"¿Pueden pensar las máquinas? Una evaluación del problema de la inteligencia artificial mediante el equilibrio refractivo"

Trabajo presentado para optar al título de
Profesional en Filosofía
Escuela de Ciencias Humanas
Programa de Filosofía
Universidad del Rosario

Presentado por:
Orlando Uribe Cerdas

Director: Carlos G. Patarroyo G.

Semestre II de 2012

Contenidos:

0. Introducción

La posibilidad del pensamiento en las máquinas es una inquietud que se ha planteado desde hace tiempo; tanto la ciencia ficción como la ingeniería y la filosofía han buscado proporcionar respuesta a la pregunta "¿Pueden pensar las máquinas?". En la bibliografía pueden encontrarse exponentes famosos tanto de respuestas afirmativas, expuestas por Turing o Kurzweil, como negativas, siendo Searle o Penrose ejemplos famosos. A pesar de que cada uno de estos representa una postura diferente, existe una característica común a las diferentes propuestas: en todas ellas "pensar" se encuentra definido por la experiencia netamente humana de hacerlo. Al presentar el *Juego de la imitación* como un criterio para determinar si las máquinas piensan, se está declarando que ellas sólo lo harán en el momento en que puedan llevar a cabo comportamientos característicos de la manera de pensar humana; al negar que una máquina puede pensar debido a que, de encontrarnos en su posición, existirían características de nuestra inteligencia -tales como contenido semántico o demostración de verdades matemáticas- a las cuales no tendríamos acceso, estamos solicitando que, para una máquina "pensar", esta pueda llevar a cabo todas las funciones de pensamiento humano. Proceder de esta manera es natural en tanto nuestra experiencia como sujetos pensantes es aquella a la que tenemos acceso con mayor facilidad, sin embargo esto parece traer consigo un juicio injusto para otras posibles formas mentales.

Supongamos que también nuestra experiencia en primera persona fuera igualmente importante para establecer aquello que significa, por ejemplo, "respirar". En ese caso tendríamos como un componente fundamental de la definición la posesión de pulmones y una estructura similar a la humana. Bajo esta definición peces o células nunca serían considerados como exponentes de sujetos que respiran, dejando de lado los fenómenos de respiración celular o branquial. Del mismo modo en que dejaríamos de lado muchos otros sujetos respirantes mediante la exploración de cómo se presenta el fenómeno exclusivamente en los humanos también creo que es posible dejar de lado otras mentes al explorar el concepto de pensar exclusivamente desde nuestra perspectiva. Si se considera que es importante entender formas de pensamiento que vayan más allá de la humana, entonces será necesario construir una herramienta la cual permita la configuración del concepto de tal modo que exista más información que lo configure además de nuestra experiencia.

En el presente trabajo procuraré ejemplificar la forma en que el equilibrio refractivo[1] de Goodman puede ser justamente esta herramienta. Si bien el equilibrio refractivo de Goodman fue propuesto como una forma de otorgar justificación a las reglas de inferencia, este proceso podrá adaptarse de tal modo que permita alcanzar una definición. Al poner en equilibrio las reglas que permiten determinar si un objeto "piensa" por un lado y los ejemplos de objetos que "piensan" por el otro, el resultado del equilibrio se consideraría como la definición de pensamiento. Si bien podría parecer circular proceder de este modo, esto no es un accidente. Parte importante al momento de aplicar el proceso del equilibrio refractivo es la inclusión de aquello que no estaríamos dispuestos a negar sin importar los resultados de una formulación específica; esto es, aquello que el sentido común podría indicar como necesariamente perteneciente o ausente en un momento del equilibrio (en el caso planteado por Goodman, reglas de inferencia las cuales no estaríamos dispuestos a modificar e inferencias que no estaríamos dispuestos a aceptar como verdaderas). De esta manera, aplicar el proceso de equilibrio refractivo a una definición será el proceso de explorar quiénes y cómo hacen algo para definir qué es ese algo, siempre teniendo en cuenta al sentido común como garante de coherencia en el proceso. Por esto es un proceso adecuado para construir un concepto de pensamiento en el cual las otras mentes puedan ser tenidas en cuenta; si parte de las solicitudes que hace el sentido común al proceso es la posibilidad de existencia de objetos pensantes los cuales no lo hagan de la manera humana, entonces el proceso de equilibrio se encargará de conservar dicha posibilidad.

