemos que dependa de la mente de nadie son los planetas. En cambio es objeto de debate si uno puede encontrar un triángulo por allí tal como encontramos un planeta. No nos interesa si la mente humana o las MTs son reales. Trivialmente, aunque las propiedades de todos los objetos que componen el universo dependan de la construcción mental que *alguien* hace de ellos, esta proposición no descarta cierta manera de realismo, el de la realidad de la mente que piensa los objetos. En cambio sí cabe que hagamos la pregunta, si en un universo completamente anti-realista, la mente o las mentes que determinan las propiedades del universo en cuestión, es computacional o no.

3.1.2 Funcionalismo

El funcionalismo es una posición en filosofía de la mente que procura responder a la pregunta *qué es lo mental*.

A trazos gruesos, el funcionalismo dice que una mente es aquello que *funciona* como mente. Esta posición resulta informativa siempre que sepamos exactamente qué hace una mente. Vamos por pasos. Un riñón es aquel órgano que *funciona* para filtrar la sangre de un organismo cualquiera. Para el funcionalismo, un riñón atrofiado, es decir, que no filtra más la sangre, se sigue llamando *riñón* quizá sólo por *respeto a la memoria*, porque en ese momento el *riñón atrofiado* tiene tanto de riñón como lo tiene un matamoscas. Para el funcionalismo si lográsemos construir un filtro de sangre hecho de silicón y chips, entonces ese aparato tendría más de riñón que un riñón orgánico atrofiado.

En el caso de la mente entendemos que lo que hace a una mente, es *intuir, conocer, ignorar, imaginar, desear, sospechar, calcular, inspeccionar, doler, disfrutar*, etcétera. Un evento de

intuición, o de *saber*, o de *imaginar*, etc. es un evento mental. Explicar así los estados en los que *algo conoce*, es explicar una parcela importante de estados mentales, si suponemos que tener estados mentales de *saber* o de al menos de *creer* es condición necesaria para tener una mente. A estos estados mentales los llamamos *estados cognitivos*. Para el funcionalismo pues algo tiene un estado cognitivo si funciona como tal.

No nos ponemos de acuerdo aún en la parte de qué es exactamente *creer*, *saber*, *sospechar*, *intuir* y otros tantos estados cognitivos posibles. Los estados cognitivos no son necesariamente actitudes proposicionales porque las actitudes proposicionales son más numerosas de lo que lo son los *estados cognitivos*. Además, no es conveniente presuponer en un análisis filosófico que todos los estados cognitivos sean relacionales, sea entre justificaciones, estados de cosas y proposiciones o algunas otras relaciones.

Ahora pongamos en claro lo siguiente. Una función en el sentido matemático es un objeto abstracto que puede o no tener estructura interna. Las MTs son objetos abstractos que tienen estructura interna. Una MT que implementa la función de la suma, le toma de hecho cero segundos computar la suma 2+2 en el reino de lo abstracto. Para el ámbito de la aritmética, *"2+2"* es solamente otro nombre para el número 4, lo que significa que con todo rigor 2+2 no es una operación, sino una etiqueta más compleja para el número 4. Computar con *lápiz y papel* 2+2 puede tomar un segundo o más, pero computarla no es exactamente la suma, de la misma forma que computar es una serie de movimientos dentro de la red causal espacio-temporal, y 2+2 es el nombre de un objeto abstracto, a saber, el número 4. Este es un sentido de funcionalismo especialmente presente en la literatura computabilista, pero no la única.

Otra manera de confundir la expresión "la función de", es al entender "función" como sinónimo de *rol que algo juega al interior de un sistema*. Por ejemplo, cuando decimos que *masticar es una función que puede sustituir una licuadora*, lo que decimos es que el rol de triturar los alimentos dentro del sistema digestivo, puede ser realizado por, por ejemplo, una licuadora. No decimos que de hecho la licuadora *mastique los alimentos*.

Una forma más en la que podemos confundir igualdad funcional de dos cosas, es si entendemos "función" como *objetivo* o *propósito* o *intención*. Cuando decimos que *la función del refrigerador es conservar los alimentos*, otras situaciones u objetos pueden sustituirlo. Una hielera, por ejemplo, puede suplir funcionalmente a un *refrigerador*.

Así, el funcionalismo que nos interesa ahora, si bien es bastante general y poco interesante para un investigador empírico, sí nos permite entender qué relación conceptual guarda con el computabilismo y en particular con el Turing computabilismo.

3.1.2 Máquina de Turing, función y funcionalismo, y criterio extensional de igualdad

El computabilismo es un tipo de funcionalismo. *Funcional* puede engañosamente entenderse como *teleológico* aunque también puede entenderse como no teleológico. No es que de hecho un riñón tenga la finalidad o intención de filtrar la sangre, aunque en otro sentido podría ser descrito. El riñón lo hace sin poder evitarlo. El riñón no puede dejar de filtrar la sangre por aburrimiento, o por sabotaje al resto del cuerpo. Pero que haya una diferencia entre función y propósito, no quiere decir que sean completamente incompatibles. Incluso puede ser que todo propósito resultara modelable como cierto tipo de funciones complejas. Este tema no requiere de que entre en demasiados detalles.[32]

Piccinini (2004) por ejemplo dice que no toda descripción de una *tabla de máquina* es una descripción funcional o *análisis funcional*. El computabilismo implica que la mente humana y el cerebro humano son alguna MT o la MUT. Lo que distingue a una MT de otra es el catálogo de *inputs* y *outputs* que relaciona, no cómo hace para relacionarlos, es decir, cómo *realiza tal función* de inputs a outputs. Así, una descripción de la tabla de máquina del cerebro humano no es igual que un *análisis funcional* del mismo.

Una tabla de máquina es la descripción del funcionamiento interno de una MT. Funcionalmente, una *tabla de máquina* o una *tabla de transiciones* es la especificación de la función matemática que vincula *inputs* y *outputs* para una MT, pero esto no implica que conozcamos cómo el cerebro hace para vincular el *input* con el *output* neurológicamente; sabríamos qué función abstracta implementa el cerebro, pero no exactamente qué neuronas enciende o apaga. La MT es una construcción abstracta que en un sentido siempre hace lo mismo fuera del entramado espacio-temporal actual; moverse de izquierda a derecha, leyendo e

[32] Por ejemplo: para el análisis funcional de estados teleológicos, tenemos que un bombero B^1 tiene el propósito de apagar un incendio *si* o *si y sólo si*, para el conjunto de incendios cercanos B^1 hace por disminuir el fuego. Podemos hacer análisis cada vez más prolijos de las condiciones cerebrales y fisiológicas de lo que es *querer apagar fuegos urbanos*, **aún** sin suponer que las personas estén en condiciones óptimas para lograrlo.

imprimiendo en una cinta infinita. Sin embargo, en otro sentido una *máquina de Turing* MT^1 puede, moviéndose diferente de otra MT^2, llegar a resultados idénticos. La *tabla de máquina* de dos MTs que se mueven distinto entre cuadros para calcular la misma función aritmética, distingue no necesariamente entre la función implementada, sino la tabla de transiciones de las máquinas. Lo que distingue a una MT^1 de otra MT^2 es solamente los pasos que siguen, los cuadros que visitan y los movimientos que ejecutan. **De** una función idéntica a otra de acuerdo solamente al criterio de *inputs-outputs*, pero no de acuerdo al criterio de *tabla de transiciones*, le decimos que son la misma función matemática. En cambio de dos funciones que son idénticas en el nivel de *transiciones de tabla*, decimos son idénticas tubularmente, que son la misma MT. *n*+2, donde *n* es sustituido por 2, es igual a 2+2, 4. Esta es una manera de calcular el resultado. Sin embargo *n*+2 sustituido por 2, es igual a 2+1+1, que también es 4. Ambas formas computan la misma función matemática, pero tabularmente son diferentes. La función es distinta, aunque la función calculada es la misma en ambos casos.

La *tabla de máquina* del cerebro nos diría específicamente cuál es ese mecanismo, mientras que el análisis funcional apenas basta para decirnos que se implementa una función.

Hay dos sentidos en los que el *funcionalismo* y el *computabilismo* admiten que la *función* que une un *input* con un *output* sea implementada por objetos completamente diferentes (por ejemplo, un riñón orgánico a uno artificial). A esto lo conocemos como *realizabilidad múltiple* (*multiple realizability*). Hay un sentido en el que todo computabilismo es una forma de funcionalismo. Las *máquinas de Turing* implementan funciones en el sentido de *inputs* o *entradas* y *outputs* o *salidas* (Turing, 1950). No obstante lo anterior tenemos una manera de explicar esto; las MTs implementan una función. El análisis funcional teleológico no es incoherente con el análisis funcional consistente en ser una *tabla de máquina* de MT. Las funciones, sean entendidas teleológicamente o no, pueden ser entendidas similarmente, analógicamente presupone Putnam según Piccinini (2004, p. 817):

> En este pasaje, Putnam señaló que las teorías psicológicas podrían ser formuladas de dos modos: uno describe disposiciones conductuales y mecanismos fisiológicos, el otro describe "estados mentales" e "impresiones". Entonces Putnam sugiere que si fuera posible formular una teoría psicológica "abstracta" en términos de estados mentales, entonces esa teoría sería para una teoría psicológica que describe mecanismos fisiológicos en la misma relación que el programa de una MT está con respecto a

> descripciones de sus implementaciones físicas. La sugerencia de Putnam fue ofrecida no como un argumento sino como una analogía con descripciones de MTs.[33]

Podemos hacer análisis de la mente y el cerebro que sean funcionales, en distintos niveles y todos expresables como tablas de máquina, si es el caso, o simplemente como rangos y dominios de entradas y salidas. Depende qué objetivos tenemos y con qué recursos contemos. Para los efectos de esta tesis, toda teoría funcionalista de la mente se considerará como una teoría computabilista de la mente.

Un aspecto más que es importante tener en claro es el aspecto normativo que pueden tener algunos de los elementos que consideramos aquí. Para dar sólo un ejemplo, las funciones de la aritmética tienen con respecto a sus respectivas implementaciones una relación tanto descriptiva, valor de verdad cuando son isomórficas, como uno normativo, cuando sirven para corregir un procedimiento: hay errores matemáticos, no sólo funciones. Si una MT que suma dos, alimentada del input 2, arroja como resultado 5, esa MT hace mal la suma de dos o simplemente no sumó dos aunque haya implementado *alguna* función. En el caso de las funciones entendidas como objetivo o de las funciones como el *rol que juega algo dentro de un sistema*, pueden servir como normas para comparar otras MTs o implementaciones de MTs, o realizaciones de otras funciones. Eso no quiere decir que *sumar* sea una norma necesariamente de tipo moral para una MT, o que de hecho un riñón que incumple con *el objetivo* de filtrar la sangre, incumple una norma de tipo moral o jurídico. El riñón tiene como *objetivo*, i.e., *es su función*, filtrar sangre. Un policía tiene como *objetivo* y además *juega el rol* de proteger a la ciudadanía[34]. No obstante, cuando un riñón no filtra la sangre, no decimos que sea un *mal riñón* en el mismo sentido en el que cuando un policía *falla en proteger* a la población es un *mal policía* o incluso, que tal policía *es inmoral o corrupto por fallar en cumplir su función*. Cuando una función en cualquiera de los sentidos es mal implementada o computada, hay diferentes formas de *error* allí que es crucial distinguir con cuidado.

[33] Mi traducción de: In this passage, Putnam pointed out that psychological theories could be formulated in two ways: one described behavioral dispositions and physiological mechanisms, the other described 'mental states' and 'impressions'. Then Putnam suggested that if it were possible to formulate an 'abstract' psychological theory in terms of mental states, then that theory would stand to a psychological theory describing physiological mechanisms in the same relation that TM programs stood to descriptions of their physical realizations. Putnam's suggestion was offered without argument, as an analogy with descriptions of TMs.

[34] Por lo menos esto los justifica en principio.

Con jerga formal, una función *f* es implementada por otra función *f'* solamente si para cada elemento del dominio, arroja exactamente al mismo elemento del contradominio. Así, para resumir, un sistema *s* implementa una MT^{1} si una descripción cuidadosa de la *tabla de transiciones* de MT^{1} es isomórfica con la descripción cuidadosa del sistema en cuestión, de tal forma que para cada *input*, ambos sistemas arrojen el mismo *output*[35]. Una máquina de Turing MT^{1} es implementada por un *sistema s*, solamente si para cada *input i'* del sistema *s*, obtenemos el mismo *output j'* correspondiente al output *j* de la MT que se obtendría del input *i* correspondiente al input *i'*. Si nos importan aspectos *internos* de la función como tiempo, recursos usados, etcétera, podríamos también considerar el tiempo y recursos que usan tanto el sistema como la MT: aunque para la máquina los recursos son *n* pasos desvinculados de tiempo y espacio, para el sistema el número de pasos pueden también medirse temporal y con respecto a los recursos. El punto aquí es que podemos hacerlo, aunque de hecho no interese a alguien hacerlo en un caso específico. Dicho de otro modo: dado que el vínculo causal que une al *input* y al *output* puede ocurrir en el espacio y en el tiempo para *s*, que dos máquinas sean indistinguibles por cómo relacionan *inputs* y *outputs* no quiere decir que sean la misma, razón por la que hemos introducido la semejanza de entre la descripción del sistema *s* y *la tabla de transiciones* de MT^{1}. Podemos especificar aspectos específicos, como que s use cierta otra función o mecanismo, y establecer distintos grados de igualdad entre sistemas que implementan funciones o MTs.

Un *lema* más; un par de implementaciones *s*'y *s*'' de una MT^{i}, son idénticas a otra MT^{j}, solamente si además de asociar a cada *input* el mismo *output*, lo hace en el mismo tiempo y con cantidades de energía y tabla de transiciones similares, o incluso con aspectos arquitectónicos específicos (por ejemplo, si se usan neuronas, o *siliconas*—neuronas de silicón –), o criterios semejantes. El grado con el que decimos que dos sistemas s' y s'' son funcionalmente iguales es hasta cierto grado variable. Este es el criterio de *igualdad extensional estricto* para MTs, funciones, constantes y demás. Al criterio de igualdad entre MTs, teorías o predicados que respeta extensionalidad y no considera el tiempo o energía requeridos por la implementación de la función, lo llamamos *igualdad extensional laxa*. La igualdad extensional laxa o estricta, continua, para dos MTs es más o menos clara.[36]

[35] Para algunos detalles más sobre esta discusión véase: Chalmers (1993, 1994) y Brown (2004).
[36] También hay igualdad extensional entre dos teorías sobre sistemas físicos si ambas permiten exactamente las mismas inferencias, sean para explicar o para predecir la conducta de un sistema físico. Un par de conjuntos de enunciados incoherentes cerrados bajo deducción, son teorías extensionalmente indistintos, aunque lo enunciados

3.1.3 Criterio de igualdad representacional entre MTs, y el tema de la realizabilidad múltiple (*multiple realizability*)

Hay una noción útil para comparar MTs y teorías Turing computables, que es especialmente interesante. La igualdad transicional entre dos tablas de transición de dos MTs es el isomorfismo. Si aceptamos un criterio de igualdad diferente al extensional, podemos establecer positivamente la igualdad incluso entre dos conjuntos mal fundados en general, o conjuntos de input, output y tabla de transiciones en particular. Para más detalles sobre este tema véase Barwise y Etchemendy (1987, pp. 34-58). Este análisis de conjuntos resulta particularmente interesante a la hora de evaluar las diferentes condiciones de igualdad como es la igualdad interteórica, ya sea para comparar *descripciones de tabla de máquina*, o para explicar propiamente fenómenos intensionales como los *puzzles* (rompecabezas) de Frege y las propias paradojas de auto-referencia. Es una consecuencia del computabilismo que, eliminando terminología inútil pero conservando el vocabulario relevante, dos teorías verdaderas, y completas sobre un fenómeno mental, deben ser isomórficas entre sí.

La comparación de igualdad entre dos sistemas computacionales pasa por la comparación de todas las teorías que describen veraz, completa, coherente y óptimamente a los sistemas en cuestión ¿Qué problemas pueden surgir al respecto de la comparación entre dos sistemas mediante dos teorías verdaderas T^1 y T^2? ¿Qué tipo de igualdad debe haber entre T^1 y T^2 para que refieran al mismo sistema? La respuesta es que dos teorías son iguales si al sustituir todas las variables de T^1 y de T^2, ambas teorías son satisfechas por los dos sistemas. Hay una relación de isomorfismo óptimo entre dos teorías T^1 y T^2 solamente si s^1 y s^2 son extensionalmente iguales[37]. Ahora, las teorías T^1 y T^2 pueden ser generales al referir a las 'mentes humanas'. Las teorías pueden ser parciales al especificar un aspecto del fenómeno de la mente humana, como es los estados cognitivos, o los estados de creencia, o de conocimiento, o de imaginar, etc.

---

describan, *con supresión mínima de la contradicción*, cosas completamente distintas. Este es un caso extraño de la igualdad de dos teorías que no tienen por qué compartir mucho.

[37] En Hofstadter (1979, pp. 573) hay un cuadro de relaciones de isomorfismo entre computadoras, cerebros y sistemas físicos, sin perjuicio de que entendamos por sistema físico un término que incluye a computadoras y cerebros, y atendiendo a la diferencia que hay entre dichas teorías y no necesariamente al referente de sus términos.

La relación de isomorfismo es una función biyectiva, transitiva, simétrica y reflexiva que mapea dos sistemas, físicos o no, teóricos o no, preservando proporcionalidad de las relaciones de los elementos dentro cada uno de ellos. Un mapa de la Ciudad de México es isomórfico con la Ciudad de México si es un buen mapa y me dice entre otras cosas que, de Norte a Sur, la Torre Latinoamericana está al Norte de Ciudad Universitaria. Conviene tener claro que isoformismo es un tipo de identidad, pero de estructuras (o sistemas). Dos teorías pueden ser a su vez isomórficas entre sí o con respecto a hechos, situaciones u objetos. Por ejemplo, una teoría de números puede ser isomórfica con una teoría de literales, y así una estructura de números serlo con una estructura de numerales. Un holograma de la ciudad de México es isomórfico con un mapa de la ciudad de México. Pero hay un cierto *nivel de descripción* en el que la Ciudad de México hoy es más isomórfica con la ciudad de Nueva York, de lo que es la Ciudad de México hoy isomórfica con la Ciudad de México en 1600. Además la relación de isomorfismo como la entenderé aquí admite gradualidad, tal como en el ejemplo anterior, la Ciudad de México en 2010 *es más isomórfica* a Nueva York, que a la Ciudad de México en 1600.