Lo que se busca al usar el equilibrio refractivo como herramienta para definir aquello que significa pensar es cambiar el peso que tiene para la definición la experiencia humana. Los estudios en inteligencia artificial parecen indicar que tomamos nuestra experiencia de sujetos pensantes como el único criterio para definir las características que cualquier cosa debe cumplir para ser catalogado como pensante. Al producir una definición a través del equilibrio refractivo de acuerdo con los planteamientos que se presentarán en el cuerpo de la monografía se buscará que ahora esta experiencia sea solamente una herramienta más para la definición. Nuestra experiencia humana será una guía para la definición, pero la forma en que extraemos de ella información deberá cuidarse de dar la oportunidad a otras mentes de ser tenidas en cuenta a pesar de su realización material. Esto no quiere decir que el proceso de equilibrio refractivo tenga que otorgar como resultado la existencia de mentes diferentes a la humana, de hecho un resultado del

[1] Ver nota al pie 15 (infra, pág 22).

proceso podrá ser que, de acuerdo con la información disponible en el momento, solamente tenemos ejemplos de pensamiento en nosotros. Sin embargo, la diferencia se encontraría en las razones por las cuales excluimos a las otras mentes; actualmente parece ser que la razón principal se encuentra en una imposición de nuestra perspectiva, mediante el equilibrio se encontrarán características representativas (e investigables en más sujetos que nosotros mismos) las cuales otorgaran una definición. De esta manera pasaríamos de preguntarnos "¿Pueden otros objetos pensar tal y cómo lo hacen los humanos?" a "¿Cómo llevan a cabo los otros procesos de pensamiento?" sin por esto suponer que ellos piensen, ya que el proceso de equilibrio permitirá que la investigación del cómo modifique lo que significa el concepto.

La primera sección de mi monografía presentará una reconstrucción del problema de la inteligencia artificial en filosofía a través de la pregunta "¿pueden pensar las máquinas?" para observar cómo las diferentes vías de respuesta terminan siendo una imposición del "pensar tal y como lo hacen los humanos". En la segunda sección buscaré mostrar la forma en que el equilibrio refractivo puede ser usado como una herramienta para definir el concepto "pensar" teniendo en cuenta las otras mentes. Para esto, primero justificaré por qué tener en cuenta las otras mentes, posteriormente presentaré el equilibrio como mecanismo de definición conceptual para, finalmente, aplicarlo y proponer una definición inicial desde la cual es posible iniciar el proceso de refinamiento característico del equilibrio. Finalmente evaluaré y presentaré el papel que podrían jugar los computadores personales en una investigación sobre el pensamiento en estos términos así como perspectivas generales de investigación.

Todas las citas presentes en el cuerpo del trabajo son traducciones propias del inglés.

1. Turing y el origen del problema

Antes de resolver cualquier duda acerca de las capacidades de una máquina, parecería natural en primer lugar determinar los límites de aquello a lo que se está refiriendo con "máquina". En el problema tradicional de la inteligencia artificial la investigación se encuentra enfocada en el estudio de las computadoras digitales presentadas por Turing en su artículo *Computing, Machinery and Intelligence* (*CMI*, 1950). Si bien esta definición presentada por Turing se da de una manera lo suficientemente abierta como para incluir dentro de ellas gran variedad de desarrollos posteriores en la ingeniería computacional, esta perspectiva acarreará algunos problemas que serán expuestos posteriormente.

El momento histórico en el que se escribe CMI es uno en el cual las computadoras personales aún no tenían una presencia marcada en el mundo, la mayoría de máquinas computacionales pertenecían a institutos de investigación y estaban fuera del alcance del público general. Para dar a entender lo que se está comprendiendo por una computadora digital, Turing las define mediante una analogía con una máquina humana (los procesos que podría llevar a cabo una persona bajo un sistema cerrado de instrucciones)[2]; desde la cual se definen las capacidades y límites con los que se encontrará la máquina. Esta es una aplicación de la definición estricta presentada en su famoso "On Computable Numbers With an Application to the Entscheidungsproblem" (Turing, 1936) en el cual las máquinas se presentan como una herramienta para solucionar un problema propio de las matemáticas. Los computadores digitales estarán compuestos de tres partes principales "(i) almacenamiento, (ii) unidad de ejecución y (iii) control" (Turing 1950, p 437) las cuales conforman un objeto funcional. El *almacenamiento* representa el espacio en el cual los datos son presentados, la información que se va a manipular y el resultado se encuentran allí; la *unidad de ejecución* es la encargada de llevar a cabo los procesos y el *control* son las reglas fijas, presentadas a modo de tablas, que debe seguir el sistema.