Si un sistema cognitivo es *en cierto grado* isomórficamente igual a otro, éste puede según el computabilismo, ser realizado por sistemas (físicos o no) distintos. Por ejemplo un sistema cognitivo de carbono (humanos) o de sílice (robots), o de *baba verde* (marcianos). Una computadora, cuya descripción al nivel químico, podría estar realizando el proceso de dolor o de *creer que* de un cerebro humano para cierto nivel de descripción, de acuerdo a una noción quizá demasiado grosera de *creer que* o *dolerse*. Según esta caracterización, si las categorías mentales son funcionales, un hormiguero *podría* padecer de una migraña, dada cierta noción grosera de migraña que podríamos tener.

Tampoco debiera sorprendernos descubrir que la noción correcta de *migraña* sólo puede ser implementada por neuronas de primates superiores, por ejemplo. Estamos lejos de tener conocimiento completo y más lejos aún de la ausencia de polémica al respecto de cómo suceden incluso las funciones más básicas de las mentes y los cerebros.

Nociones mentales como la de *dolor de cabeza* o *ver un color rojo* tienen componentes fisiológicos (que sin mucho problema aceptamos como Turing computables) y componentes subjetivos, que quizá con más problema aceptemos como Turing computables. La relación

entre un componente subjetivo y los posibles análisis empíricos (no subjetivos) de los casos de estados mentales de dolor es compleja Aunándole el tema de la *realizabilidad múltiple* tenemos una complicación más. Una teoría computabilista de la mente parece comprometerse con la tesis de que no es esencial el sustrato físico de un evento, siempre que dos sistemas cumplan con la función mental dada ya sea en sentido matemático (pares ordenados input-output), en el sentido del propósito o incluso en el sentido del rol que juega algo en un sistema más grande. Un radiador puede ser suplido por un clima gélido para el objetivo de enfriar, pero también para el *output* de enfriar como para el rol de enfriar un motor.

Es una obligación del computabilismo explicar casos raros de la *realizabilidad múltiple* y cómo sucede ella, además de hasta que grado es admisible. Asimismo de cumplir con el criterio funcional de pares *entrada-salida*, dos sistemas deben tener una estructura que sea en cierto nivel isomórifica si queremos evitar decir que, por ejemplo, la *bocca della veritá* del capítulo 1, es un sistema cognitivo sobre bases que no parezcan demasiado *ad hoc*. Por lo demás debe ser posible también que haya descripciones isomórficas entre sí y divergentes entre ellas de estos dos sistemas. Por ejemplo, una teoría verdadera, óptima y completa en Mandarín sobre la traslación y rotación de la tierra, es isomórfica con una teoría verdadera, óptima y completa en Castellano sobre el mismo fenómeno.

Determinar exactamente cuál es ese nivel, es un problema que ha apuntado sobre todo Putnam (1988, pp. 72-73, 120-125) entre otros, y una respuesta ha sido abordada por Chalmers (1996), Maudlin (1989) y Block (1981). No obstante para que el computabilismo no sea trivial y no cualquier sistema implemente un estado computacional de dolor para cualquier teoría de lo que es dolor, creo que se pueden establecer controles teóricos. Por ejemplo para explicar funcionalmente el dolor debemos apelar a un nivel de descripción de tal sistema que sea un efecto causal del mismo no de alguna interpretación o de algún sistema conjunto, y que psicológicamente sea, el dolor, la causa física de aquello de lo que el dolor es causa incluyendo de preferencia, a los elementos subjetivos también si es que estos son componentes esenciales de la neuralgia. Si no lo son, pues no. Para usar un ejemplo, no porque el movimiento de las hormigas de una colonia *se pueda interpretar* como dolor, quiere decir que *implementan de hecho* un estado mental de dolor; hace falta más evidencia, de modo que no se pueda interpetar como dolor sino que *por las condiciones, deba interpretarse como dolor*. Aunque en el peor de

los casos, pensar que cierto comportamiento colonial de las hormigas debe interpretarse como dolor, i.e., la colonia tiene dolor.[38]

Con Chalmers (1996) estoy de acuerdo en que la relación que une un estado cerebral y/o un estado mental dado, con otros debe ser un nexo causal suficientemente dilucidado para poderse atribuir una creencia o un dolor de tal manera que queden especificadas condiciones suficientes para establecer el vínculo de causalidad inequívocamente. En otras palabras, necesitamos por lo menos una buena noción o una teoría verdadera de *dolor* o *del significado de los términos que componen una creencia* para discernir entre casos genuinos de *dolor* y meros casos posibles de *dolor*, y entre casos de creer que el agua quita, de hecho, la sed.

La función que rige *el paso a* y *de un estado mental a* otro estado de cosas u otro estado mental, no es sólo una relación tan general como lo es un condicional material. La especificación de la función que une computacionalmente a un estado mental dado con otro o con otros hechos y efectos no mentales como acciones, no tiene que ser ~~es~~ exactamente la misma que une a un *input* con un *output* de una MT. Los estados mentales a veces no causan reacciones sino otros estados mentales. Así que toda la transformación interna o externa que hay entre un *input* o un estado mental pasando quizá por otro estado mental, finalmente hasta el *output* puede, en teoría, ser explicado por una psicología computabilista suficientemente rica y provista de evidencia empírica, y auxiliada de generalidades acerca del mundo *en* el que están las mentes. Esto equivale a decir que aunque ignoremos hoy día las características específicas de la causalidad psicológica, no implica que no las podamos establecer en algún momento futuro, a posteriori, sin variar gran cosa nuestra física y epistemologías actuales. Si las relaciones que nuestra teoría de la mente sólo son entre estados mentales, podemos hacerles análisis funcional, si sólo son sobre estados cerebrales, igual se puede, en caso de que sean mixtos, también y aún si sólo hacemos análisis de conductas también quedan dentro de los límites del funcionalismo que tomare como tal para los efectos de esta tesis.[39]

---

[38] La idea general de Putnam contra el funcionalismo mediante *realizabilidad múltiple* es que no hay un hecho intrínseco a un sistema *s* (físico, para Putnam) al respecto de si está teniendo un estado mental de, por ejemplo, *creer que el agua quita la sed*. Si así fuera, entonces habría una explicación interna al sistema *s* al respecto del significado-referencia de *agua* en su creencia. Sin embargo, el significado-referencia del término "agua" es el agua fuera del sistema *s*. Así, los estados funcionales de creencia están crucialmente determinados por hechos externos al sistema *s* en cuestión.

[39] Considero importante subrayar aquí un asunto conceptual. Que dos estados cerebrales sean un mismo estado mental, implica que un estado físico *e*1 es igual a uno *e*2, en algún nivel de descripción. Y silo que caracteriza a los estados mentales no es alguna cualidad meramente intrínseca (fenomenológica) sino su entramado causal con otros estados mentales o cerebrales, o hechos, parece que estaríamos aquí postulando que debe haber por o menos un nivel de análisis en el que dichos dos estados cerebrales *e*1 y *e*2, más sus causas y sus efectos son iguales. Es

Así el computabilismo que no se refuta con *realizabilidad múltiple* supone que una vez eliminada la arbitrariedad en la interpretación de *inputs*, *transformaciones de estados de máquina,* y *outputs* hay una descripción tal que justifica suficientemente el isomorfismo entre las funciones que constituyen, por ejemplo *creer que, saber que* y *dolerse*, y una vez así especificado dicho isomorfismo, sepamos cómo es realizable por sistemas, qué recursos bastan al interior del sistema y cuáles fuera de éste. No es que el comptutabilismo sea completamente no trivial, tiene partes triviales como quizá ciertas definiciones o implicaciones, pero tiene partes no triviales que constituyen verdadero avance en el conocimiento empírico de la mente. Para el funcionalismo como lo entendemos aquí es suficientemente sencillo al menos en principio realizar el trabajo de investigar a la mente. Que en algún caso particular se haya hecho mal, no debería achacarse al funcionalismo en general, sino sólo a esa especificación particular. Todo lo que comparten, pues, estados mentales, cerebrales o causas psicológicas iguales es al menos cierta igualdad funcional. Cuando menos la presente tesis supone que tal descripción funcional existe. De cualquier modo, con todo y esto, las ideas funcionalistas y computabilistas aunque no carecen de problemas, tienen también grandes fortalezas: no cualquier objeción sirve para echar por tierra esta clase de teorías sobre la mente y el cerebro humano. Esta manera también nos permitiría saber cómo específicamente *escribir* o *comprender* la oración Gödel, es causalmente posible o imposible que un agente sea MT u óptimamente describible mediante una teoría completamente Turing computable. Con toda la adaptabildiad y plasticidad de esta clase de teorías, todavía tienen una característica más: hay dos modos de interpretarlas.

3.2.1 Dos tipos de computabilismo

---

decir, parecemos estar postulando la existencia de al menos una causa. Cuál es la cantidad mínima de causas psicológicas para que el computabilismo no sea refutado en principio por un argumento como el de Putnam (1988) con base en la *realizabilidad múltiple*, sin que por ello mismo la *realizabilidad múltiple* quede eliminada de principio como una característica deseable de un programa de filosofía de la psicología naturalista. Tal cantidad mínima es: uno, *un evento*, probablemente *causado* o simplemente *superviniente*. En este sentido hay una teoría posible de ello y más si el evento causado o superviniente es *ex hypothesi* computacional. Los eventos psicológicos físicamente instanciados que sean del mismo tipo tienen que implementar físicamente una misma función, de lo contrario en efecto la *realizabilidad múltiple* parecería destinada al fracaso; no debe ser posible que para un agente es una función de dolor, es *in totto*, una de júbilo para otro agente. Incluso un masoquista debe sentir dolor, para significarlo placenteramente y hasta ahí debe haber una semejanza funcional entre los estados de dolor del masoquista y del que no lo es. De lo contrario, no sería masoquista.

En este subcapítulo abundaremos en cada una de dos grandes maneras de sostener el computabilismo al respecto de la naturaleza metafísica de la mente. Aquí se explica de qué manera es ambigua la atribución de computabilismo para la mente humana. Una de las dos formas dice algo así como: no sé si la mente humana es una computadora, pero se comporta (funciona) como una. La otra forma dice: la mente humana se comporta (funciona) como una computadora y es una computadora.

3.2.2.1 Grado de arbitrariedad de la representación

Imaginen un equipo de futbol soccer. A cada uno de los 24 jugadores de un equipo, incluyendo a la banca, les asignamos un número. El número debe ser visible en la camisa del jugador y éste facilita la identificación que hace el árbitro de infractores y el control de la gente que puede jugar en nombre del equipo. Quizá sirva para otros fines, pero no es forzoso para un jugador que se le asigne algún número de acuerdo a la posición que juegue o si habrá de jugar siquiera, por ejemplo.

Los números de las playeras de los jugadores de futbol soccer representan a los jugadores de forma arbitraria, y según las reglas de la FIFA no describen de ninguna forma ninguna peculiaridad del jugador, posición, habilidades o demás. Esta relación por supuesto cambia de equipo a equipo y de acuerdo a diferentes anécdotas. Al futbolista alemán Michael Ballack le ha tocado el número 13 en el equipo Chelsea Premier Club de la Liga Premier, pero Ballack no tiene muchas características del número trece; Ballack no es *número primo* y Ballack tampoco es la mitad del número veintiséis, Ballack no sirve para contar objetos que están en número entre doce y catorce, y así. A la representación numérica que no describe, como en este ejemplo, la llamo *representación finita nominal*. Es una representación finita, porque el número del futbolista es un número finito, por ejemplo “13” de Ballack, pero es nominal porque funciona como *nombre propio*: no describe necesariamente.

En cambio, podemos asignar la etiqueta *trece* a un conjunto de trece cosas, o al treceavo elemento de un conjunto. El *treceavo presidente* de México desde que entró en vigor la Constitución de 1917 es Gustavo Díaz Ordaz. Gustavo Díaz Ordaz es representado *ordinalmente* por el número *trece* en la lista de presidentes de México a partir de la Constitución de 1917. Una docena de huevo más uno de pilón, son trece. Este conjunto de

huevos son *cardinalmente* representados por el número *trece* en la cantidad de huevos que componen a tal conjunto. A la representación numérica que sí describe, como en este par de ejemplos, la llamo *representación finita descriptiva*. Es una representación finita, porque el número de huevos o el relevo presidencial en México desde 1917, es un número finito, "13", pero es descriptiva porque da información acerca del referente.

Pensemos en una complicación posible esclarecedora. Hay algunos edificios cuya planta número trece se ha bautizado diferente en consideración a los triscadecafóbicos. El nombre que se le da a la planta número trece es, de acuerdo a la Wikipedia[40], piso "12A" o "14". Otra forma de lidiar con el asunto es clausurando el acceso público a la planta que es la número trece, al reservarla para ser *cuarto de máquinas*. Quienes temen al número *trece* y no sólo a las grafías que lo representan (como "13", "10+3" o "XIII") y a todo aquello que es descrito por tal número temen al piso 12A o 14 de tales edificios. Quienes temen a la noción denotada por el numeral "13", pues dependiendo de su filosofía de las matemáticas y de su metafísica general, quizá nunca tengan un episodio de miedo irracional al trece en todas sus vidas. Pero quienes temen al numeral que denota al trece, "13" o "XIII", pues probablemente se queden tranquilos con las medidas de rebautizar a la planta trece como "12A" o descaradamente como "14". No obstante esto, en los edificios en los que al piso trece se le bautiza arbitrariamente como piso "14", pierden la función descriptiva que esta numeración normalmente suele tener. Imaginen a una persona que tiene que subir a pie al piso 14 de un edificio sensible a los triscadecafóbicos. Si esta persona ignora tales medidas, puede pensar que subir al piso 14 es justo demasiadas escaleras, pero no las escaleras que llevan al piso 13. Esta persona pensaría erróneamente que subir a pie al piso 14, es subir demasiadas escaleras y quizá, imaginemos, podría perder una entrevista de trabajo o de negocios por un error en esta etiquetación.

3.2.2.2 La finitud de la representación

Podemos representar un estado de cosas mental, pero también podemos representar conjuntos de estados de cosas mentales, y estados de cosas mentales y no mentales como objetos finitos. Los límites en las combinaciones no importan. La representación de una causa, no obstante, no es la representación de *un* estado de cosas, sino del conjunto de al menos dos estados de cosas,

[40] "Thirteenth Floor", Wikipedia en inglés, visitada el 29 de Octubre del año 2010. <URL = http://en.wikipedia.org/wiki/Thirteenth_floor>.

uno siendo la causa, y otro siendo el efecto. Éste análisis puede variar en tamaño y detalle dependiendo de casos concretos. Los intereses teóricos y la cantidad de subrutinas a considerar los determinaría por ejemplo la comunidad científica.

3. La Francia pre-revolucionaria era opresiva y la desigual.
4. La Revolución Francesa buscó promulgar leyes de protección social.

Podemos pensar que 3 es causa de 4 sin demasiados problemas. Las condiciones de miseria de un país pueden terminar llevando a ese país a un desgaste social tal que las instituciones se debilitan. Esto, le puede parecer a alguien, suele ser causa suficiente de un proceso social revolucionario armado. A un historiador quizá le interese enterarse de más sucesos acaecidos entre 3 y 4. A un psicólogo además puede interesarle específicamente la subrutina de un sistema físico específico; el cerebro de Robespierre. Ambos teóricos pueden estar queriendo entender mejor todo el proceso causal que vincula a 3 con 4. Sin embargo eso no quita que esperemos contar con representaciones finitas, sea nominal o descriptivamente, de los estados de cosas que nos llevan de 3 a 4. Estas representaciones en donde están las descripciones son los libros de historia.

Aunque la revolución francesa haya sido causada por sucesos continuos (infinitos), quizá queramos que nuestros libros de historia del periodo sean sucintos (al menos nominalmente), independientemente de que hagan alusión o no a características no computables (al menos nominalmente). Un libro de aritmética en cambio, es finito (tiene un número finito de páginas) pero alude (al menos nominalmente) a fenómenos que son descriptivamente infinitos. No hay textos que sean descriptivamente infinitos: no tenemos el papel ni el espacio para hacerlos y guardarlos.

3.2.3 Causalidad psicológica

Para los objetivos de esta tesis, donde hablo de *computabilismo* simpliciter, en realidad es importante tener en cuenta que se trata de *computabilismo* y no de computabilismo no-finitista (*hipercomputabilismo* p. ej.). También es importante notar que aunque la *relación de causalidad* no es necesariamente una relación computacional, sí es representable como una relación computacional que vincula *inputs* con *outputs*. Las particularidades metafísicas que

componen a la relación de causalidad y cómo se distinguen de las relaciones computacionales o sintácticas pueden ser muy distintas pero al menos la representación de un vínculo causal no tiene por qué no ser computable. Parte del interés en esta tesis es el de establecer si la mente humana *puede saber si la oración Gödel* es verdad y qué implicaciones metafísicas tiene esto, en concreto si *saber que la oración Gödel es verdad* es, para todas las nociones de saber, un proceso computable o no, o si el estado mental correspondiente es él mismo no Turing computable sin importar si el proceso causal que lo genere es computable.

Al vínculo que une a un evento cualquiera con un estado mental humano como efecto, o bien al vínculo que une un estado mental humano a otro, lo llamaremos *causalidad psicológica*. Este vínculo lo podemos ver como una función como ya vimos y sacar las consecuencias. Queremos saber en particular si la causalidad psicológica es un proceso computable, en especial si la descripción óptima de lo mental es computable o no. Así escribimos en adelante las dos tesis computabilistas que vimos en el capítulo 1 de la presente tesis.

> 2.1 Las mentes humanas son *consustanciales* al menos a una MT o la MUT.

Queda reformulada como sigue:

> 2.1* Los estados mentales son estados de cosas finitos. Las relaciones causales que unen a estados de cosas con estados mentales y a estados mentales con otros estados mentales, son secuencias finitas de otros estados de cosas y/o estados mentales, o ningún estado de cosas de plano.

Y la otra interpretación es:

> 2.2 Las mentes humanas son *óptimamente representadas* por alguna teoría dentro de los límites de al menos alguna MT o la MUT. (Esto es consistente con la posibilidad de que haya algún nivel o algunos niveles de buenas descripciones materialistas de las mentes)

Queda de la siguiente forma:

2.2A* Los estados mentales son estados de cosas óptimamente representables finita y descriptivamente por teorías finitas Turing computables. Las relaciones causales que unen a estados de cosas con estados mentales y con otros estados mentales, son *óptimamente representables* finita y descriptivamente por teorías finitas Turing computables.

Y

2.2B* Los estados mentales son estados de cosas óptimamente representables finita y nominalmente por teorías finitas Turing computables. Las relaciones causales que unen a estados de cosas con estados mentales y con otros estados mentales, son *óptimamente representables* finita y nominalmente por teorías finitas Turing computables.

Una diferencia importante entre una teoría óptima finita descriptiva (TOFD, en adelante) y una teoría óptima finita nominal (en adelante TOFN), es que la TOFD supone menor generalidad que TOFN. Por ejemplo la física trabaja con TOFD y las matemáticas y la lógica trabajan normalmente con TOFN. Una característica extra crucial que tiene una TOFD es que nos permite extraer conclusiones metafísicas a partir de la teoría misma, mientras que TOFN, no necesariamente.

3.2.4 Causalidad computacional

No importa cuál sea el énfasis que queramos dar a los estados mentales, si como los vínculos que unen a comportamientos, o los vínculos que los unen más ciertas otras características psicológicas, ese vínculo causal entre cierto estímulo y cierta reacción, simple o complejo, si es un proceso finito con relaciones nominal o descriptivamente contables entre sus elementos, debe considerarse computabilista en alguno de los sentidos que vimos antes. Así un sistema es causalmente computacional si los estados del sistema y la transición causal entre ellos son finitos ya sea representacionalmente o en sí mismos. Esto quiere decir que uno termina de describir los estados del sistema aunque para un lenguaje provisto de una cantidad infinita numerable o no numerable de constantes, la mayoría de las cuales no tenga referente en alguno

de los estados del sistema físico en cuestión, tengamos una cantidad infinita numerable o no numerable de oraciones sean falsas entre una cantidad finita de oraciones verdaderas[41]. Las descripciones finitas de un sistema físico deben ser positivas. Esto es, los enunciados deben establecer vínculos positivos entre objetos con referente y clases u otros objetos con referente. Si el lenguaje usado para describir tal sistema físico contiene una cantidad infinita numerable o no de constantes, entonces la mayoría de las oraciones positivas del lenguaje serán falsas.

Quizá sea aceptable que los estados de un sistema físico o de un conjunto de sistemas físicos, puedan ser infinitos numerables, pero que la conexión causal entre ellos sea, no obstante, computacional. La causalidad computacional implica que el tiempo está constituido también de unidades discernibles. Una hipótesis que es compatible con esta forma de causalidad, es su división temporal o espacial sin fin. Es decir, aunque el paso causal de un estado físico dado a otro, contenga siempre una cantidad finita de pasos causales, la cantidad de pasos causales totales bien puede ser prolongada infinitamente, en la realidad o al menos en teoría.

Una forma de entender al computabilismo nos diría entonces que la causalidad psicológica computacional es causalidad computacional. A esta tesis la llamaremos computabilismo metafísico. No es una muy violenta equiparación entre *sistema psicológico computacional* y *causalidad psicológica computacional*, ya que finalmente cada estado computacional es computable a su vez. Dos estados computacionales uno de los cuales es *entrada* y el otro es *salida*, están unidos por fuerzas computacionales *in se*.

3.3.1 Computabilismo teórico

Hay otra manera de entender el computabilismo además de la forma anterior, i.e., que la causalidad psicológica no es causalidad computacional. El computabilismo teórico (o cognitivo) se compromete solamente con que aunque los estados de los sistemas cognitivos y sus relaciones causales, no estén compuestos de *átomos* (en el sentido etimológico del término), es decir, que sean divisibles al infinito, pero que las teorías óptimas que tenemos de

[41] Entre el centímetro 0 y el cm. 100, hay 100 centímetros. Si el espacio fuera indivisible a partir de los milímetros, entre el cm. 0 y el cm. 100, habría mil milímetros, y sería falso que habría diez mil diezmilímetros, o cien mil cienmilímetros, o un millón de millónlímetros, y así. No los hay, los podemos imaginar, pero no los habría.

ellos son computables. ¿Cómo es esto posible? La razón es simple, aunque requiere alguna elaboración.

Veamos con un ejemplo. Supongamos que hay un universo U^1 en el que sucede lo siguiente. Un sistema s^1 viaja del punto A al punto B, en una cantidad de tiempo *t*. Este es un evento que consideraríamos perfectamente capturable por una teoría T^1 con una cantidad de enunciados positivos no mayor que $\aleph_0$. La velocidad es de A-B/*t*, donde A-B es la distancia que hay entrambos puntos. Supongamos que la distancia de A a B, es infinitamente divisible en U^1. Así, si A es el punto de inicio, 0, y B el de llegada, 1, entonces la cantidad de puntos entre A y B, es la misma cantidad que números racionales hay entre 0 y 1. La cardinalidad de racionales que hay entre 0 y 1, es $\aleph_0$. Si el evento causal del tránsito de s^1 de A a B se da de hecho en *t*, ese evento puede ser reconstruido mediante la función f = n/2 + n/4 + n/8…. En este evento los estados son infinitos, no obstante hay una TOFD y TODN para modelarlo. Supongamos que el evento es el paso de un estado cerebral a otro, o de un estado mental a otro. El paso bien puede ser gradual. Si el paso responde a ciertas reglas de gradualidad ello no tiene por qué falsear a toda representación teórica de cardinalidad finita. En concreto, tal evento puede ser descrito de la siguiente forma:

5. s^1 en $t - t$ está en A.
6. s^1 en $t - t/2$ está en B/2.
7. s^1 en t está en B.

La teoría $T^{5\text{-}7}$ la componen los enunciados 5 a 7, que no son más que tres; todas las relaciones entre ellos son obvia y completamente computables. Prácticamente cualquier computadora actual puede escribirlos y realizar muchos cálculos con ellos. Los valores numéricos de los enunciados de 5 a 6 no describen al número de metros o millas en particular. Así que puede ser cualquier valor numérico. Esta teoría finita óptima para describir el evento de tránsito de s^1 en U^1 de nuestro ejemplo, tiene términos que refieren nominalmente a un estado de cosas, pero no lo describen porque no nos permiten saber exactamente en qué mundo posible son de hecho verdad.

Así, la mente humana que sea óptimamente representada por una teoría finita nominal o descriptivamente, puede ser de hecho que lo único computable sea la teoría, pero no

necesariamente el fenómeno representado. Que el fenómeno sea de hecho una secuencia infinita de estados, tampoco quiere decir que la teoría que lo describe no sea computable. Aunque la causalidad psicológica en un universo así sea de hecho no Turing computable para todas las interpretaciones, ello no quiere decir que una teoría sobre tal causalidad sea necesariamente no-computable también. El fenómeno puede ser, por darme una licencia en la expresión, una hipertarea (un caso de hipercomputación, veremos más adelante qué significa esto), pero satisfacer óptimamente a una teoría *de todo a todo* computable.

Una teoría computabilista de la mente de este tipo diría algo así: aunque la causalidad psicológica sea teóricamente computable, no necesariamente es ella misma un fenómeno computable. Así, la teoría que representa un fenómeno como el que tratamos en este subcapítulo puede ser computable, pero no representa óptimamente al evento. Para representar óptimamente al evento debe contener una oración positiva para cada átomo temporal y/o espacial. Es decir, para que un evento sea óptimamente representable por una teoría de cardinalidad menor a $\aleph_0$, es necesario que la descripción de todos los estados que componen al evento sean finitamente describibles, y que la cantidad de estados que componen al evento sean también finitos. Una representación finita de un evento en U^1 como el traslado de s^1 que vimos en este subcapítulo, no es óptima, sino sub-óptima. La teoría de 5-7 es en este sentido sub-óptima para todo evento en U^1.

3.3.2 Causalidad hipercomputacional

En principio podemos pensar en al menos un tipo de causalidad no computacional, aunque sea también sistemática y mecánica. Pensemos en otro evento dentro del universo U^1 con tiempo y espacio infinitamente divisibles y sin ciertas restricciones físicas como las que tienen los sistemas físicos en universos donde una física finitaria es estrictamente correcta (con límites en la velocidad a la que puede viajar una partícula y donde la suma de vectores de fuerza es el caso, incluso cuando el vector resultante es más grande que la velocidad de la luz, etc.). Consistente con el tipo de causalidad no computacional del capítulo anterior, pero representable sub-óptimamente por teorías computables, tenemos una noción de causalidad hipercomputacional. Dicho en otras palabras, en U^1 es posible realizar tareas infinitas en una cantidad finita de tiempo. La hipercomputabilidad sólo es posible en U^1 y no en universos

donde nuestras leyes físicas macroscópicas además del límite de la velocidad de la luz y límites semejantes, son el caso.

Retomemos el ejemplo del subcapítulo anterior; el traslado de s^1 en U^1 de A a B en tiempo *t*. La teoría $T^{5\text{-}7}$ representa finita, nominal y sub-óptimamente al evento en cuestión. Para que ese evento sea representado óptimamente por alguna teoría, ella debe tener una cantidad de enunciados positivos verdaderos de cardinalidad $\aleph_0$. Si el evento en U^1 es de tal forma que el paso del enunciado 5 al 7 es recursivamente definible, entonces la teoría puede ser además nominalmente verdad, o también, descriptivamente verdad. Si la teoría óptima del evento en U^1 que va del enunciado 5 al 7 no es recursivamente enumerable (como sería por ejemplo la enumeración de todos los valores decimales de π), no es tampoco axiomatizable ni es por ello mismo finitamente expresable, y por ello mismo no es descriptiva tampoco. Llamaremos a una teoría que representa óptimamente a un evento en U^1 y que represente óptimamente en sentido descriptivo TOInfD. Llamaremos a una teoría que representa óptimamente a un evento en U^1 y que represente óptimamente en sentido nominal TOInfN. Si un evento en U^1 satisface TOInfD entonces satisface TOInfN, pero la inversa no se sostiene.

3.4 Observaciones teóricas generales

Dado que la numeración Gödel, necesaria para probar la incompleción de Gödel, hace que cada enunciado de LPO+PA tenga un *nombre numérico* en sentido **por lo menos nominal, aunque no es claro que no lo sea también en sentido descriptivo** el número de Gödel de cada oración de LPO+PA puede ser finitamente descomponible hasta llegar a la oración que él representaba. Esta situación hace que la oración Gödel de LPO+PA sea recursivamente enlistable de hecho, tanto como el número que la representa. En cambio si la numeración Gödel refiriese sólo nominalmente a cada fórmula bien formada de LPO+PA, la derivación haría parecer *ad hoc* la prueba de la oración G respectiva que todo agente o teoría hiciera. **Lo anterior equivale a** decir que un agente dado que tiene razones hipotéticas para *saber que G*, necesita vivir en un universo como U^1, de tiempo y espacio infinitamente divisibles, pero más aún, en el que la causalidad psicológica no es computable.

Lo cierto es que el primer teorema de incompleción de Gödel, usa auto-representación de una manera tal que cada fórmula del sistema, en nuestro caso LPO+PA, es representada de una

forma recursivamente enumerable, pero además, los números de Gödel contienen datos extra sobre la longitud de las fórmulas representadas, y en esa medida de una característica esencial de las derivaciones de una fórmula. No sobra decir que la noción de prueba formal de un cálculo *hilbertiano* es la razón de que las pruebas en un cálculo así tengan una longitud finita. LPO+PA es del estilo de cálculo hilbertiano, y la representación de LPO+PA dentro de la misma teoría conserva esa característica. Así es que la auto-representación de LPO+PA es descriptiva en este sentido también.

La aplicabilidad física y metafísica de los teoremas de incompleción de Gödel parece violar la supuesta *neutralidad tópica* sobre todo de LPO, pero también de PA. Pero tanto LPO como PA aunque no *tratan* acerca de objetos propiamente temporales, ni propiamente espaciales, sí *se hacen* de una manera que es mesurable al menos en teoría. A medir sistemas formales se encarga toda la meta-lógica y para realizar esta labor requiere de átomos a partir de los cuales hacerlo. Tales átomos tienen implicaciones metafísicas; las derivaciones formales finitas, son precisamente, finitas. *Finitud* es una noción metafísica. Para que las pruebas finitas sean finitas, requerimos que haya átomos probatorios, si se quiere, quizá no temporales o quizá no espaciales, pero sí átomos finitamente enumerables de una forma tal que nos permita medir de alguna manera, su finitud. Los números de Gödel de LPO+PA son descripciones definidas de las fórmulas de LPO+AP pero además, son recursivamente producibles y nos ofrecen información acerca de la longitud de las derivaciones correctas de LPO+PA.

Es importante notar pues que los *AGs cognitivos* son aquellos que suponen que el computabilismo sostiene que la mente humana apta para LPO+PA, sólo es verdadera y óptimamente representable por una teoría computable, y deja espacio para la posibilidad de que de hecho el fenómeno de lo mental sea no discreto, quizá continuo. Si es así, para el caso en que un *AG cognitivo* tuviera razón, queda abierta la posibilidad de que por azar tuviéramos la teoría correcta de la mente humana, pero ignoraríamos muchas cosas acerca de si tal teoría es completa y coherente. Con el éxito de este tipo de AG, todo proyecto de investigación psicológica pierde propiedades deseables, no obstante podría haber alguna teoría computable sobre la mente humana que la describa verdadera y óptimamente. En cambio, para un *AG metafísico* el problema es que no **tenemos en principio una** tal teoría, así sino que toda teoría computable de la mente humana sería falsa, porque la mente humana no es una alguna MT. Para el *AG metafísico* no podría haber una teoría computabilista de la mente que sea verdad, y las teorías de las capacidades aritméticas de la mente humana Gödel susceptible serían falsas.

Con el éxito del segundo tipo de AG (metafísico), hay que desechar al computabilismo completamente. Con el éxito del primer tipo de AG (cognitivo) hay que desechar sólo cualquier aspiración a que la psicología sea una ciencia rigurosamente certera y finitaria como es la física incluso en principio.

Conclusiones del capítulo III

En este capítulo aclaramos dos grandes fuentes de confusión entre tesis computacionales. La premisa del AG general que indica que las mentes humanas son MTs, son ambiguas entre dos posibilidades: que la mente sea metafísicamente una MT, lo cual tiene implicaciones metodológicas en particular que deben usarse metodologías computacionales para investigarla, y que la mente humana sea un fenómeno no computacional en sí mismo, pero sí adecuadamente representable como tal. Aunque la metodología computacionalista también podría funcionar para esta segunda posibilidad, el vínculo cognitivo entre lo que arrojaran investigaciones computacionales sobre la mente, y lo que ella realmente es, sería más débil.

Establecimos la relación que hay entre ciertos hechos y las implicaciones infinitas que tienen, haciendo algunas concesiones sencillas de análisis sobre los conceptos involucrados. En el caso de los AGs, por ejemplo, demostrabilidad, conocimiento, proposición, S4 (o introspección positiva en lógica epistémica), etc.

En este capítulo explico la forma en que el funcionalismo es un tipo de computabilismo. Es importante explorar cómo esta igualdad sucede. Dos funciones pueden ser iguales por razón de la extensión, o por razones de la intensión. La igualdad extensional da sentido cómo es que, al menos subjetivamente, dos estados mentales pueden parecer distintos, y no obstante ser el mismo. Pero lo que descartamos aquí es que si dos funciones son extensionalmente iguales, dos teorías o posturas subjetivas puedan estar en lo correcto y no ser la misma función.

En este capítulo he establecido las relaciones que hay entre establecer procesos funcionales, ya sea descriptivos o normativos, teleológicos, o meramente descriptivos. Hasta este momento hemos encontrado que todas las teorías y posturas pueden, al menos a primera vista, llevarse al computabilismo ya sea para describir máquinas, cerebros, o mentes tratando de describir o tratando de actuar de acuerdo a cierta norma.

Discuto también algunos de los problemas que hay con respecto a la igualdad en el terreno de las MTs. Una característica que parece tener la igualdad de implementación entre dos MTs, que pueden ser implementadas por dos sistemas completamente distintos entre sí, por ejemplo, una computadora y un cerebro humano, de modo que dependiendo del análisis y del nivel del análisis computacional sobre tal o cual sistema alegadamente computacional, podríamos tener dos grandes formas en los cuales una teoría computacional verdadera se puede relacionar con el objeto descrito. La primera, describe al objeto de algún modo. De esta clase de relaciones de representación correcta aprendemos cosas del objeto descrito al aprender características de las teorías en cuestión. La otra forma en que una teoría puede representar a un objeto, es por medio de una representación nominal, o no descriptiva. Con esta clase de teorías no aprendemos necesariamente nada del objeto representado. También, estas teorías podrían no ser racionales.

Estas dos formas de describir computacionalmente un objeto son interesantes para los AGs por una razón básicamente: en particular para la representación descriptiva de las teorías computacionales sabríamos, de ellas, que una teoría finitaria verdadera, pero incompleta del objeto descrito, sabemos que es incompleta no por su falsedad, sino por las verdades que no nos dice. En cambio si la representación del objeto computacional es nominal, sabemos con mayor certeza aún que una teoría óptima es incompleta, pero además no sabríamos cómo es que describe al objeto o qué características tiene. Por ejemplo, no sabríamos qué mecanismos causales o cómo es que nuestra teoría es de estas mentes es verdad, o en qué sentido o por qué la mente se comporta computacionalmente.