Esta es una manera de introducir de manera simplificada las máquinas de Turing para el uso dentro de la teorización; la principal ventaja que trae consigo introducir estas máquinas es que para ellas se aplica la tesis Church-Turing. Bajo esta tesis se entiende que cualquier algoritmo que

[2] O, en palabras de Turing "La idea detrás de los computadores digitales puede ser explicada diciendo que esas máquinas están concebidas para llevar a cabo cualquier operación que pueda ser llevada por un computador humano. El computador humano deberá seguir reglas fijas, no tiene autoridad para desviarse de ellas en detalle alguno." (Turing, 1950, pág. 436).

sea computable (se pueda resolver de manera recursiva) será Turing-computable (podrá ser resuelto por una máquina de Turing). Lo que quiere decir esto es que cualquier proceso sistematizado, lo cual vendría ser a grandes rasgos un algoritmo, podrá ser llevado a cabo por la máquina de Turing. La definición de computabilidad puede leerse como "Aquello que puede ser calculado por un ser humano abstracto trabajando de manera rutinaria es computable" (Gandy, 1980, pág. 124), al leerse de esta manera el planteamiento de Turing se puede observar con claridad que las máquinas a las que se hace referencia en *CMI* son las mismas que introdujo para dar cuenta de los números computables.

El lograr llegar a esta claridad no quiere decir que se hayan dejado de lado los problemas acerca de la definición de las máquinas, o inclusive que sean claros los límites que tiene cada una de las definiciones. Ejemplos de esto se encuentran en la literatura posterior, tanto desde la perspectiva de las matemáticas donde tenemos desarrollos como el de Robin Gandy quien en 1980 plantea una transformación del argumento de Turing respecto de los números computables. Bajo esta nueva definición, las máquinas serían estrictamente "dispositivos mecánicos deterministas de estado discreto" [Discrete deterministic mechanical devices] los cuales plantean de manera más clara los efectos dentro de la formulación de las limitaciones de un mundo en el cual las leyes de la física tienen injerencia[3]. Existen diferentes variaciones más en la definición, cada una de las cuales será una modificación en el concepto de la máquina de Turing:

> La noción de algoritmo es más rica en estos días que en los de Turing. Y hay algoritmos […] que no están cubiertos directamente por el análisis de Turing, por ejemplo algoritmos que interactúan con el entorno, algoritmos cuyas entradas son estructuras abstractas, y algoritmos geométricos o, de manera más general, no-discretos.[4] (Blass & Yuri, 2006, pág. 31)

Dado que inicialmente el concepto de algoritmo es la base desde la cual es construida la noción de máquina en Turing, al existir modificaciones en lo que se comprende por algoritmo se tendrá un espacio diferente sobre el cual podremos entender las máquinas. Los avances en investigaciones teóricas en matemáticas traerán consigo modificaciones en lo que se comprende

[3] Inicialmente podría parecer que esto simplemente es una reformulación de aquello planteado por Turing, sin embargo, el ejercicio propuesto es encontrar las formas en que las limitaciones físicas afectan el planteamiento teórico de la máquina. Así, limitaciones sobre la cantidad de actividades que una máquina puede llevar a cabo en un solo paso o efectos de la causalidad local son tenidos en cuenta de manera explícita en las matemáticas evaluadas para este planteamiento.

[4] "The notion of algorithm is richer these days than it was in Turing's days. And there are algorithms ... not covered directly by Turing's analysis for example, algorithms that interact with their environments, algorithms whose inputs are abstract structures, and geometric or, more generally, nondiscrete algorithms." (Blass & Yuri, 2006)

por una máquina y, así mismo, en el campo de acción en el cual ellas se desenvuelven. Tomemos de la cita anterior el caso de la existencia de algoritmos los cuales interactúan con el ambiente, esto es una modificación a la manera en que se comprende el funcionamiento de una máquina. Para que una máquina pudiera entrar en acción dentro del mundo sin la existencia de esta clase de algoritmos, en ella debería poderse introducir dentro de su configuración el estado de cosas del mundo para así responder a él; el mundo no interactúa con la máquina, simplemente se comparan los diferentes inputs que de él se obtienen con la programación inicial para llevar a cabo una acción determinada. Cuando los algoritmos interactúan con el mundo esto quiere decir que, en primer lugar, los procesos pueden ser modificados por la información externa a la máquina y, más importante aún, que para la programación de una máquina no se requiere conocer la información del mundo para posteriormente programar una máquina de acuerdo con ella (esto es, no se requiere que el mundo sea computable, dado que la programación de la máquina es independiente de aquello con lo cual interactúa; la programación inicial solo requiere poder importar del mundo algunos datos, no requiere en la programación inicial de la máquina el estado de cosas en el mundo, lo primero requiere la existencia de eventos computables y que se puedan procesar en el mundo, lo segundo requiere que toda la información relevante sea computable para así poder programar la máquina). El primero de los puntos no se debe confundir con el simple uso del mundo externo como un input adicional, en este aspecto se permite que nueva información adquirida sobre el entorno se convierta en una parte operativa del algoritmo, sin que ella estuviera considerada de manera previa por el mismo. Una máquina programada de esta manera no toma el mundo como un material el cual es procesado mediante la configuración previamente establecida, la información del entorno se incorpora dentro del algoritmo generando nuevos resultados. Esta solo es una de múltiples posibilidades en como una modificación en aquello que se entiende por algoritmo puede llegar a modificar la forma en que entendemos a las máquinas y las conclusiones que se pueden extraer de que ellas lleven a cabo sus procesos en el mundo.