Si atendemos solamente a teorías computacionales que no son refutables con variaciones sencillas en otros niveles de descripción o, en general, por problemas de descripción del objeto descrito aunque todas las descripciones sean verdaderas, implicaría que existe realmente un sentido en el que dos máquinas de Turing distintas MT1 y MT2, son iguales entre sí, idénticas, aunque una sea implementada por un cerebro y la otra, por una computadora de sílice, una tarde más o una tarde menos que otra. Esto sirve para que tenga sentido cómo es que, por ejemplo, aunque dos personas son capaces de los mismos frutos uno del otro, uno pueda ser más rápido que el otro, y ser, un mismo tipo de "máquina" por ponerlo en estos términos.

Finalmente mostré cómo necesitamos una teoría de la mente humana que no sólo nos muestre cómo es la función que une inputs, con outputs, sino que debe poder mostrarnos de forma

descriptivamente interesante cómo a veces tratamos de establecer con funciones cognitivas ciertos inputs con ciertos outputs, algunas veces quizá de modo que tengamos que explicar la función a través de los outputs. Por ejemplo, si la mente humana es capaz de darnos conocimiento sobre el infinito, pues deberíamos explicar cómo puede darnos un "fruto" como ese, y no sólo que nos da ese fruto, o una etiqueta que refiere no descriptivamente a ese objeto.

En este capítulo también exploramos la naturaleza del vínculo representacional entre una teoría computacional con una mente metafísicamente computacional y el vínculo que hay entre una teoría computacional con una mente posiblemente no computacional, pero sí representable óptimamente como un objeto computacional. Todo esto sirve para evaluar el alcance de las críticas que se hacen a los AGs por la vía de afirmar cualquiera de estas dos posibilidades.

Dicho lo anterior de un modo distinto: el fenómeno que buscamos igualmente, es el de la naturaleza de los estados mentales en sí mismos, y en los efectos que tienen. Sea como sea, buscamos la naturaleza de la causalidad psicológica de estados cognitivos y de otros estados, o bien, de los estados cognitivos mediante los efectos que éstos pueden tener. Ya que dos mentes pueden ser iguales al nivel computacional por cierta descripción computacional de las mismas, es importante tener en claro que no todo nivel de análisis computacional debe ser tomado como un análisis suficientemente bueno. Al mejor análisis computacional lo llamamos óptimo. Suponemos que un análisis computacional es óptimo si siempre que es verdad de un objeto O, y de otro objeto T, ambos objetos O y T tienen las mismas características mentales.

4. Explicación y justificación de las condiciones de solidez de los Argumentos Gödelianos

4.0 Introducción; la clasificación general

En este capítulo se dan algunas razones o condiciones para la evaluación de la fortaleza (solidez y admisibilidad) de varios tipos de AGs. Tres de cuatro tipos adolecen de varios defectos señalados ya en discusiones conocidas, mientras que hay un tipo de AG que no es susceptible a los errores de los demás. Este tipo de AG no obstante no está exento de polémica. En el presente capítulo presentamos tanto los tipos de AGs, como algunos de sus principales defectos, y las virtudes comparativas del tipo de AG que no cae en los vicios en los que sí caen demás. El tipo de AG que tiene mejores probabilidades de ser bueno (sólido) es resistente a las críticas más fuertes contra los AGs. Tales críticas estás contenidas básicamente en el texto de LaForte, Hayes & Ford (1998) mismo que enfatiza ciertos supuestos epistémicos que el éxito de cualquier AG requiere. Argumentaré que tales supuestos epistémicos son menos implausibles de lo que parece, y ciertamente menos implausibles que otros que harían sólido al AG del tipo específico del que hablamos.

En este capítulo vamos a analizar lo que llamo *una familia* especialmente fuerte de AGs que puedo incluir dentro de *cierta* tradición cartesiana, aunque no por lo *dualista*. Hago esta alusión histórica meramente para ubicar posibles defensores de peso en la tradición filosófica, aunque este vínculo siempre importa riesgos de interpretaciones equivocadas. Pidiendo se disculpe ese riesgo por unas líneas, rescato lo que Descartes escribió:

> Porque de hecho, incluso si la idea de sustancia está en mí como resultado del mero hecho de que yo soy una sustancia, la idea de una sustancia infinita no estaría en mí, dado que yo soy finito, a menos que la idea se derive de una sustancia que sea realmente infinita.
>
> Tampoco debería yo pensar que yo percibo el infinito por una idea genuina sino sólo por la negación de una idea finita, tal como yo percibo al reposo y a la oscuridad por la negación del movimiento y la luz; por el contrario, de forma manifiesta a mi entendimiento hay más realidad en la sustancia infinita que en la finita, y que por ello la percepción de lo infinito en mí debe ser de alguna forma previa a la percepción de lo finito: la percepción de Dios, en otras palabras, es anterior a la

> percepción de mí mismo.[42] (Descartes, 1641, pp. 32-33, mis negritas; lo que está alrededor es meramente contexto, y podría decirse que no lo suscribo *in se* de forma tajante y en especial me refiero a su paráfrasis teológica de la última oración: "… en otras palabras…")

En la cita de Descartes parece presuponerse que *la idea del infinito es más real* porque *el infinito es más real* que la mera *negación de lo finito*. Pero esta cita contiene ideas interesantes cuyo espíritu captura parte de *las razones de éxito de esta familia de* AGs: *cierto conocimiento matemático* y lógico –en especial la incompleción de Gödel, no es un fenómeno que se pueda explicar del todo con una *metafísica finitista computacional*. El problema no viene de la parte *computacional in totto*, sino del computabilismo finitista. No obstante lo anterior, la solidez de los AGs no es necesariamente cartesiana en el sentido que suele aplicarse a este término; por el dualismo. Los AGs que resultaren exitosos son completamente compatibles tanto con el monismo como con cualquier dualismo, o *n*-ismo metafísico.

Deshacerse *a priori* del presupuesto infinitsta para tratar de verificar esta clase de computabilismos parece ser una aplicación correcta de *la navaja de Ockham*. El problema de esta fortaleza es que ya no es una buena fortaleza para que la ciencia consiga *la verdad*, por decirlo así, sino para que la ciencia cumpla con *cierta idea de parsimonia*. No podemos preferir un principio epistémico aún en los casos en los que respetarlo nos haría perder la verdad sin incurrir en dogmatismo y sobre todo en falsedad. Para que un principio de parsimonia metafísica funcione, necesitamos antes que todo tener dos explicaciones competidoras igualmente efectivas al explicar, y si tengo razón con los AGs, el grupo de explicaciones computacionales finitistas, sean metafísicas o no, no pueden explicar *cierto conocimiento* lógico y aritmético; así la parsimonia no es un verdadero problema para un AG exitoso. Tales modelos finitistas tampoco podrían explicar conocimiento aritmético no especializado en el modo intuitivo en el que juzgamos que se presenta. Por ejemplo, si alguien está contando montones de naranjas, puede saber que si en el montón A contó 15 y en el montón B contó 17, entonces en la suma de A y B hay 32 naranjas. Si juntos ambos montones y vueltos a contar resultaran ser mayores o menores que 32, es muy difícil pensar que la suma general 15+17=32 tuvo una instancia falsa. Nos parece más *intuitivo juzgar* que cometimos un error de conteo, a pensar que la aritmética que conocemos es falsa. Sopesamos más importante el conocimiento aritmético que la mayor parte de los otros tipos de conocimientos, por decirlo así.

[42] Mi traducción de: "For indeed, even if the idea of substance is in me as a result of the very fact that I am a substance, the idea of an infinite substance would not therefore be in me, since I am finite, unless it derived from some substance that is really infinite. / Nor should I think that I perceive the infinite not by a true idea, but only by negation of the finite, as I perceive rest and darkness by the negation of motion and light; for on the contrary, I manifestly understand that there is more reality in infinite than in finite substance, and that therefore the perception of the infinite in me must be in some way prior to that of the finite: the perception of God, in other words, prior to that of myself." (Descartes, 1641, pp. 32-33.)

En el espíritu de aplicar el principio de la navaja de Ockham, ¿por qué no mantener *la finitud empírica* (finitud actual) que es *más simple* que la *infinitud empírica actual* o que la finitud empírica, pero el misterio de cómo podemos referirnos por ejemplo, al infinito y obtener conclusiones verdaderas de un vínculo que nos es completamente misterioso, al postular cierta clase de fantasma (dualismo finitario-no finitario, racional-no racional)? ¿Qué es más parsimonioso metafísicamente hablando; fantasmas o infinitos actuales? Esta pregunta se responde con elementos ya vertidos antes a lo largo de este trabajo (Capítulos I a III). Aunque el reino de los números fuera accesible a nosotros por virtud de cierto *fantasma etéreo*—alguna forma de *sustancia inmaterial*-, todavía tendríamos que presuponer que ese fantasma etéreo tiene acceso al infinito, y entonces la explicación computabilista, sea finitista o no, seguiría siendo pertinente ahora con respecto al fantasma, o incluso con respecto al *fantasma aunado al cerebro*.

Si alguien pensara en serio que el fantasma no es un objeto de conocimiento natural, no cuantificable de forma finitaria, no ejemplificable o indexable (incluso por medio de artificios como el *dasein* o el *yo-aquí-y-ahora* de la fenomenología continental), tampoco parece que los AGs pierdan aplicabilidad, aunque es obvio que la solidez de los AGs no implicarían *sobrenaturalidad* del fenómeno de la mente humana (aquí he de entender "sobrenatural" como la imposibilidad de que algo sea objeto de investigación científica, sea empírica o con otra metodología sistemática, sea finitaria o infinitaria, cualitativa o cuantitativa, etc.). Incluso es un signo de poca seriedad pasar de la solidez de los AGs a la existencia de fenómenos tradicionalmente considerados *sobrenaturales* y validar con ello a la astrología, la quiromancia, cafemancia, la cábala, y otras artes esotéricas. Es además un asunto irrelevante para esta discusión, me parece, el de que Gödel creyera en la existencia de fantasmas, espectros de la personalidad que sobreviven a la muerte del cuerpo físico: estamos evaluando AGs, incluyendo el de Gödel, pero no únicamente el de Gödel. Bien puede ser que la característica de la *hipercomputacionalidad* de las mentes humanas—o quizá de sólo ciertas mentes humanas, pudiera implicar algún tipo de fantasmagoría, aunque esto igualmente está lejos de los alcances de esta tesis y, parece ser aún así que no hay razones fuertes para considerar sobrenatural a lo no-finitario. Pero aunque la *hipercomputacionalidad* implicara cierta fantasmagoría de la vida mental humana, eso no cualifica por ese hecho que no podamos hacer investigación científica sobre un fenómeno que habría sido natural todo el tiempo, sin que tuviéramos manera de comprobarlo; algo así como una revolución científica *Kuhneana*.

Una observación más al respecto de la aplicabilidad del principio de parsimonia. Aun suponiendo que alguna teoría computacional finitaria de la mente humana *pudiera* explicar todos los fenómenos mentales, puede volverse más complicada incluso en términos de la *navaja de Ockham* que una teoría computacional no-finitaria. Por ejemplo, una persona que suponga que en el universo hay sólo una

cantidad limitada de materia y energía podría estar *a primera vista* en problemas para explicar ciertos fenómenos como, pensemos, los agujeros negros. Así, si los agujeros negros tuvieran de hecho gravedad infinita (como sugieren algunos expertos) que pueda ejercer la fuerza suficiente para evitar que escape la luz, una explicación finitaria de estos fenómenos podría verse forzada a postular que no es gravedad infinita la que tienen los hoyos negros, sino que siendo finita, es, por ejemplo, muy grande y que eventualmente desembocaría toda es energía absorbida en algún lugar lejano, quizá prácticamente invisible para nuestra comunidad científica. Esta sería una consecuencia altamente indeseable para una explicación finitaria en términos de la parsimonia o, incluso, en términos de otros principios epistemológicos y metafísicos.

A continuación entonces expongo los cuatro tipos de AGs emergidos del total de combinaciones posibles entre *Turing-testeabilidad* y *computabilismo metafísico*; i) AG *Turing-testeable* y *computabilismo metafísico*, ii) AG *no-Turing-testeable* y *computabilismo metafísico*, iii) AG *Turing-testeable* y *epistémico* (no necesariamente metafísico), y finalmente iv) AG *no-Turing-testeable* y *epistémico* (*ídem*). En capítulos anteriores, vimos ya a qué se refiere con detalle que un AG sea *Turing-testeable* o no, y qué significa que sea *metafísico* o *meramente epistémico*. Para los efectos de esta tesis, un AG que sea *epistémico* considera la posibilidad teórica al menos de que el AG sea no metafísico. La relación entre presupuestos metafísicos del computabilismo, y su negación no nos compromete *ipso facto rationis* con la alternativa de los AGs epistémicos. Estamos suponiendo aquí que los éxitos científicos futuros y pasados de las ciencias naturales (*Naturwissenschaften* o *natural sciencies*) o en *ciencias del espíritu* (*Geisteswissenschaften* o, de forma parecida *moral sciences*) que presuponen el modelo computabilista o uno finitario isomórfico hacen del finitismo una característica por lo menos *metodológica* y, en ese sentido, una característica también *de las teorías* aunque no necesariamente de los *fenómenos* a los que las teorías describen o modelan.

4.1 AGs metafísico y *Turing* testeables* (tipo I)

El primer tipo de AG toma la premisa computabilista como indicando la igualdad metafísica entre la mente humana y una MT. Llamaré a este argumento gödeliano, "*AG Turing metafísico*". No es que la MT *represente* a una mente, sino que la mente humana *es* cierta MT, tal como cualquiera otra. También, esta clase de argumentos supone que es posible TTear las características mentales. Así, si una mente digamos, es *capaz de aritmética* esta aptitud, para quien ofrece un AG del primer tipo, implica que puede ser *Turing testeada* tal como se entiende en esta tesis (en el futuro, *"Turing* testeada*").

Si el mundo presupuesto por el tipo I de AGs (sección 4.1 de esta tesis) es el mundo actual, aún podríamos preguntarnos si la evidencia para probar posesión de características mentales por medio del TT es parcial o, si la posesión de facultades mentales es en ciertas ocasiones la implementación parcial de funciones. Por decirlo así, parece que habría dos subtipos de AGs *Turing metafísico*.

En los capítulos anteriores hemos visto ya una serie de distinciones que servirán para explicar el primer tipo de AGs, y así mismo, escogido este tipo por ser el más sencillo de refutar y, de hecho, el más y mejor refutado en la literatura especializada[43]. En términos de este AG, una implicación clara de lo que dice un AG metafísico *Turing* testeable* exitoso es que tanto los estados como la causalidad psicológica no pueden ser consustanciales a la causalidad computacional como estados computacionales. Si la causalidad psicológica fuese consustancialmente computacional, ello implicaría que la veracidad de la oración G no puede ser establecida sólo sobre los recursos conceptuales usuales de PA+LPO, a menos que ya incluyamos *a manera de* axioma a cierta oración G, algo así como una forma de innatismo de las matemáticas, consecuencia que nos parece altamente implausible. Un sistema Turing-computacional Gödel susceptible instanciado físicamente y consistente no podría establecer la veracidad de la oración G o la oración G particular del sistema *so pena* de la venganza del mentiroso: en este caso, un predicado de demostrabilidad que incluyera, como axioma extra, la oración G, para con ello poder construir una oración G', indemostrable ella también, y así al infinito.

Hay una observación que debe hacerse. Dado que las MTs son objetos abstractos, por ciertas leyes de entropía no pueden estar dentro de un universo material sin que se sometan a desgaste y corrosión. ¡Que una idea abstracta decaiga físicamente, parece inadmisible! Sostener que la mente humana, o la parte cognitiva de la mente humana es consustancial a por lo menos alguna MT, parece comprometernos de entrada con cierta forma de dualismo que dado que la mente humana es consustancial a una MT, y que estas no pueden estar en un universo físico finito *ex hypothesi* como el nuestro, pues la pura presencia de la mente humana sería prueba de la existencia de una sustancia no física, una forma de dualismo. Para la filosofía es un movimiento dialógico peligrosamente cercano a ser argumentativamente falaz (la falacia verbal o algún tipo de error retórico del estilo de *petición de principio*). Claro, en especial en el

[43] La mejor formulación de este tipo de AG, que es *conductual* y *Turing* testable*, es la de Bringsjord (1992).

sentido en el que confundiríamos inferencia con prueba, o argumento válido con argumento sólido. Cuando un computabilista sostiene razonablemente que por lo menos las funciones cognitivas de las mentes humanas son MTs, con ello tiende a colapsar el estatus metafísico de tales funciones con el de las MT solamente, y no extrae más implicaciones ni sobre el mundo real ni sobre la naturaleza de los objetos abstractos. Así que los supuestos de los AG metafísicos concluyen *antes de la discusión* la falsedad del computabilismo. El tema de la naturaleza metafísica de los objetos abstractos es polémico, y así, si las MTs son objetos abstractos, por lo menos parte de su estatus metafísico está *bajo fuego –filosófico– intenso* de la misma forma en que lo está el dualismo o cualquier metafísica inconsistente con la metafísica subyacente a la física contemporánea.

Quienes sostienen un AG *Turing* testeable metafísico* en esta sección, sin duda tienen una ventaja a su favor: las conclusiones a las que llega parecen tener un alcance mayor del que tiene *prima facie* el AG no *Turing* testeable*. La conclusión del tipo más débil de *AG*, el que es *Turing* testeable* y *metafísico*, tampoco es contraria a las llamadas tesis de la identidad (entre cerebros y mentes humanas). Sin embargo, el éxito de los AGs metafísicos sí nos comprometería con el abandono de cualquier modelo que suponga que lo mental puede ser reducido óptima y completamente a lo digital, sean teorías, experimentos o interpretación de evidencia.

Sobre este tema Maudlin (1996) apunta:

> 1.3 La estrategia general para conectar los resultados de Gödel a la física tendría que mostrar que ciertos movimientos actuales de los cuerpos no puede ser acomodado dentro de teorías físicas de cierto tipo, por razones de principio. De la misma forma como el análisis puede mostrar que el comportamiento físico de los planetas cuyos procesos de traslación (órbita) no puede ser explicado por la teoría gravitacional newtoniana, así Penrose parece afirmar que toda la física clásica y cuántica (tal como una clase grande de extensiones posibles o enmiendas de esas teorías) no pueden explicar los movimientos físicos de ciertos cuerpos físicos conocidos: los cuerpos de los matemáticos humanos. ¿Cómo, en detalle, puede ser hecha esta conexión entre un teorema matemático y una acción física?[44]

[44] Mi traducción de: "1.3 The overall strategy for connecting Gödel's result to physics would have to be to show that some actual motion of bodies cannot in principle be accommodated within a physical theory of a certain kind. Just as analysis can show that the physical behavior of planets whose orbits precess (*sic*) cannot be accounted for by Newtonian gravitational theory, so Penrose seems to claim that all of classical and quantum physics (as well as a large class of possible extensions or emendations of those theories) cannot account for the physical motions of

Así, esta clase de AGs aunque es clara, es fácil de problematizar. Un AG de este tipo, es cuando menos una sugerencia débil de que hay un fenómeno mental inexplicable físicamente, pero no apunta específicamente a ese fenómeno: una aportación débil a la discusión. No obstante es clara porque bastaría con especificar qué *movimiento* constituye esta excepción a dichas teorías físicas, y por ello, las cuestiona o bien, de plano las refuta. (La escritura de un teorema, entender un teorema, entender la verdad de un teorema, imprimir un teorema, identificar teoremicidad, etc.) Es problemática, porque es difícil establecer un vínculo entre el movimiento físico (Turing testeable) y la fuente de ese movimiento físico como un evento no computacional.

Dependiendo de qué aspectos específicos caracterizan al AG propuesto, es el movimiento corporal o estado mental que se consideraría como una excepción al modelo computabilista. Algunos de los candidatos más fuertes son: imprimir el teorema (la impresora de Smullyan, 2001) o *responder sí o no a la pregunta de si la oración G relevante es teorema* (Turing, 1950), *inferir –con pensamiento ontológicamente computacional— el teorema de incompleción* (véase p.ej. Kirk, 1986), y uno de los más sutiles es *saber la verdad del teorema* (Penrose, 1994, por lo menos, McCall, 1999, Haifman, 2000, Gödel, 1951, McCall, 1999, y es presupuesto por las críticas de LaForte, Hayes & Ford, 1998). La idea básica es que ya que este movimiento o estado mental (comprensión, escritura o impresión de teorema, pensar o comprender la validez del teorema) es imposible de acuerdo al computabilismo y algunos supuestos más, entonces no sabríamos que el teorema de incompleción es un teorema o si lo supiéramos, no podríamos explicar sin aludir a eventos misteriosos cómo sabemos que lo es. Quizá lo sospecharíamos, pero *saberlo*, no. ¿Cómo podríamos estar en esta situación? *Ex hypothesi* sería imposible metafísicamente. Así, o bien la presuposición computabilista metafísica es falsa (no vivimos en un mundo como lo presupone el computabilismo) o bien *el movimiento corporal o el estado mental* (impresión, comprensión, inferencia, etc.) no pueden ser el caso.

Veamos, ¿qué evidencia podemos admitir para refutar ya sea que el teorema sea impreso, que haya sido, sea o será comprendido, que sea, haya sido o será inferido, o algún otro estado

---

some known physical bodies: those of human mathematicians. How, in detail, could this connection between a mathematical theorem and physical action possibly be made?"

mental al respecto de su veracidad? Primero, concluir que esta clase de movimientos o *estados* físicos de impresión, comprensión o inferencia, *inter alia*, de la veracidad de la oración G relevante es imposible requiere tanto de la veracidad de la tesis computabilista como de la indemostrabilidad o no-imprimibilidad de la *oración G* en las condiciones en las que se demostró. Dado que la verdad de G sería automáticamente tenida si admitimos que es imposible probarla de forma *Turing-testeable* (en particular finitariamente o de forma relacional finitaria input-output), sólo quedaría, dentro de la disyunción de si el computabilismo es el caso o no, admitir que no vivimos en un mundo computabilista o bien no sabríamos si G es verdad, ni sabríamos todo lo que la verdad de G implica.

La impresión del teorema como candidato a *movimiento imposible* es difícil de poner en duda y tiene sólo un interés limitado. Por ejemplo, pensemos en un AG *metafísico* basado en Smullyan (2001), en el que suponemos que tenemos una impresora que puede imprimir todas las secuencias de comandos verdaderos por medio de los cuales se le solicita impresión. Así la orden análoga a la oración G, "no imprimirás este comando", no sería físicamente imprimible, dado el supuesto de que la impresora sólo imprime verdades. Sería imprimible suponiendo que la impresora funciona bien (que sigue siendo Gödel-susceptible) solamente en el caso de que sea falso el supuesto de que imprime sólo verdades. La ausencia de impresión, para este análogo a la oración G, muestra, según quienes sostengan AGs de este tipo, que las computadoras que son Gödel-susceptibles, son Gödel-limitadas. Pero pongamos una excepción a la regla de que imprime sólo verdades, y pensemos que, fuera de la oración G relevante, todo lo demás que imprime la máquina, son verdades: ¿Demostraría esto que la impresora superó el límite de Gödel? Por supuesto que no, porque habría impreso una falsedad. La computadora sería Gödel ilimitada más bien si no imprime falsedades, y además, pudiera darse cuenta de esto de algún modo. Por eso, proponer un AG en la línea de la impresión de teoremas y así, es una estrategia que pocos considerarían seria. Pero, otros sí la consideran seria no ya imprimiendo teoremas, sino robots humanoides contestando preguntas como: "Mueve la cabeza de arriba abajo si reconoces la verdad de la oración G relevante." Una robot tendría que saber que no podría mover afirmativamente la cabeza en ese caso, so pena de volverse inconsistente y por ello mismo, poco interesante para refutar el computabilismo.

Los seres humanos, en cambio, no estarían asimismo limitados según quienes sostienen esta clase de argumentos. Un ser humano impresor, un escriba, que se guiara por las mismas y exactamente las mismas reglas que la impresora no podría imprimir la análoga oración G, a

menos que violara alguna de las reglas. Si esto es así, entonces, nuestro escriba o probador de teoremas no estaría Gödel limitado, pero tampoco sería Gödel-susceptible por su inconsistencia. Según este tipo de AGs, tanto humanos como TMs que son Gödel-susceptibles, son también Gödel-limitados.

Alguien podría agregar que la mera impresión de la oración G no es físicamente imposible y que allí donde la impresora y el escriba se ven limitados, el escriba puede no obstante notar que hay una verdad que no puede escribir, pero quizá sí gritar en un arrebato de estupefacción, siguiendo las reglas de la impresora, mientras que la impresora no estaría en ese supuesto. Esto sería un cambio de juego o un cambio de tipo de AG, obviamente y saldría de los supuestos ya sea del computabilismo metafísico, o bien, de la Turing* testeabilidad.

Las computadoras pueden imprimir la oración G, los estudiantes de lógica pueden escribir el teorema en sus libretas o pueden incluso decirlo en voz alta en alguna clase o en una plática informal. Si queremos ser caritativos con quienes sostienen esta clase de AGs, lo que es físicamente imposible en todo caso es algo mucho más sutil: *cierta característica mental*, por ejemplo, cierta comprensión del teorema y sus implicaciones, la patencia subjetiva de su veracidad, o la certeza de que el *impresor-escritor* que lo imprime o escribe es, después de todo no es Gödel-susceptible. Si estos AGs son *TTeables*, entonces tenemos que admitir que hay algo extraño en el estado cognitivo o en nuestra impresión del teorema de incompleción de tal forma que lo escribimos, pero no lo comprendemos, lo comprendemos pero no aplica a nosotros, aplica a nosotros y no lo podemos saber, en cualquier combinación de situaciones posibles: sabemos acaso una oración G tan general que no aplica de forma concreta a nosotros mismos, sabemos una oración G muy alejada de la que nos limita a nosotros mismos, no sabemos nada, simplemente creemos que sabemos la oración G.

El ejemplo típico de un AG tipo I se presenta en Bringsjord (What Robots Can and Can't Be) y Penrose (TSOM), y caracterizado en Maudlin (1996).

4.2 AGs cognitivos y *Turing* testeables* (tipo II)

Tenemos otra posibilidad: vivimos en un mundo que quizá no sea computacional metafísicamente hablando pero es *Turing testeable*. El énfasis aquí está en el "quizá", ya que

indica no la probabilidad de que el mundo sea no computacional metafísicamente hablando sino que no podemos saberlo con algún grado alto de certidumbre, seguridad o garantía. En esta sección voy a refutar esta idea; si el mundo en el que vivimos puede ser no-computacional, podría ser Turing-testeable solamente si esa posibilidad es lógica, pero no metafísica. Dicho de otro modo, un mundo es *Turing testeable* solamente si es metafísicamente computacional. Podría ser el caso que en cierto nivel de subdesarrollo científico, alguna teoría computacional describa bien porciones del mundo que son computacionales, pero de ahí no podríamos generalizar la consecuencia metafísica de que como, por decirlo así, la física newtoniana es computacional y describe bien ciertos segmentos de la realidad, entonces todo el mundo es computacional. De ser este el caso, tendríamos que contentarnos con teorías del mundo que son correctas con reservas; no podríamos aspirar a explicar todos los fenómenos naturales, y en especial los frutos de las capacidades cognitivas de nuestros cerebros, pero aun así disfrutar de estos frutos. Esto nos orillaría a vivir en cierta clase de conformismo científico notoriamente acrítico; *el mundo tiene misterios, como son nuestras capacidades cognitivas, y preferimos no tratar de profundizar más allá*. Esto es parecido a estar contentos con el hecho de que le exigimos a los físicos un nivel de rigor altísimo en sus predicciones y explicaciones, y a los psicólogos un nivel mucho más bajo aún en los casos en los que no haya obstáculos éticos o de recursos de cálculo para conseguir un mismo nivel de rigor, claro, si es que el computabilista tuviera la razón.

Me explico. Esta clase de AGs tratan de refutar la tesis de que las mentes humanas sean óptimamente representables por medio una teoría computable, y que de esto nos podemos percatar por medio de cierta evidencia *Turing testeable*; por ejemplo, responder a preguntas, observar irrestrictamente el cerebro de los matemáticos, programar a una computadora con todas las habilidades metafísicamente posibles para la aritmética y la lógica y preguntarle a ella si la oración G es el caso o no, etc.

El éxito de los AGs metafísicos y *Turing testeables*, además de dadas ciertas condiciones de epistémica, ellos implican el éxito de los AGs cognitivos *Turing testeables*. La inversa no se sostiene para esos mismos casos. Es perfectamente posible que las mentes humanas sean sistemas cognitivos que por razones contingentes se comporten de forma TT incluso en el caso de que tener cierta evidencia al contrario.

Los AGs cognitivos y *Turing testeables* son interesantes porque son metafísicamente más débiles y admiten que si bien metafísicamente las capacidades cognitivas de las personas pueden ser no computacionales, ellas se comportan así por *alguna razón* de tal forma que podemos tener ciencia rigurosa y computabilista sobre tales capacidades, aunque esa ciencia tenga límites de principio. Por ejemplo, que siempre tendremos misterios al investigar las mentes humanas usando como metodología el computabilismo. También hay una pérdida más: con esta forma de computabilismo perdemos de entrada sin argumento independiente de cuál es *esa razón*, la idea intuitiva de que el objeto conocido se comporta de acuerdo a nuestras teorías no de forma azarosa. Dicho de otro modo: cambiamos la posibilidad de dar explicación racional (computacional, después de todo) de nuestro aparato cognitivo, por la imposibilidad de dar explicación racional (computacional) a much $\forall x(\Phi(e,x)\downarrow \Rightarrow \Phi(x,x)\uparrow)$ os de los frutos de tal aparato cognitivo. Esto no parece un buen trato, en especial si entendemos que dar una explicación racional de nuestros aparatos cognitivos implica poder explicar todo lo que *"pasa o sale através de ellos*".

4.3 Críticas a los AGs *Turing testeables*

La crítica más devastadora a esta clase de AGs es la de Hayes, LaForte & Ford (1998). Esta crítica se ha hecho al AG *Turing testeable* y básicamente consiste en poner en duda que existen los estados cognitivos al respecto de la oración G aplicable o relevante para el sujeto cognoscente. Si uno admite de entrada el principio de que todo evento, para existir debe suceder dentro de los linderos conceptuales de la *Turing-computabilidad*, entonces parece sencillo simplemente afirmar que cierto estado que parece no-computacional en realidad no puede existir, de modo que no es un estado genuino. En particular, dudar de que siquiera sepamos que somos coherentes. Pero esto sólo es devastador si de entrada uno admite la veracidad del computabilismo. Hayes, LaForte y Ford (1998) no lo ponen exactamente así, sino en términos más generales:

> One can show quite rigorously that Penrose's notion of what it is to know oneself to be sound [i.e. Gödel-unlimited] cannot itself be sound. The computable analogue of believing that the procedure indexed by *e* is sound is, in the context of the diagonal arguments we have been considering, to use the formula to make decisions by merely deducing first-order consequences from this formula. This procedure defines a

> computable function, *f* which takes a program index *e* to the index *f*(*e*) for a program which uses first-order logic and the soundness assertion for procedure *e* to make decisions about which programs halt on their own arguments. This is a more precise description of what Penrose refers to as "automating Gödelization". We might think of *f* as being the formal analogue of understanding the meaning of "*e* is sound", since it represents the notion of asserting what follows from *e*'s soundness. So, for a sound *e*, *f*(*e*) gives a sound procedure, while for an unsound *e*, *f*(*e*) gives an unsound procedure. If we look at *f* in this fashion, we can get a clear idea of what it would mean to use one's understanding of the meaning of soundness in one's reasoning. (*Op*. *Cit*., pp. 271-272)

Para ilustrar este fenómeno, Hayes, LaForte y Ford (1998) usan una analogía conveniente y elocuente incluso a quienes no suscribimos la generalidad de sus refutaciones. Al creer con fuerza epistémica alta (certeza, seguridad, garantía, etc.) que podemos identificar todos los procesos de razonamiento propios y su confiabilidad cometeríamos el siguiente error:

> "… [C]onsider the fact that anyone can stand on anyone else's shoulders, but no one can stand on their own shoulders." (*ídem* p. 271)

Nadie puede dar ciertas validaciones a sus recursos epistémicos, usando esos mismos recursos epistémicos por decirlo de algún modo: no sabemos si somos consistentes, y esto mismo nos hace no sólo Gödel-limitados, sino no Gödel-susceptibles para empezar. Esta clase de errores según la cual aún si admitiésemos que cierto agente racional fuera Gödel-susceptible, para que además esté Gödel-limitado es necesaria una pieza de conocimiento imposible de conseguir: prueba (proceso para obtención de creencia con una fuerza epistémica alta) de que uno mismo *es* consistente. La tesis de que uno mismo *es* consistente tiene asimismo dos lecturas que quedan debidamente contempladas en las dos subsecciones anteriores, la primera es que *nuestras facultades cognitivas son consistentes* en sentido metafísico, y la otra, uno lo es en sentido meramente cognitivo. No obstante lo anterior, es altamente dudoso que haya algo, en sentido metafísico que sea contradictorio en sí mismo. El segundo sentido de *'ser consistente cognitivamente'* es un tanto cuanto más débil que el primero, y por ello parece más razonable ya que permitiría algo así como la posibilidad de poseer una teoría verdadera y coherente sobre

nuestras propias capacidades racionales[45], pero perderíamos garantías epistémicas altas como son; certeza, garantía o seguridad, ya que éstas garantías requieren, para ser tenidas, de establecer claramente coherencia. Parece que podemos suponer que al ser finitas nuestras facultades mentales, ellas son consistentes dados todos los conjuntos finitos de *inputs* a los que puede enfrentarse aunque nunca lo podríamos saber con certidumbre, por ejemplo, si estas mismas facultades tuvieran que juzgar conjuntos no-finitos de evidencia o casos: por ejemplo, podemos considerar una cantidad finita de numerales de los números naturales e inferir que hay al menos un numeral que es el más grande de todos, pero esta misma operación no la podemos hacer cuando llevamos el conjunto a una cantidad infinita de numerales. Mientras que el establecimiento de la consistencia en sentido metafísico está mediada, parece obvio, por el tipo de teoría y la justificación que tengamos al respecto del objeto o característica que son nuestras facultades cognitivas.

Si admitimos que la evidencia que puede probar o refutar los AGs sólo es de tipo computabilista, entonces o bien no sabríamos si nuestras facultades cognitivas son consistentes, lo que nos encierra dentro de la Gödel-limitación o bien lo sabríamos a costa de concluir que nuestras facultades cognitivas son, metafísicamente inconsistentes, lo cual no nos deja en claro que seamos Gödel-ilimitados en ningún sentido epistémicamente valioso para nuestras facultades cognitivas. En ambos casos, nuestras facultades cognitivas resultan severamente lesionadas (en un sentido epistémico y metafísico-epistémico).

Pero debemos tener en cuenta cuál es el objetivo final de los AGs propuestos: refutar el computabilismo. Así, no sólo es completamente razonable poner en suspenso el presupuesto computabilista, sino que es argumentativamente necesario, por mor de la discusión: evitar en la medida de lo posible incurrir en petición de principio. Los AGs que transigen con el presupuesto computabilista de forma irrestricta, son altamente ingenuos y por ello son fácilmente refutables por una exposición limpia y rigurosa como la que hacen Hayes, LaForte y Ford (1998). Es por ello que no profundizaré más en este asunto ni en esta clase de AGs que tienen tremendos vicios de origen.