Por lo tanto se aprecia que una definición desde la resolución de algoritmos nos deja con un significado flexible de lo que es máquina: un cambio en la comprensión de lo que puede ser un algoritmo modifica directamente las posibilidades de acción de las máquinas. La idea general se conservará, pero debido a que no existe una exploración completa sobre aquello que significa un algoritmo –es una noción la cual se enriquece constantemente- cualquier conclusión a la que se

llegue acerca de las capacidades operacionales de una máquina (de Turing) deberá ser revisada en la medida en que se modifique nuestra comprensión de los algoritmos.

Siendo esta una aproximación funcionalista, lo importante al momento de reconocer y comprender las limitaciones de la máquina será la definición que de ella se dé, la ejecución física será algo irrelevante. Se considera que los logros en ingeniería serán aquellos que dictaminarán las limitaciones físicas de la máquina al momento de su ejecución pero no por ello se estará de manera alguna limitando lo que una máquina *puede* lograr. Esta perspectiva abre camino a dos de las problemáticas generales que serán posteriormente criticadas por escépticos del proyecto de la inteligencia artificial (el ejemplo principal que trataré en esta sección es John Searle): La importancia del medio para la identificación de facultades mentales y el mito del homúnculo. Más adelante se trataran más a fondo estos problemas.

Por el momento es necesario observar una cualidad del objeto de estudio que ha resultado evidente hasta el momento: se está trabajando con una definición blanda. Modificaciones tanto en la comprensión de las matemáticas subyacentes como en ejecuciones particulares dentro de la ingeniería traen consigo una nueva idea de lo que es una máquina. Estas son modificaciones de las cuales se deberá ser consciente al momento de trabajar el tema para tratar de manera correcta el tema y hacer justicia a los alcances de la propuesta de uno u otro autor.

Hasta el momento hemos mostrado las dificultades que existen al momento de hablar de las máquinas y usarlas como un tema de estudio mostrando la manera en que, al proponer una pregunta de investigación sobre ellas, las respuestas podrán depender de lo que se comprenda por una máquina y sus propiedades principales. Sin embargo la discusión central que se presenta sobre ellas es acerca de sus capacidades mentales: el principal foco de discusión acerca de la inteligencia en medios artificiales es, desde la primera presentación de Turing, una discusión sobre su capacidad para producir resultados análogos a los presentados por un humano; así se presenta el test de Turing. La idea detrás del test de Turing es plantear las condiciones en las cuales es posible establecer si una máquina puede ser confundida con un humano. Un criterio como este es para Turing mucho menos ambiguo y delimitable que intentar responder de manera directa la pregunta “¿puede una máquina pensar?” ya que, en este planteamiento inicial, tanto las partes que se están evaluando como la mecánica de evaluación parecen estar dadas de manera clara. El juego de la imitación proporcionará una forma de resolver la pregunta, ya que de fondo se tiene una idea conductista acerca de lo que significa “pensar”. De esta manera, si la máquina aprueba el

test de Turing, se considerará que la programación fue acertada y, por lo tanto, es representativa de la forma en que se dan los procesos mentales: “Ahora nos hacemos la pregunta “¿Qué pasará cuando una máquina tome el papel de A [el hombre] en este juego?” […] estas preguntas remplazarán nuestra pregunta original “¿Pueden las máquinas pensar?”” (Turing, 1950, pág. 434). Este método tiene el problema, que para Turing es irrelevante(Comparar con el “argumento desde la conciencia”, Turing, 1950, pág 445), de ignorar los procesos subyacentes y enfocarse exclusivamente en el output que la máquina puede generar. De esta manera se intuye en el trabajo Turing un concepto de Inteligencia en el cual ésta es una propiedad que sólo puede ser adjudicada desde la perspectiva de un observador; esta afirmación tiene respaldo en la frase de turing en CMI “Si el significado de las palabras ‘máquina’ y ‘pensar’ se encontrase examinando cómo son usadas comúnmente, es difícil escapar la conclusión de que el significado y la respuesta a la pregunta ‘¿pueden pensar las máquinas?’ se debería buscar a través de una estudio estadístico como una *Gallup Poll”* (Turing, 1950, pág. 433). El proceso establecido por Turing para definir qué se entiende por máquina, así como el establecer una forma de responder a la pregunta, busca al mismo tiempo alejarse y estar de acuerdo con el sentido común. Se busca alejar en tanto la respuesta deja de ser la primera consideración que un sujeto pueda tener, deja de ser la opinión no formada sobre el problema; se pretenden dejar claras las condiciones correctas para que se pueda emitir un juicio sobre, sin incorporar presupuestos teóricos en el evaluador, las capacidades mentales de la máquina. En lugar de resaltar la imposibilidad de reducir y determinar los conceptos, se está mostrando que es un componente importante de cada uno de ellos el criterio evaluativo que cada uno de los individuos pueda tener; esta evaluación es la forma en que el sentido común se hace presente en el proceso planteado por Turing. No por esto se convierte el concepto en algo relativo, las condiciones para afirmar o negar la propiedad (en este caso “pensar”) son parte de los usos naturales del lenguaje y, se asume que, quienes lleven a cabo el juicio podrán aplicarlas para discernir si es o no correcto llamar a la máquina pensante. La evaluación por individuos no trae consigo la individualización del juicio que tendría una encuesta directa, se presupone que un acto como llevar a cabo una conversación será algo que no dependerá de las condiciones particulares del evaluador. Al establecer de esta manera a los sujetos como jueces no se está imponiendo las consideraciones particulares, se está comparando la máquina con un humano en un comportamiento que es común a toda la especie y, por lo tanto, cualquier sujeto es un buen juez para determinar si la máquina cumple con el juego de la imitación, observando la naturalidad del acto y emitiendo un juicio desde el sentido común. Puede

que en ningún momento del proceso se hagan explícitas las cualidades que deben reconocerse al momento de decir que la máquina piensa, pero se considera la comunicación como un acto característico en el cual un humano cualquiera podrá identificar a una contraparte competente y, por lo tanto, considerarla como pensante.

1.0. Formas de respuesta.

Después de este planteamiento inicial las críticas en filosofía hacia el planteamiento de Turing se han presentado de dos modos principales: Cuestionando la validez del juego de la imitación como un criterio para asignar la propiedad de pensar a las máquinas, y establecer que debido a las limitaciones de las máquinas estas no podrán superar el test. Claramente cada una de estas aproximaciones se da desde presupuestos diferentes, el primero de ellos es un cuestionamiento sobre los requisitos necesarios para catalogar a un sistema como pensante, el segundo es un planteamiento sobre las incapacidades técnicas de las máquinas. Uno de los representantes más conocidos de las dudas del primer tipo en la tradición filosófica es John Searle, quién, desde su crítica establecida en "Minds, Brains and Programs" (MBP), con el argumento de la Habitación China busca mostrar las razones por las cuales el producir un comportamiento que apruebe el test no será suficiente para decir que la máquina piensa. Searle mismo es también un ejemplo de los argumentos del segundo tipo con el ataque que hace hacia el cognitivismo en "Is the brain a digital computer" (Searle, 1990) o en *The Rediscovery of mind* (Searle, 1994), donde busca, mostrando cómo el cerebro no funciona como un computador digital, mostrar por qué la máquina no podrá llevar a cabo las labores que este realiza. Esta crítica manifestada por Searle está lejos de ser la única, pero al estudiar la forma en que se producen estos argumentos y revisar la forma en que otros funcionan podremos observar una noción común al momento de investigar el pensamiento en las máquinas: el pensar se torna equivalente al pensar humano.

1.1. El criterio humano en la habitación china.

La discusión que John Searle plantea con la posibilidad de que una máquina pueda pensar se puede leer en dos momentos principales: en primer lugar en su crítica presentada en MBP y posteriormente en su libro *The Rediscovery of Mind* (Searle, 1992) (Particularmente el capítulo 3 "Breaking the Hold: Silicon Brains Conscius Robots and Other Minds). Si bien existen elementos comunes en ambas etapas de la discusión, el trabajo inicial de Searle es presentar la crítica sobre

aquello que encuentra problemático en la lectura tradicional del problema mientras que en la segunda etapa es donde construye una teoría para dar cuenta del mismo. Desde el primer momento en que se plantea el argumento de la Habitación China la respuesta dentro de la comunidad académica ha sido extendida y aún permanece activa (e.g. Donald Nute, 2011); sin embargo más que reconstruir estás para buscar como Searle se encuentra en lo correcto o se equivoca, en el presente texto se buscará hacer explícito el peso que cobra dentro de este debate nuevamente la perspectiva humana.