4.4 AGs no Turing testeables

[45] Kirk (1986).

La evidencia que no se obtiene usando TTs, por más ingeniosos que sean los *tests* propuestos, es especialmente difícil de clarificar. En el capítulo 2 de esta tesis escogí definir la evidencia que no es *Turing testeable* como evidencia subjetiva, por la razón de que no es simplemente cuantificable, no siempre es relacionalmente contrastable (input-output, siendo un caso especial de relación de contraste), no presupone estados finitos, no al menos en el nivel óptimo de descripción de un proceso, y tampoco está casada con la noción de que un juez, experto o no, una comunidad de jueces expertos o no, puedan determinarlo usando cualquier forma de TT. La evidencia no es exactamente lo mismo que el hecho, y por ello es que la evidencia subjetiva es un poco más débil que los estados subjetivos en cuanto a qué tesis se pueden fundamentar con ella. No es lo mismo mi estado subjetivo de que veo una mano, al hecho de que alguien me diga que tiene un estado subjetivo de creer ver una mano. La última, es evidencia subjetiva, la primera, es algo más que mera evidencia subjetiva. Aunque hay una manera trivial de reducir los estados subjetivos como portadores o actos en los que se tiene evidencia subjetiva. Esta noción trivial depende de una operación lógica de segundo orden:

La noción de *evidencia subjetiva* no implica lógicamente la imposibilidad de alguna reducción objetiva. La noción de *evidencia subjetiva* tampoco debe ser entendida de tal forma que implique lógicamente que los *estados subjetivos* tengan a fortiori alguna reducción objetiva. Ambas posibilidades son irrelevantes para los efectos de esta tesis, aunque haya una familia de reducciones objetivas que sí deberían quedar canceladas y ellas son las reducciones computabilistas. Al final del día la evidencia subjetiva puede equipararse correctamente con otros tipos de evidencia, por ejemplo evidencia TT.

La evidencia TTeable sirve en especial para justificar tesis *sintéticas*, ya que las analíticas son demostrables también de forma no TTeable y para teoremas como el de incompleción, justamente lo que el teorema indica es que no es demostrable con esquemas como este. Las relaciones computacionales en estados mentales, por ejemplo, establecen relaciones *funcionales* del tipo:

Input p, output q
$p \rightarrow^{comp} q$

Cuando la relación computacional es trivial, analítica o necesaria[46], tendríamos relaciones del siguiente tipo:

Input p, output p
$p \rightarrow^{comp} p$

Si la relación computacional es un tipo de *condicional* dentro de los márgenes del condiconal material, tendríamos que las tesis computacionales triviales son tautologías o conjuntos de tautologías en última instancia. En el sentido metafísico (y no sólo cognitivo) de la tesis computabilista son poco informativas; no podemos tener estados computabilistas contradictorios en términos metafísicos. Si la lectura de la tesis computabilista es la cognitivista, tendríamos que ver en qué sentido podría ser falsa, qué clase de evidencia bastaría para falsear la tesis computabilista. En ambas lecturas parece exagerado circunscribirnos al esquema computabilista de prueba. Un TT que nos pueda mostrar cómo es que sabemos la oración G relevante y que pueda ser falseado es difícil de imaginar. Ahora, si pensamos que la relación "$\rightarrow^{comp}$" no es en sí misma analítica (como todos los teoremas de PA+LPO podrían no serlo, en *cierto sentido, usando distintos conjuntos de axiomas, por ejemplo*), ¿qué clase de evidencia nos muestra que PA+LPO es completo y correcto? Usar los propios medios y requisitos de PA+LPO nos lleva de nuevo al contexto de incertidumbre al que nos lleva la aspiración del programa de Hilbert, PA y LPO. Para evitar este círculo, suspendamos el esquema de TT ya sea porque exige que las verdades necesarias de los sistemas formales sean meras trivialidades o bien porque nos circunscribiría a la incertidumbre que en general imprime la incompleción de Gödel a PA+LPO, o al Turing-computabilismo, que sea isomórfico a PA+LPO.

¿Cómo podemos mostrar entonces sin métodos de análisis funcionales (Turing computables, Turing Testeables) las condiciones de solidez de algún AG? En capítulos previos hemos admitido que ciertas afirmaciones pueden ser verdad *sin justificación* necesariamente finitista o funcional; verdades apodícticas. Una verdad apodíctica se define como una afirmación cuya verdad podemos observar por el mero hecho de pensarla, sin pasos previos, sólo usando la intuición. Las verdades apodícticas no necesariamente son axiomáticas, ya que las verdades axiomáticas se definen con relación al rol que juegan en sistemas formales, muchos de los

[46] La relación computacional es necesaria *de re*, y no de dicto ya que *de dicto*, las relaciones causales cuya causa y efecto son el mismo hecho o partes del mismo hecho, pueden ser sintéticas

cuales pueden ser computables y así mismo Turing Testeables. Sin mencionar además que el fenómeno de la revancha del mentiroso surge para el acercamiento de tratar a las verdades apodícticas como los axiomas. Pensemos sin justificación por el momento que las verdades axiomáticas no son necesariamente verdades evidentes a nuestra consciencia o intuición[47], pero que las verdades evidentes pueden conocerse sin saber qué rol inferencial juegan en el sistema o sobre el modelo al que se refieren. Si lo anterior es el caso podemos entender claramente que una capacidad cognitiva no funcional, podría darnos acceso a verdades que no son computacionales por razones de principio. Al ser esto al menos consistente, tenemos que considerarlo una posibilidad de explicación del fenómeno de la completud no computacional de un sistema dado.

Otra forma en la que podemos considerar una tesis no idónea para TT consistiría en la capacidad de realizar tareas superiores al límite de Turing, pero no sólo esto, sino también, que no puedan ser reconstruidas como meramente Turing computables pero no finitarias, por ejemplo, al echar mano de recursos infinitos como memoria infinita, tiempo ilimitado para realizar una tarea, aceleración ilimitada en la ejecución de tareas, etc. No toda extensión al infinito de una tarea computable bastaría para explicar el estado cognitivo de *saber que G*.

De esta forma entenderemos como afirmación no TT un tipo de afirmación que por razones necesarias no puede ser TT, so pena de incurrir en alguna contradicción, como es el caso, obviamente, de las oraciones G, aunque no necesariamente sólo de oraciones G o semejantes. La explicación o la justificación de la oración G para un AG no TT, es una reconstrucción psicológica o cognitiva de la consciencia o de un proceso no computable, o bien una reconstrucción conceptual (epistemológicamente normado) del proceso de demostración.

Sobre todo la reconstrucción en términos científicos, psicológicos o cognitivos, puede ser contingentemente inaccesible por razones de varios tipos, o puede no ser generalizable a todas las psicologías; después de todo, parece cierto, aunque altamente dudoso, que no todas las personas son Gödel-susceptibles para empezar. También puede haber algo de verdad al respecto de la inconmesurabilidad de distintas experiencias conscientes, de tal forma que no

[47] "Apodíctico" viene del griego antiguo "ἀποδεικτικός", que significa '*demostrable*', 'incondicionalmente cierto', pero se ha usado para indicar '*auto-evidencia desde la perspectiva de quien juzga*', '*verdades subjetivas necesarias*.' Así, tenemos axiomas difíciles de verificar por medio del '*ojo de la mente*' o apelando a métodos subjetivos. Por ejemplo el axioma de las paralelas en geometría euclidiana, pero también el conocido como *axioma de introspección positiva* y leyes lógicas válidas como la ley del borracho, la identidad entre condicional material y la disyunción incluyente, y el primer axioma de la lógica proposicional de Łukasiewicz.

pudieran generalizarse de forma alguna las vidas conscientes de las distintas mentes y pensando en que ello fuera requisito necesario para la tarea de una psicología o ciencia cognitiva objetiva. No obstante esto, una explicación epistemológica puede sernos suficientemente accesible o bosquejable por lo menos a manera de proyecto de investigación. Ambas, la reconstrucción psicológica tanto como la epistémica, sea cualitativa o no, tendrían que coincidir aunque sea de forma gruesa en alguna teoría con algunas concesiones de precisión.

Quienes han obtenido la conclusión de que hay estados cognitivos *no Turing testeables* a partir de su evaluación de AGs específicamente son Storrs McCall (1999) y Haim Gaifman (2000). Gaifman (2000) usa su evaluación positiva de ciertos AGs como elementos para justificar cierta forma de *monismo anómalo*[48]. Estos son sólo dos ejemplos, pero podría haber otros AGs justamente de esta clase. En las subsecciones siguientes, veremos mejor caracterizados dos ejemplos de AGs no *Turing* testeables*, el de tipo metafísico y el de tipo cognitivo.

4.4.1 AGs metafísico y no Turing* testeables (tipo III)

Hay muchas formas en las que una propiedad, por ejemplo, la computacionalidad de un sistema cognitivo, puede determinarse sin que algún TT sea epistémicamente idóneo para ese efecto, sino algún método sin ese presupuesto. Tales propiedades pueden ser reales, es decir, pueden estar instanciadas en este mundo, por algún sistema sin que nosotros sepamos de ello. Lo anterior, la afirmación de que estas propiedades pueden ser reales, me compromete con la idea de que las propiedades, que aquí apenas he caracterizado, son lógica y metafísicamente coherentes. El *Test* de Turing* o la TTeabilidad sea metafísica o metodológica, presupone una serie de cosas de la propiedad que es TT: mesurabilidad al menos idealmente, finitud, su expresión en el esquema *input-output*, un agente artificial o un objeto del cual hipotéticamente se dice que posee cierta propiedad, cierto velo cognitivo que no predisponga a juez alguno y un juez que puede o no ser experto pero que tenga al menos cierta idea de lo que busca juzgar. La ausencia de TTeabilidad de la que aquí hablamos no tiene nada que ver con las últimas

[48] Aunque no haya un consenso sobre lo que es el monismo anómalo, podemos tomar provisionalmente la definición que de esta posición hace la *Stanford Encyclopaedia of Philosophy*, en el artículo '*Anomalous Monism*', al caracterizar a esta postura como una posición en la que se defiende que la psicología como ciencia no puede aspirar a postular leyes indefectibles como hace la física. Esta *debilidad* de las leyes de la psicología podría ser un asunto contingente o bien una mezcla de problemas de cantidad y calidad de información objeto de la psicología además vicios de racionalidad (cfr. Yalowitz, Steven, 2012).

características, sino en especial con la negación de alguna de las tres primeras: mesurabilidad (contabilidad), finitud y formulación *input-output*. Así, una propiedad real que no sea TTeable, es o bien inmesurable, infinita o no puede ser expresada como una relación *input-output*.

Estos AGs tratan de refutar la tesis de que las mentes humanas sean alguna MT o sean una *máquina universal de Turing*, aunque a la vez tales AGs no aceptan que el TT en ninguna de sus variantes sea una buena forma de darse cuenta de esto de forma plena o justa.

Un ejemplo de esta clase de argumentos es el que viene McCall (1999). Storrs McCall afirma (mi traducción)[49]:

> Una máquina de Turing puede *saber* aquello que puede probar, esto es, aquello que deduce de sus axiomas usando bien-definidas reglas de inferencia. Su base de datos axiomática podría ser grande, y sus reglas de inferencia pueden ser eficients y poderosas. Pero para una máquina de Turing, *cognoscibilidad = demostrabilidad*. Cuando la demostrabilidad y la verdad se separan, como es el caso de los teoremas de incompletud de Gödel, una máquina puede habérselas con una cosa, pero no con ambas. (McCall, 1999, p. 525)

Al final de su texto explorando la posibilidad de programar a alguna MT para lidiar con las verdades del sistema más que con sus teoremas, y habiendo repasado varias estrategias y encontrándolas él mismo inadecuadas, concluye (mi traducción):

> Ahora ¿qué de máquinas más poderosas, equipadas con bases de datos axiomáticas más poderosas y más capaces de generar grandes conjuntos de teoremas? ¿Qué pasaría con máquinas equipadas con un operador T (por *truth*)? ¿Qué pasa con máquinas que tienen enormes bases de datos? Es difícil ver por qué la situación, para ellos, es siquiera diferente de aquello que es una máquina PA. Aún quedan dos categorías de teoremas y no-teoremas, y queda aún para cada grupo, una oración análoga a A10 [i.e. un teorema que dice que: "si PA es consistente, entonces su oración de Gödel G no es demostrable"]. Esta oración no es demostrable pero verdadera. El talón de Aquiles de todas las máquinas de Turing es que para ninguna de ellas es el caso que posean una categoría que diga "verdad pero no es teorema" y

[49] A Turing machine can *know* what it can prove, that is, deduce from axioms using well-defined rules of inference. Its axiomatic database may be large, and its rules of inference may be efficient and powerful. But for a Turing machine, *knowability = provability*. When truth and provability part company, as in the case of Gödel's incompleteness theorem, a machine can cope with one but not with the other. (McCall, 1999, p. 525)

> "teorema pero no es verdad". El golfo que separa a las máquinas de las máquinas en cuanto a la forma de pensar, es el mismo golfo que separa la demostrabildiad de la verdad. (McCall, 1999, pp. 531-532, mis negritas)[50]

La estrategia de McCall (1999) es novedosa en el sentido de que abarca dos posibilidades que raramente se discuten con esta profundidad y obtiene las conclusiones epistémicas de que, en principio, las facultades cognitivas de las MTs son distintas de las facultades cognitivas de los humanos (idealizados o de los mejores seleccionados, quizá). Y si las facultades cognitivas, *sub specie aeternitatis*, tienen un catálogo distinto de *creencias* o *teoremas*, entonces son facultades cognitivas distintas, siendo superiores las humanas ya que tienden mejor hacia la verdad, si es que eso es lo que importa para la inteligencia, la racionalidad y en general, para cualesquiera facultades cognitivas actuales o posibles. No reproduciré detalles técnicos, pero si la forma general de los AGs la tenemos por supuesta, entonces McCall simplemente agrega la siguiente información, al respecto de la premisa de que *los humanos podemos saber que G es verdad, pero indemostrable*.

La estrategia del los AGs no TTeables es moderna en el sentido de que poniendo en duda todas nuestras tesis metafísicas acerca del mundo, encontramos no obstante esto, *contra* el escepticismo radical (que pone en duda la veracidad de todas nuestras tesis metafísicas y sus más comunes fuentes de demostración), una serie de verdades desde las cuales podemos aspirar a conocer el mundo. Y dado que el inventario de cosas en el mundo está vacío (puesto en suspenso), pues partimos en la lista con el primer elemento: tenemos al menos una verdad cierta. El *cogito cartesiano*, por ejemplo, la tesis de que *quien duda de todo existe*, en tanto dude de todo, sigue esta estrategia que he llamado "moderna." Pues los AGs no TTeables se inscribirían en esa misma tradición: ponemos en suspenso todas nuestras tesis acerca de la naturaleza de la mente humana y sus funciones cognitivas. La mente cognoscente puede ser grande o chica, material, no material o mixta, describible como se describe un evento físico o indescribible como se describe un evento físico, puede ser finita o infinita. Suponemos que es finita metafísicamente hablando, y ello tiende a comprometernos con alguna forma de

[50] What now of stronger machines, equipped with more powerful axiomatic bases and able to generate larger sets of theorems? What of machines equipped with the truth operator T? What of machines with huge data banks? It is difficult to see why the situation, for them, is any different from what it is for a PA machine. There still remain the two categories of theorems and nontheorems, and there still remains, for each of them, a sentence analogous to A10 [i.e., theorem that says: "if PA is consistent, then its Gödel sentence G is not provable"]. This sentence is unproven but true. The Achilles heel of all Turing machines is that for none of them does the categories of "true but not a theorem" and "theorem but not true" exist. The gulf that separates human from machine thinking is the gulf that separates provability from truth. (McCall, 1999, pp. 531-532)

computabilismo: el computabilismo que llamo metafísico. Suponemos que la mente puede o no ser finita, pero al menos es describible de forma finitaria, y ello tiende a comprometernos con alguna forma de computabilismo: el computabilismo que llamo cognitivo. Si los AGs tienen la razón, entonces al menos perdemos la tesis de computabilista metafísica. Y este ya es un hallazgo importante.

En el caso de McCall (1999) y cualquiera que siga esos mismos pasos, el AG general que expuse en el capítulo I de esta tesis, modificaría interesantemente la forma general de los AGs, agregando a que *los humanos pueden saber la verdad de G* la siguiente conjunción:

> Si las MTs computan G, las MTs computan una falsedad, y si las MTs no computan G, entonces G es una verdad incomputable.

Después, McCall (1999) agrega que los humanos no se ven constreñidos necesariamente a creer en esas falsedades (la oración G demostrada) en el mismo sentido en el que para una MT Gödel-susceptible es un hecho que creen en cierta clase de falsedades. Cualquiera de las dos posibilidades diferencian a las MTs de los humanos, ya sea metafísicamente o, como veremos en la sección siguiente, también cognoscitivamente. En el caso del primer disyunto, tenemos un *jaque mate más.* Éste consiste en que aunque *en cierto sentido*, *G es demostrable* (fuera del sistema, con herramientas distintas a las de LPO+PA, asumiéndolo como axioma, usando facultades distintas a las describibles por PA y LPO, o como sea).