En MBP Searle abre su crítica definiendo el campo hacia el cual se encuentra dirigida, para esto establece la división entre la AI fuerte y la AI débil:

> De acuerdo a la AI débil, el principal valor del computador en el estudio de la mente es que nos brinda una herramienta poderosa. Por ejemplo, nos permite formular y probar hipótesis de forma más precisa y rigurosa. Pero de acuerdo con la AI fuerte, el computador no es meramente una herramienta para el estudio de la mente; en lugar de esto, el computador programado de manera apropiada es una mente, en el sentido que, de los computadores con los programas adecuados, puede decirse literalmente que entienden y poseen otros estados cognitivos. En la AI fuerte, debido a que el computador programado posee estados cognitivos, los programas no son meramente herramientas que nos permiten probar explicaciones psicológicas; en lugar de esto, los programas son en sí mismos las explicaciones. (Searle J. , 1980, pág. 2)

En este sentido las máquinas que aprueben el test de Turing, aceptándolo como criterio para determinar la capacidad de una máquina para pensar, se considerarían como pertenecientes a la AI fuerte; los programas diseñados de tal manera que aprueben el test serían en sí mismos explicaciones de cómo llevamos a cabo estas tareas mentales. Por otra parte máquinas diseñadas para llevar a cabo una tarea (por más compleja que sea) siempre y cuando no se le intenten asociar propiedades cognitivas serán consideradas como AI débil. En general, cuando la máquina esté diseñada de tal manera que se pretenda responder de manera afirmativa la pregunta "¿Esta máquina piensa?" será considerada como un AI fuerte.

Searle considerará problemática la tarea de la AI fuerte precisamente debido a la pretensión que ella tiene de catalogar a las máquinas como pensantes. La crítica no se plantea desde los resultados que pueda arrojar un programa particular, la crítica cae en la forma en que ellos funcionan. Para Searle la capacidad de manipulación sintáctica por parte de la máquina no quiere decir que esta tenga un conocimiento debido a la ausencia de semántica. Esta dificultad será la que Searle ejemplificará mediante su argumento de la Habitación China, el cual presenta las dificultades que se tendrían para adjudicar un proceso de pensamiento a lo ocurrido en un procesador.

El argumento se aplica a los computadores digitales entendidos de la misma manera en que son entendidos en CMI, esto es, mediante la analogía con la máquina humana. En el caso de Searle, estará compuesto por una unidad de almacenamiento que es el papel donde se introducen las preguntas y se escriben las respuestas; el lenguaje en el que estarán escritas será en chino. La unidad de control será una persona introducida en una habitación cuya única manera de comunicarse con el mundo es mediante el almacenamiento, esta persona no tiene conocimiento previo de chino y solamente habla inglés. La unidad de control será un grupo de manuales los cuales indican la manera en que se pueden modificar los caracteres que son introducidos en el sistema para producir las respuestas adecuadas cada vez que se introduce una pregunta; estos manuales se encontrarán escritos en inglés para que puedan ser comprendidos por la unidad de ejecución (Searle J. , 1980, pág. 418) Si bien el autor no asigna directamente esta correspondencia con los elementos de una computadora digital, la intención de asimilar la habitación china con una computadora es evidente y acá lo único que hago es poner en explícito los nombres de cada componente. La pregunta, que Searle responde de manera negativa, es "¿Al producir respuestas en chino *conoce* la persona de la misma manera en que lo hace cuando responde en inglés?", la cual se responde no mediante una demostración, sino mediante un argumento que es convincente[5].

> Una formalización posible del argumento que plantea Searle es:
> La historia del "argumento de la habitación china" parece contener el siguiente argumento:
>
> 1. Quien ocupa la habitación no sabe chino.
> 2. Quien ocupa la habitación sabe Inglés.
> 3. A quien ocupa la habitación se le entregan paquetes de piezas escritas en chino, {Ci, Cj,…, Cn}.
> 4. A quien ocupa la habitación se le entregan instrucciones formales en inglés que correlacionan parejas de piezas en chino, hCi, Cji.
> 5. A quien ocupa la habitación se le entregan instrucciones formales en inglés para entregar algún Ci en particular dado un Cj particular.
> 6. La habilidad de quién ocupa la habitación para manipular sintácticamente las piezas de chino es comportamentlamente indistinguible de aquella de un hablante completamente competente de chino.
> 7. Si 1–6 son posibles en conjunto, entonces la sintaxis no es suficiente para el contenido mental.
> 8. 1–6 Son posibles en conjunto.
> 9. Por lo tanto, la sintaxis no es suficiente para el contenido mental.[6]

[5] "No he demostrado que esta afirmación sea falsa, pero ciertamente parecería una afirmación increíble en el ejemplo" (Searle J. , 1980, pág. 4)

[6] "The CRA story appears to contain the following argument:/ 1. The room occupant knows no Chinese./ 2. The room occupant knows English./ 3. The room occupant is given sets of written strings of Chinese, {Ci, Cj,…, Cn}/ 4. The room occupant is given formal instructions in English that correlate pairs of sets of Chinese strings, hCi, Cji./ 5. The room occupant is given formal instructions in English to output some particular Ci given a particular Cj./ 6. The room occupant's skill at syntactically manipulating the strings of

Esta formalización es especialmente explicativa en los primeros seis puntos donde resume de manera efectiva las solicitudes que se tienen para llevar a cabo la argumentación. La forma en que esta formalización completa el argumento de la habitación china no es estrictamente igual a como lo plantearía Searle, aún así deja claras las intenciones del argumento. El punto 7 en particular podría ser considerado como el más complicado, pero éste presenta la solicitud principal del argumento: que las premisas anteriores dan las condiciones de una operación netamente sintáctica de la información. Por otra parte es problemático decir que el argumento se centra en la suficiencia de la sintaxis para generar contenido mental, Searle en MBP constantemente recuerda cómo su argumento está enfocado al "entendimiento" ("Does it even provide a necessary condition or a significant contribution to understanding?" (Searle J. , 1980, pág. 418)), extender el argumento a toda clase de contenido mental es carecer de cuidado por los límites (sobre todo si no se presenta una definición de que se entenderá por contenido mental). Teniendo en cuenta estas salvedades, esta formalización es una manera eficiente de mostrar la forma en que se presenta el argumento de la habitación china.

Este argumento es una respuesta directa al Test de Turing. En él se está mostrando cómo si se presenta un sistema el cual apruebe dicho test, aún así esto no es garantía de que él posea conocimiento. "El ejemplo [la habitación china] muestra que puede haber dos "sistemas", los cuales ambos aprueban el test de Turing, pero sólo uno de ellos entiende; y no es un argumento contra este punto decir que, ya que ambos aprobaron el test de Turing, ambos deben entender, dado que esta afirmación no responde al argumento de que el sistema en mí que entiende inglés tiene un entendimiento mucho mayor al de aquel que solo procesa el chino."[7] . La comparación entre un sujeto que aprende a hablar chino y responde mediante la forma tradicional en que lo hacen los humanos, y uno que se encuentra en la habitación permite observar la diferencia existente entre dos objetos que, al ser evaluados, pasarían el test de Turing.

Chinese is behaviorally indistinguishable from that of a fully competent speaker of Chinese./ 7. If 1–6 are jointly possible, then syntax is not sufficient for mental content./ 8. 1–6 are jointly possible./ 9. Therefore, syntax is not sufficient for mental content." (Nute, 2011, págs. 431-432)

[7] "The example [the Chinese room] shows that there could be two "systems," both of which pass the Turing test, but only one of which understands; and it is no argument against this point to say that since they both pass the Turing test they must both understand, since this claim fails to meet the argument that the system in me that understands English has a great deal more than the system that merely processes Chinese" (Searle J. , 1980, pág. 6)

La fuerza del argumento se da en una intuición común[8], al encontrarnos en la posición de la unidad de control consideraríamos que, con claridad, no sabemos chino. En este caso se entiende que una persona sabe un idioma cuando lo maneja en la misma manera en que un hablante nativo podría manejarlo, y se asume que la distinción es clara para quien experimenta la diferencia; esto será en general lo que se considera es entender el idioma. En este punto se observa la forma en que el punto de evaluación nuevamente es aquello que un humano puede hacer, solo que en vez de una comparación directa con las personas para poder emitir un juicio (cómo se haría en el test de Turing) se usan las características de la experiencia individual como criterio evaluativo. El entendimiento se presenta en las condiciones personales, las consideraciones que como humano tenga acerca de la presencia o ausencia de conocimiento serán una condición suficiente para decir que allí se presenta el conocimiento. Así, al ponerse en el lugar de la máquina y evaluar si ella tiene un conocimiento la pregunta deja de ser exclusivamente sobre el funcionamiento de máquina; ahora se cuestiona principalmente es por la posibilidad de un humano, al encontrarse en la posición de la máquina, de tener un conocimiento.

De este modo, al ser una respuesta directa, tanto Turing como Searle recurren a un mismo criterio para evaluar si una máquina piensa: la evaluación por un observador humano. En el caso de Turing, su Test requiere directamente de tres evaluadores para dar un parte acerca de las habilidades de la máquina, por el lado de Searle, el lector debe dictaminar su postura sobre lo que ocurre dentro de la habitación china. Aún así lo que cada autor solicita del evaluador humano es totalmente distinto, el primero pide que lleve a cabo una actividad cotidiana para, posteriormente, dar su opinión acerca de lo normal que ésta fue; el segundo pide imaginarse en la posición de la máquina y emitir un juicio de acuerdo con aquello que podría considerar en esas condiciones. Pese a esta sutil diferencia, se puede rastrear que en ambos casos el criterio principal desde el cual se hace referencia al pensamiento es mediante una comparación con las capacidades humanas.