Las facultades cognitivas pueden ser tomadas como *objetos con características*. La actividad de las cuales sirve de soporte metafísico de hechos, sean o no hechos computacionales y psicológicos, tales como *tal humano sabe que G,* o bien, *tal MT no puede saber que G*, o *tal MT sabe que G*, debe poder explicar cómo *soporta* o *causa* tales hechos epistémicos. En el caso del AG de McCall (1999), la diferencia entre MTs y humanos, estriba en que sea o no verdad la oración G, general o relativa al agente racional, si las MTs creerían compulsivamente la veracidad de la oración G relevante, dado que la demostrarían de acuerdo a cierta noción de *ejecución de una demostración*, entonces tendríamos una diferencia metafísico-epistémica entre los dos tipos de agentes, Turing-mecánicos y, humanos. La diferencia sería metafísica: para la MT creer o sostener, de cualquier modo en que una MT pueda creer, sería evidencia suficiente para arrojar una cantidad grande de dudas sobre sus propios procedimientos de demostración. Dadas ciertas relaciones lógico-epistémicas, incluso estaríamos frente a una

merma en la capacidad de la certeza de la MT en cuestión tan grande, que no podría diferenciar el peso epistémico de una verdad necesaria por sobre una contingente, digamos, 2+2=4 por sobre *veo una mancha negra sobre la pared.*

La solidez intuitiva de estos argumentos descansa en una relación cuya verdad no es simple de aprehender, en caso de que lo sea. La falacia verbal consiste en inferir una característica del referente de un término a partir de que esa característica la tenga la descripción del referente. Parece un error craso tratar de encontrar una "r" en una rosa, o sostener que una rosa es más pequeña que un tulipán, por el sólo hecho de que "rosa" es más pequeño que "tulipán". Pero un escenario de contraste podría poner la perspectiva correcta. Algo podemos aprender del entorno en el que están los estados cognitivos, a partir de los estados cognitivos, de la misma forma en la que sabemos qué come un animal dentado, a partir de conocer bien los dientes y el sistema masticatorio. Así, en el caso del AG de McCall, el asunto es doble: si el estado cognitivo preciso sucede de la forma en la que él dice suceder, entonces las máquinas de Turing no pudieron producirlo, o, si el estado cognitivo preciso fue producido por alguna máquina de Turing, entonces ese estado cognitivo no es lo que él dice que es.

El problema más grande con este tipo de AGs es que la conclusión a la que llegamos es bastante más fuerte que una conclusión semejante, pero, por ejemplo, condicionada a la corrección exacta del más sutil AG en consideración.

4.4.2 AGs cognitivo y no Turing testeable (tipo IV)

La tesis computabilista entendida como una tesis metafísica tiene la desventaja general de que implica que las cosas en sí mismas son computables. Que si uno aprende lo suficiente de un objeto metafísicamente computacional, uno no encontrará otra cosa que cierta cantidad y cierto tipo de elementos, y relaciones entre ellos. En el caso de los objetos metafísicamente computacionales, uno no encontraría otra cosa que estados finitos, transformaciones contables de estados (relaciones de causalidad), por mencionar un par solamente. En cambio, si uno suscribe una tesis no metafísica del computabilismo, esa carga simplemente no está allí: el criterio que queda para determinar la verdad de una tesis computabilista no metafísica, es qué tanto permite a satisfacción del científico o de la persona que juzga sobre la veracidad, una cantidad adecuada de explicaciones y predicciones, o qué tanta coherencia con el corpus

científico tiene, o con las tradiciones de las prácticas científicas hay o así, pero el compromiso del computabilista cognitivo (no del metafísico) es mucho más débil, y en este sentido es compatible su éxito con la existencia de fenómenos no computacionales, por pocos o improbables que sean comparados con el resto de los fenómenos.

Algunos autores que suscriben esta clase de AGs aunque no son del todo explícitos, son por ejemplo Haifman (2000), Tiezsen (2011, *et aut.*) y McCall (1999). Y con algunas concesiones relativamente menores de interpretación, el propio Gödel (1951), Penrose (1995) y más.

Sucede también con el presupuesto epistemológico de la *Turing* testeabilidad*. Afirmar que un fenómeno es *Turing* testeable* es tanto como decir que lo podemos conocer por medio de pruebas de cierto tipo. Pero el avance metodológico que representa el marco del *Turing* test* no debe soslayarse a la ligera.

No obstante lo anterior, si tenemos razones suficientemente buenas para hacernos dudar de ambos presupuestos, tanto el computabilismo metafísico como la *Turing* testeabilidad*, pues debemos hacerlo. Debemos pues dudar de ambos presupuestos en especial si podemos aspirar a profundizar nuestra comprensión de los fenómenos mentales como el conocimiento aritmético y su relación con otros tipos de conocimiento. Este es el caso aquí; para dar cuenta del conocimiento aritmético en el nivel metafísico mínimo que presuponen las ciencias naturales que lo estudian, pues tenemos que prepararnos para abandonar la metodología computabilista. Las ciencias naturales que estudian los fenómenos mentales y cognitivos son la psiquiatría, neurología, la lingüística, psicología y la *inteligencia artificial*.

Una teoría verdadera y computabilista de un fenómeno no computable, no puede ser óptima, aunque no esté equivocada. Lo vimos en el capítulo anterior; si tal teoría es óptima, debe contener dentro de sí términos que refieran o describan todos los procesos no sólo actuales sino que nos ayuden a determinar los procesos posibles mediante sus elementos constitutivos y las funciones que los vinculan; no sólo necesitamos un listado de los estados cognitivos de ciertos sujetos, sino especialmente conocer las formas en que cada uno causa y se vincula a los demás (no sólo una lista de correlaciones por más que sepamos que son relaciones biyectivas). Si la teoría de la mente verdadera y computabilista es descriptiva o no, no importaría porque sabríamos que no es óptima, y en particular no puede describir ni representar mínimamente

todas las relaciones causales entre estados mentales cognitivos, dado que siendo verdadera describe o refiere a fenómenos no computables.[51]

No debe tenerse a esta clase de argumentos como argumentos cuyas condiciones de solidez son tan poco comunes que quizá no valiera la pena invocarlos para hacer alguna modificación metodológica seria o bien adoptar alguna posición eliminativista particular. Primero, el uso de la aritmética aunque parece bastante acotado en la vida mental humana, es crucialmente importante para la misma. Nos permite usar y aplicar la inteligencia a situaciones de vida y muerte, de florecimiento y búsqueda de la felicidad, o al menos, evitar la infelicidad. Lo mismo aplica para el uso de la lógica en las mentes y vidas humanas. Segundo, las facultades cognitivas aritméticas y funciones cognitivas semejantes (recursivas) están presupuestas siempre que comprendemos lingüísticamente alguna expresión, por ejemplo, la normatividad del significado. La normatividad del significado nos permite estar en la disposición de aplicar un término lingüístico en una situación distinta a una anterior, a veces variando entre ellas casi sólo ordinalmente; quizá yo nunca me he tomado la cerveza que tengo yo frente a mí en este momento, pero ello no evita que yo sepa que "estoy tomando una cerveza más" de las cervezas posibles reales, actuales, pasadas o potenciales. También, tercero, la discusión y comprensión de los AGs nos ayudará a entender mejor las características metafísicas de la mente humana y de sus facultades cognitivas, conocer un tipo más de eventos que pueden suceder o que de hecho suceden en el universo. Penrose (1995) mismo afirma que si su AG es correcto, ello obligaría a revisar un presupuesto de la física actual: el finitismo.

---

[51] Piénsese en una verdad aritmética y ciertos estados mentales correspondientes a ellas. La creencia de que *2+2=4*, se representará como un estado epistémico distinto del de la creencia *2+1+1=4*, así

C(2+2=4) es distinta de C(2+1+1=4)

aunque

2+2=1+1+2

y

⌜2+2=4⌝≡⌜1+1+2=4⌝

Considerando la equivalencia anterior, el fenómeno mental

C(*siempre hay un número racional entre dos números racionales cualquiera*)

impone el requisito para el marco conceptual de una cantidad por lo menos potencialmente infinita de fórmulas de las ciencias cognitivas que permitan no sólo expresar la diferencia entre estados de creencia C(x) sino de la posibilidad de discernir relaciones tanto entre las creencias, como entre sus contenidos dado que necesitemos dar cuenta de los procesos mentales que las producen, sostienen, causan o incluso justifican. Nuestra psicología podría describir bien todos los estados de creencias aritméticas cuyas relaciones tengan como resultados números naturales enteros positivos en una cantidad finitaria de instancias, y hasta ahí que la psicología cognitiva humana sea exitosa, pero a la vez, esta psicología podría no tener resueltos temas normativos de la teoría de números como son precisamente compleción y corrección, o creencias correctas indeterminadas sobre tamaños de tamaños de infinitos. Para más detalles sobre este asunto dirigirse a Shapiro (1998, pp. 273-274, 280 y 300), que cita a Dummett (1963) o al propio Gödel (1951).

Los AGs del tipo que tratamos aquí, tratan a la premisa computabilista (las funciones cognitivas de la mente humana son computables) como una afirmación modalmente más débil que en los AGs metafísicos y no Turing* Testeables. Así, el conocimiento que sobre las funciones cognitivas nosotros tenemos de nosotros mismos o de otras funciones cognitivas.

Una combinación interesante aquí es aquella en la que las funciones cognitivas se pueden representar como teorías computables, aunque no óptimamente, y que la teoría que nosotros tenemos de ella es meramente nominal. Podríamos discutir hasta qué grado una teoría puramente nominal reporta conocimiento. Pero las teorías pueden ser parcialmente nominales, en el sentido de que sepamos cómo es el mundo nominalmente a la perfección y sepamos a qué se parecería eso, o cómo debería verse eso de forma descriptiva. Un ejemplo de esta clase de teorías del mundo es por ejemplo el aparato matemático con el que describimos a los fenómenos cuánticos; parece ser muy difícil de imaginarse. No obstante esto, la jerga técnica cumple el propósito. Quizá otro caso, también sacado de la física, la óptica: nadie de nosotros sabe *cómo se ve el color infrarrojo* de la misma manera en que sabemos *cómo se ve el color rojo* o *el verde*. La teoría física que nos dice que el rojo y el infrarrojo son sólo dos puntos distintos en una misma escala, implica que ignoremos cierto conocimiento directo de los mismos, aunque sí tengamos idea de cómo funcionan mediante teorías del mundo debidamente comprobadas.

Regresando al tema del caso curioso: si estuviéramos en un mundo que verifica las premisas y valida el AG cognitivo y no Turing* testeable, tendríamos ciertos aspectos de la psicología del conocimiento que conoceríamos de la misma forma en la que conocemos la existencia del infrarrojo. Lo que es más, no podríamos aplicar algún TT para saber de la ocurrencia de algún *estado mental cognitivo* que indique la ausencia de Gödel limitación por alguna razón en las líneas siguientes: no sabemos qué buscar exactamente. Por ejemplo, no sabríamos qué cantidad de información empírica cerebral, mental, lingüística o lo que sea, basta para determinar Gödel limitación. Este caso encuadra bien en el tipo de AG que tratamos ahora: ¿Qué clase y cantidad de evidencia determina ausencia de Gödel limitación? ¿Qué fenomenología (evidencia subjetiva) tendría el convencimiento *correcto* de que cierto agente no está Gödel limitado? Si no apeláramos a alguna fenomenología, sino a cierta evidencia traducible a evidencia Turing computable, entonces esa evidencia decidiría la solidez de otras clases de AGs, no este necesariamente uno específico como el tipo de AG que presento aquí.

Así que un argumento del tipo que tratamos ahora, probablemente involucre un escrutinio personal, subjetivo, del tipo de actitud proposicional que tal sujeto tiene con respecto a la oración G del sistema del que el sujeto en cuestión forma parte: ¿cómo me convenzo de que la oración G relevante que me involucra a mí mismo, es el caso? Si creo que tal oración G es verdad, ¿lo creo sin justificación alguna? ¿mi justificación demuestra necesariamente o sólo contingentemente? Suponiendo que tenemos evidencia subjetiva (*sabemos*) que cierto agente, quizá uno mismo, no está Gödel limitado, todavía cabría preguntarse de qué tipo de conocimiento es éste. Si el conocimiento tiene contenido empírico o no, si nos enseña algo de psicología humana o no, si nos enseña algo de lógica o no, si nos enseña algo de los conceptos de *conocimiento*, *justificación* o lo que sea. Dicho de otro modo, supongamos que esta clase de AG es exitosa, lo que implicaría que ciertas premisas son verdad y la presuposición computabilista es falsa; ¿qué condiciones metafísicas y/o conceptuales *garantizan* o *respaldan* la verdad y falsedad de las premisas que lo son?

Quizá la siguiente situación que guarda ciertas relaciones de semejanza ayude a entender las dudas arriba expuestas:

> Esta afirmación es indemostrable. Por tanto, la afirmación anterior es indemostrable.

¿Si alguien dudara de la verdad la afirmación de indemostrabilidad, cómo podríamos convencerle? Parece que es imposible "demostrarle" la verdad de la afirmación en cuestión, puesto que incurriríamos por lo menos en vicios retóricos obvios, aunque los vicios no sean, en este caso en especial, notoriamente lógicos para la conclusión del argumento y aunque parecen ser semánticamente comprensibles.

Una posibilidad más es que por razones prácticas sea imposible establecer la teoría del mecanismo computable conforme al cual funciona la mente humana en general o alguna mente particular, la propia mente incluida. En este caso, aunque el mecanismo subyacente podría ser, metafísicamente computable, la mejor forma de describirlo, no computable. Esta estrategia es semejante a presuponer que las personas son libres, para mejorar nuestra capacidad predictiva y explicativa de su conducta, aunque en el fondo el mecanismo con el que funcionan las personas sea completamente mecánico. Al ignorar el mecanismo computacional y adoptar como marco conceptual que suponga que las facultades cognitivas humanas son no computables, podemos ganar cierta comprensión de la dificultad que enfrentan las ciencias

humanas (ciencias que lidian con las facultades cognitivas humanas y sus efectos, acciones individuales, colectivas, estados psicológicos, etc.), estaríamos incurriendo en una violación al principio de parsimonia o *navaja de Ockham*, pero sobre todo volveríamos a nuestros artificios conceptuales (teorías de las facultades cognitivas) básicamente ambiguos o vagos. Las implicaciones prácticas de la investigación científica no son sencillas de considerar, pero del hecho de que una teoría que suponga que el comportamiento humano está indeterminado, por ejemplo, quizá esté más abierta a admitir fenómenos conductuales humanos irregulares sin prejuzgarlos de forma alguna, pero ello no querría decir que no podamos tener exactamente la misma ventaja sin tener que comprometernos con supuestos tan caros metafísicamente hablando. El número real π quizá no sea *predecible*, por ejemplo, o quizá no es fácilmente explicable su secuencia de valores decimales, pero ello para nada quiere decir que sea un número arbitrario. No tenemos alguna razón fuerte para admitirlo así, y menos aún que estamos condenados a lanzar cualquier dígito como el *siguiente dígito* en algún intento de secuenciación de π. Mejor sería guardar silencio.

Regreso al punto del párrafo anterior. La naturaleza de las funciones cognitivas humanas no es arbitraria, aunque quedara indeterminable dada la fuerza de ciertas razones como las de los AGs. Pero si estuviéramos ante un caso como el del párrafo anterior, quizá sería mejor suspender el juicio al respecto de este asunto. Es bastante más claro que en caso de que no podamos explicar la maquinaria que sustenta nuestras facultades cognitivas no es una razón para descreer de sus productos *so pena* de incurrir en falacias. Por lo demás cierta clase de verdades (productos, por decirlo así, de estas facultades) son suficientemente fuertes como para no disminuirlas incluso en el caso de que no sepamos cómo es que las sabemos. Piénsese: ¿"2+2=4" dejaría de ser verdad sólo porque ignoramos qué neurona disparó un estímulo y qué otra neurona estuvo inactiva o sería dudosa porque esa afirmación saliera de un proceso computacional? Aunque tuviéramos toda la información metafísica concerniente a nuestras facultades cognitivas, ¿podríamos de allí concluir correctamente que 2+2 no es 4? No parece plausible, aunque ignoremos todo al respecto de cómo es que sabemos que lo sabemos o bien cómo es que funciona nuestra arquitectura cognitiva. Dicho de otro modo, es más sencillo conservar ciertas creencias necesariamente verdaderas que tratar de ponerlas en duda por lo que ignoramos acerca de nuestra arquitectura cognitiva, en especial si es computabilista o no.-

Siguiendo con la metáfora de la *arquitectura* tenemos que si ésta es computable, finita, y tuviéramos la oración G relevante como *built in*, como si fuera un axioma, no podríamos

percatarnos de la veracidad de la oración G, o de alguna de las oraciones G producto de revancha, aunque pudiéramos, digamos, *decidir que sí lo es* (*Entscheiden*) de forma conductual o no; lo que no podríamos hacer, en casos como este, es explicar las razones por las cuales sabemos que la afirmación de un sistema Gödel-susceptible, sea AP+LPO o bien, AP+LPO además de nuestra propia arquitectura cognitiva. Pero la justificación de por qué "2+2=4" es tan clara, que pocas personas creerían que esta afirmación aritmética es falsa, por tal de asegurar que la mente humana es un fenómeno computable.

Esta clase de argumentos es la clase más fuerte de AGs, dadas las críticas más comunes que se han presentado y reformulaciones de las mismas. También, rescatan de forma natural la posibilidad de que nuestras teorías puedan estar correctas y completas en cierto nivel (no teorías óptimas en el sentido esbozado en el capítulo 3 de esta tesis), aunque en otros niveles y quizá bajo otros criterios estén inundadas de debilidades epistémicas por razones más específicas. Pero la mera posibilidad del éxito en los esfuerzos de auto-conocimiento, como especie o como individuos, depende de que las relaciones lógicas más estrictas, aquellas que son a la vez reflexivas, simétricas y transitivas son presupuestas por actos o eventos psicológicos como la reflexión, la consciencia, el cambio de creencias, la educación, el orden de nuestras prioridades, y quizá incluso la autonomía.

Sabemos que la oración G, en general es cierta. Sabemos que la oración G aplicable al modelo de nuestra propia mente, dado que sea Gödel no limitada, es verdad porque sabemos que la aritmética de Peano es consistente, y que la LPO es consistente y completa. También sabemos que al ser consistente, la compleción de LPO y de PA, nos compromete con que la forma en la que sabemos esto es necesariamente no computacional y no TT. De este modo, concluimos que lo que sea que sea ese mecanismo, es real.

5. CONCLUSIONES

Vimos una definición de *Argumento Gödeliano* (AG). Para que un argumento sea un AG es necesario que éste tenga como premisas algunas afirmaciones dentro de las cuales está una premisa que para ser verdad, requiere básicamente de que la incompleción de Gödel en general, o para el sistema concreto desde el que se fundamenta dice o sostiene el AG en particular, sea verdad.

No todo sistema (formal o no) podría ser Gödel incompleto, porque no todo sistema formal o no es Gödel-susceptible. Por ejemplo, un sistema formal que tiene un máximo de teoremas posibles más pequeño que el infinito más pequeño, no es Gödel-susceptible. Para que un sistema pueda ser Gödel incompleto se necesita: a) ser capaz de producir los teoremas de la Aritmética de Peano o algo isomórifico, b) ser capaz de interdefinir términos (para lo que técnicamente se conoce como *numeración de Gödel*), c) ser capaz de deducción o un mecanismo que refleje deducción (*lógica de predicados de primer orden*). Una vez que un sistema (formal o no) puede ser incompleto al estilo de incompleción de Gödel, lo llamaremos Gödel-susceptible.

Dentro del conjunto de todos los sistemas Gödel susceptibles, hay algunos que son incompletos y al menos en principio hay algunos que pueden ser completos suponiendo que no pueda deducir su completud, entre quizá otros requisitos.

Cuando un sistema es Gödel susceptible, pero no es Gödel limitado de modo relevante es cuando ese sistema tiene a su disposición recursos cognitivos que superan *de algún modo* los recursos que se requieren para la Gödel susceptibilidad. En qué medida y cómo es que esto sucede, es objeto de la discusión alrededor de las condiciones de solidez de los AGs.

Las premisas que componen el AG, así, pueden ser verdaderas o falsas. En el capítulo 2 vimos las condiciones con respecto a las cuales se puede evaluar la verdad de dichas premisas. Hay dos grandes formas de determinar la veracidad o falsedad de las premisas. Una de ellas es consistente con el testeo de Turing: son formas computables para decidir la verdad de las premisas del AG, el que sea. Las otras no son necesariamente Turing testeable. A la evidencia no-Turing testeable, siguiendo a Gödel mismo, y a algunos exegetas posteriores como Tieszen,

Hao Wang y otros, la llamo evidencia subjetiva. La distinción es que la evidencia subjetiva es básicamente la percepción fenomenológica de la veracidad de ciertas afirmaciones. La explicación de cómo es este mecanismo recae dentro de la discusión más grande: si la intuició funcionara a la base de la misma como un proceso computable, entonces sería una forma de evidencia Turing testeable (TT). Por definición necesitaríamos que este no sea el caso. De modo que la evidencia subjetiva no TT sea al menos consistente. Supondremos que lo es, independientemente del mecanismo ontológico que haya detrás de la misma, siempre que, como mencioné antes, no sea TT en modo alguno.

En el capítulo 3, tratamos en el asunto de la semántica de la tesis computabilista. Una de las premisas de AG se llama *tesis computabilista*. En ella se establece que todo lo que hace nuestra mente o cerebro, es computable, y en ese sentido, TT. Esta tesis es metafísicamente ambigua. Una cosa es que el mundo sea describible en al menos cierto nivel o para ciertos objetivos, como un proceso computable, y otra cosa es que lo sea. La pasada es una interpretación metafísica del computabilismo. La previa a la pasada, es la tesis de que conocemos la mente o el cerebro tratándolo como un sistema computable. La tesis metafísica de la computabilidad de la mente implica a la cognitiva, pero no la inversa. No obstante, sabemos que un conocimiento óptimo (i.e. completo) de la mente o la arquitectura cognitiva de los agentes razonadores si es computable, ello implica que el fenómeno descrito es computable. Para que el fenómeno sea metafísicamente no computable, debe suceder que una teoría óptima del mismo, no lo es.

Así, llamamos al proceso cognitivo de razonar con aritmética *causalidad psicológica*. Si la *causalidad psicológica* es *computable o no,* en términos metafísicos o cognitivos, es el problema a superar con estas argumentaciones y otras. Un tipo de *causalidad psicológica* no computable, es la *causalidad hipercomputacional*. Que la causalidad psicológica no sea computacional no quiere decir que no sea mecánica o computacional, en general, sino que no es Turing computacional. Otros modelos de explicación serían los que suponen que la mente es un fenómeno básicamente superior en tamaño, al computacional. A ese modelo se le llama hipercomputabilidad. Hay, aún para la explicación hipercomputable de la mente humana bastantes problemas pendientes. Pero de entrada, sólo tengo espacio para señalar o caracterizar qué es o qué se consideraría *no computable*, en términos de evidencia, véase el capítulo 2, tanto como qué es en términos ontológicos, véase el capítulo 3.

Una vez sentadas las bases anteriores, definiciones, y tesis, cabe la pregunta: ¿existe algún AG que tenga la posibilidad o que sea sólido? Aunque la pregunta es ambiciosa, subdivido los AGs en cuatro subtipos. Si interpretamos la tesis computacional de forma metafísica o cognitiva, y si aceptamos como evidencia válida procesos TT o no. La combinación total de estas cuatro características, que son exhaustivas, nos permite clasificar los AGs en:

Tipo I: AG metafísico y TT
Tipo II: AG cognitivo y TT

A estos dos tipos se les han hecho muchas críticas. Una de ellas, la más devastadora, consiste en mostrar que si el modelo de la mente o del agente racional en cuestión (puede ser artificial o natural, hipotético o actual) no nos permitiría inferir si dicho agente no sólo es Gödel-susceptible, sino si es Gödel-limitado. Puesto que para saber si lo es o no, tendríamos que aspirar a que el sistema que explica o describe a dicho agente o sistema formal, es completo y coherente. El punto de la incompleción de Gödel es justamente mostrarnos que esto no puede hacerse.

Pero los tipos III y IV,

Tipo III: AG metafísico y no-TT
Tipo IV: AG cognitivo y no-TT

Son no sólo inertes a las críticas que se le hacen a los tipos I y II, sino que son las dos formas más caritativas de interpretar un AG. El objetivo, vimos en el capítulo I de la tesis, de los AGs es mostrar que no todo lo que sucede en nuestras mentes, y en esa medida, en el universo, es TT o computacional. De modo que si prejuzgáramos el éxito de los AGs como requiriendo que su evidencia o modelo sea TT, estaríamos incurriendo en petición de principio falaz.

Así, en el capítulo 4 vimos las debilidades y fortalezas de los 4 tipos de AG. De cada tipo de AG he proporcionado bibliografía sobre muy probables proponentes. Recordando quedando que *si un AG tipo III es sólido, entonces el AG tipo IV*, lo es, pero no a la inversa, es más sencillo defender la solidez de un AG tipo IV. Es más fácil saber que *p*, que *acceder* a *p*, sobre todo si *p* es inferible o accesible por algún método de introspección, o registro.

Así, podemos inferir que el AG que más posibilidades tiene de ser sólido, es el de tipo IV, ya sea porque la no computabilidad es hipercomputabilidad, es fenomenológica pero no computacionalmente accesible, o por cualquier otro motivo.

Argumento en la sección 4.4.2, que basta con que la aritmética de Peano sea consistente, y que no hay modo en que por allí haya alguna tesis cuya valor de verdad no podamos saber. "2+2=4" no tiene excepciones, porque "4=4" y "2+2" es "4". Lo mismo con el resto de las afirmaciones de la aritmética. Es más fácil aceptar que no sabemos bien cómo explicar nuestra forma de conocer las verdades de la aritmética, a aceptar que dado que ignoramos cómo funciona nuestra arquitectura cognitiva, "2+2" podría no ser "4". Y así con todas las afirmaciones de la aritmética.

Por tanto, sabemos que G es cierto, que la G aplicable a nuestro propio modelo lo es, y a la vez la ontología de nuestras arquitecturas cognitivas, siempre que sean Gödel-ilimitadas, son consistentes. Si no lo fueran, G sería cierta igual para ellas. Así, hay al menos un AG cuya solidez es palpablemente considerable.

Una estrategia a favor de la solidez de algún AG, en especial de tipo metafísico y no TT, consiste en encontrar elementos que nos permitan establecer

Es importante considerar este trabajo como un intento por discutir un conjunto de elementos con los que se puede reconstruir la prueba detallada. Pero aquí está la información mínima básica para acercarse al problema planteado por quienes sostienen AGs de forma seria. En especial, Gödel, McCall, Gaifman, Putnam (en *Reflexive Reflections*), y Shapiro.

REFERENCIAS

Capitulo 1

Enderton, Herbert B. (2001). "A Mathematical Introduction to Logic." 2nd Edition. Harcourt Academic Press.

Bringjord, S. & Xiao, H. (2000). A Refutation of Penrose's Gödelian Case Against Artificial Intelligence, en el *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 12, No. 3, pp. 307-329.

Copeland, Jack B. (2002). Hypercomputation. En Minds and Machines, Vol. 12, pp. 461-502.

Lacey, Hugh y Joseph, Geoffrey. 1968. What the Gödel Formula Says. En Mind, Vol. 77, No. 305, pp. 77-83.

Dawson Jr., John (1984). The reception of Gödel's incompleteness theorems, pp. 74-95 en Shanker, S. G. (ed., 1998)

Gaifman, Haim (2000). What Godel's Incompleteness Result Does and Does Not Show. En *The Journal of Philosophy*, Vol. 97, No. 8, pp. 462-470.

Goble, Lou (ed.) (2001). The Blackwell Guide to Philosophical Logic. Editorial Blackwell Guide.

Hodges, Wilfrid (2001). Classical Logic I: First-Order Logic. En Goble, Lou (ed.). 2001. The Blackwell Guide to Philosophical Logic. Editorial Blackwell Guide, pp. 9-32.

Shanker, S. G. ed. (1988). *Gödel's theorem in focus*, de la Philosophers in focus series, Routledge, London and New York 1989, ix + 261 pp.

Smullyan, Raymond (2001). Gödel's Incompleteness Theorems. En Goble, Lou (ed.). 2001. The Blackwell Guide to Philosophical Logic. Editorial Blackwell Guide, pp. 72-90.

Smullyan, Raymond (1992). Gödel's Incompleteness Theorems. Editorial Oxford University Press.

Capítulo 2

Bringsjord, Selmer, Noel, Ron & Caporale, Clarke (2000). "Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence." En Journal of Logic Language and Information 9 (4):397-418.

Buechner, J. (2008). *Gödel, Putnam, and Functionalism, A New Reading of* Representation and Reality. MIT Press, EUA: Cambridge.

Dummett, Michael A. E. (1978). "Truth and Other Enigmas." Harvard University Press, EUA: Massachusetts.

Kirk, Robert (1986). "Mental Machinery and Gödel." Synthese, vol. 66, pp. 437-452.

Kripke, Saul (1982). *Wittgenstein: On Rules and Private Language: An Elementary Exposition*. Cambridge, Mass.: Harvard University Press, 1982.

Harnad, Steve, (1992). "The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion." en SIGART Bulletin 3(4) (Octubre de 1992), pp. 9 - 10.

Maudlin, Tim (1995). "Between The Motion And The Act… A Review of *Shadows of the Mind* by Roger Penrose." Psyche, 2, vol. 2.

Quine, W. V. O. (1951). "Two Dogmas of Empiricism." En The Philosophical Review 60 (1951), pp. 20-43.

Putnam, Hilary (1985). "Reflexive Reflections." Erkenntnis, Vol. 22, pp. 143-153.

Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind, New Series, Vol. 59, No. 236., pp. 433-460.

Tieszen, Richard (1984). "Mathematical Intuition and Husserl's Phenomenology." En *Noûs*, Vol. 18, No. 3, pp. 395-421.

Tieszen, Richard (1998). "Gödel's Path from Incompleteness Theorems (1931) to Phenomenology (1961)." En *The Bulletin of Symbolic Logic*, Vol. 4, No. 2, pp. 181-203.

Tieszen, Richard (2005). *Phenomenology, Logic, and the Philosophy of Mathematics*. United States of America, Cambridge, Massachussets: Cambridge University Press.

Tieszen, Richard (2011). "After Gödel: Mechanism, Reason, and Realism in the Philosophy of Mathematics." En *Philosophia Mathematica* (III) 14 (2006), pp. 229–254. - Wang, Hao. 1996. A Logical Journey: From Gödel to Philosophy. Cambridge, Massachussets. MIT Press.

Capítulo 3

Block, Ned. (1981). *Psychologism and Behaviorism* en *The Philosophical Review*, vol. LXXXX, No. 1, pp. 5-43, accesible desde el sitio <URL = http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Psychologism.htm>

Barwise, Jon y Etchemendy, John (1987). *The Liar: An Essay on Truth and Circularity*. Editorial Oxford University Press.

Bringsjord, Selmer, Kellett, Owen, Shilliday, Andrew, Taylor, Joshua, van Heuveln, Bram, Yang, Yingrui, Baumes, Jeffrey & Ross, Kyle (2006). "A new Gödelian argument for hyprcomputing minds based on the busy beaver problem." En *Applied Mathmatics and Computation*, Vol. 176, pp. 516-530.

Brown, Curtis. (2004). "Implementation and Indeterminacy." En *Computing and Philosophy Conference*, Vol. 37, pp. 27-31. <URL = http://delivery.acm.org/10.1145/1090000/1082150/p27-brown.pdf?key1=1082150&key2=9354580721&coll=GUIDE&dl=GUIDE&CFID=83604717&CFTOKEN=93945725>

Chalmers, David. J. (1993). "A Computational Foundation for the Study of Cognition." Borrador en <URL = http://www.consc.net/papers/computation.html>

Chalmers, David. J. (1996). "Does a Rock Implement Every Finite-State Automaton?" En *Sythese*, 309-333. Liga no estable <URL = http://cogprints.org/226/0/199708001.html>

Cotongo, Paolo. (2003). "Physical Church-Turing Thesis." En *The British Journal for the Philosophy of Science*, Vol. 54, pp. 181-223.

Hintikka, Jaakko (1962). *Knowledge and Belief, An Introduction to the Logic of the Two Notions*. Ithaca, New York, USA: Cornell University Press.

Hofstadter, Douglas. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books.

Maudlin, Tim. (1989). Computation and Consciousness. En *The Journal of Philosophy*, Vol. 86, No. 8, pp. 407-432.

Piccinini, Gualtiero (2003). "Functionalism, computationalism, and mental states." En *Studies in History and Philosophy of Science*, Vol. 35, pp. 811-833. <URL = http://www.umsl.edu/~piccininig/Functionalism_Computationalism_and_Mental_States.pdf>

Potgieter, Petrus H. (2004). "Zeno Machines and Hypercomputation." En *Theoretical Computer Science*, Vol. 358, No. 1, pp. 23-33. Versión en línea en <URL = http://arxiv.org/PS_cache/cs/pdf/0412/0412022v3.pdf>, pp. 1-13.

Putnam, Hilary (1988). *Representation and Reality*. 7ª reimpresión. Editorial MIT Press.

Stannett, Mike (2003). "Computation and Hypercomputation." En *Minds and Machines*, Vol. 13. pp. 115-153.

Capítulo 4

Bringsjord, Selmer (1992). *What Robots Can and Can't Be*, Ed. Kluwer Academic Publishers.

Descartes, René (1641). *Meditations on First Philosophy*, traducido al Inglés por Michael Moriarty, Ed. Oxford University Press, 2008, pp. 32-33.

Dummett, Michael (1963). “The philosophical significance of Gödel’s theorem,” en *Ratio*, vol. 5, pp. 140–155, que aparece en Dummet (1978).

Dummett, Michael (1978). *Truth and Other Enigmas*, Harvard University Press.

Gödel, Kurt (1951). “Some basic theorems on the foundations of mathematics and their implications,” pp. 304–323 publicado en Gödel (1995).

Gödel, Kurt (1995). *Collected works III*, Oxford University Press, Oxford.

Haifman, Gaim, (2000). “What Gödel’s Incompleteness Result Does and Does Not Show.” En *The Journal of Philosophy*, pp. 462-470.

Kerber, Manfred (2005). “Why is the Lucas-Penrose Argument Invalid?” En *Springer Verlag, KI 2005: Advances in Artificial Intelligence, LNAI 3698*, pp. 380 – 393, Consultado el 9 de Marzo de 2014, en <URL = http://www.cs.bham.ac.uk/~mmk/papers/05-KI.html>.

Kirk, Robert (1986). “Mental Machinery and Gödel” en *Synthese*, vol. 66, pp. 437-452.

LaForte, Geoffrey, Hayes, Patrick J. & Ford, Kenneth M. (1998). “Why Gödel's Theorem Cannot Refute Computationalism.” En *Artificial Intelligence*, Vol. 104, No. 1-2, pp. 265-286.

Maudlin, Tim (1996). “Between The Motion and the Act” en *Psyche*, Vol. 2, pp. 40-51.

McCall, Storrs (1999). “Can a Turing Machine Know that the Gödel Sentence is True?” En *The Journal of Philosophy*, vol. 96, pp. 525-532.

Murphy, M. y O'Neill, L. A. J. (1997). *What is Life? The Next Fifty Years: Speculations on the Future of Biology*, Cambridge University Press.

Penrose, R. (1994). *Shadows of the Mind* Oxford: Oxford University Press.

Penrose, R. (1995). “Why New Physics is Needed to Understand the Mind,” en Murphy, M. y O'Neill, L. A. J. (eds.), 1995, pp. 115-130.

Smullyan, Raymond (2001). “Gödel’s Incompleteness Theorems”, En Goble, Lou (ed.). 2001. *The Blackwell Guide to Philosophical Logic*. Editorial Blackwell Guide, pp. 72-89.

McCall, Storrs (1999). “Can a Turing Machine Know that the Gödel Sentence is True?”, en The Journal of Philosophy, Vol. 96, No. 10, pp. 525-532

Turing, A. M. (1950). "Computing Machinery and Intelligence." En *Mind*, New Series, Vol. 59, No. 236, pp. 433-460.

Yalowitz, Steven (2012). “Anomalous Monism”, The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/anomalous-monism/>, consultada en Septiembre de 2014.