[8] Sobre la forma en que Searle construye su argumento basado en intuiciones se puede observar la crítica planteada por Dennett (Dennett, 1991, págs. 435,440) donde llama el argumento de la habitación china una “bomba de intuiciones”. Una “bomba de intuiciones” solicita al lector que se imagine un escenario en el cual estaría dispuesto a aceptar con facilidad, después se realizan modificaciones y, sin garantizar que el nuevo escenario se comprende del todo, aún se extraen conclusiones de las intuiciones iniciales. De este modo, en la habitación china se extiende la intuición inicial, el que el habitante no comprende chino, a sistemas cada vez más complejos sin evaluar realmente como estos nuevos sistemas podrían modificar la idea que tenemos sobre el entendimiento en, o de, la habitación.

Si este criterio es o no válido es una discusión que puede extenderse[9], el debate es abierto por Searle en MBP al presentar posibles objeciones a su argumento y cómo estas se pueden contestar, sin embargo, la validez o no del argumento de la habitación no es lo que concierne al presente artículo. Lo que quiero resaltar hasta el momento es cómo, en una primera aproximación, el plantear un argumento como el de la habitación china hace que la discusión siga perteneciendo exclusivamente a "pensar como lo hacen los humanos". Hasta el momento el argumento de la habitación china nos lleva a considerar nuevamente la relevancia de la experiencia humana al momento de evaluar si el objeto estudiado puede ser considerado como uno que piense.

1.2. La discusión del cognitivismo.

Searle construirá una nueva forma de su argumento, el cual se puede ver claramente expresado en el capítulo 9 de *The Rediscovery of Mind*. "En este capítulo me referiré a 1[¿Es el cerebro un computador digital?] y no a 2 [¿Es la mente un programa computacional?]. En escritos anteriores he dado una respuesta negativa a 2 debido a que los programas son definidos puramente sintáctica o formalmente" (Searle J. , 1994, pág. 200). Las dificultades que ha intentado señalar Searle hasta el momento con el argumento de la habitación china son, por lo tanto, parte de la pregunta acerca de la mente; por esto se ha enfatizado en la posibilidad de comprender los posibles contenidos mentales que una persona podría tener en el papel de la máquina. El nuevo argumento no está dirigido hacia los defensores de la AI fuerte; se encuentra dirigido hacia los defensores del cognitivismo ("la perspectiva de que el cerebro es un computador digital" (Searle J. , 1994, pág. 202)). De esta manera el proyecto de la AI quedaría relegado a la AI débil ya que si el cerebro no funciona como un computador digital y tampoco la mente es un programa, entonces el papel de la AI quedará relegado a simulaciones de procesos más no a la reproducción de los mismos.

La forma de argumentar este punto por parte de Searle depende de la aceptación de una división que marca sobre la las propiedades que son intrínsecas a los objetos y las que son relativas al observador. Esta es una cuestión que por sí misma ha desencadenado una discusión[10],

[9] Para un buen resumen de algunas de las respuestas presentadas al argumento de la habitación china puede observarse la presentación hecha en *Searle: Contemprary Philosophy in Focus* "The Thought Experiment Debate" (Moural, 2003)

[10] Se puede encontrar una discusión en "Searle's Misunderstandings of Functionalism and Strong AI" (Rey, 2002) donde se plantea cómo, a pesar de que los símbolos no se encuentran definidos en la física,

aún así, más que adentrarme en los pormenores de la discusión, quisiera acercarme a las consecuencias que puede tener el que el argumento sea o no correcto para responder la pregunta original "¿pueden pensar las máquinas?".

Como se puede observar en el planteamiento del problema, la discusión en este caso no es acerca de las capacidades de una máquina. En este caso lo que se está trabajando es una pregunta sobre el paralelo existente entre las explicaciones computacionales y los fenómenos presentes en el cerebro. Las cuatro principales dificultades que señala Searle en el capítulo mencionado son "La sintaxis no es intrínseca a la física, […] La falacia del homúnculo es endémica al cognitivismo, […] la sintaxis no tiene poderes causales [y][…] el cerebro no lleva a cabo procesamiento de información" (Searle J. , 1994, págs. 207-222). Mediante estos, más que resaltar las incapacidades para pensar de una máquina, se busca distanciar a las máquinas como forma de explicación del cerebro humano.

El primero de los puntos señala como a pesar de que en la computación cualquier cosa puede ser *tomada* por un 0 o un 1, este tipo de contenidos no son intrínsecos al espacio material. Puede encontrarse que exista una estructura isomorfa en el cerebro respecto a las descripciones computacionales, pero ésta equivalencia en la descripción de la estructura no es una garantía suficiente por la división entre propiedades intrínsecas y propiedades relativas al observador. Las propiedades intrínsecas son aquellas que son propias de los objetos, como la carga electromagnética, y de las cuales se ocupan las ciencias naturales, por otra parte están las propiedades relativas al observador, como por ejemplo el que algo sea un símbolo, las cuales no se derivan exclusivamente de las propiedades físicas; para Searle la sintaxis es una propiedad de este último tipo y, por lo tanto, la equivalencia estructural no es suficiente. Así, debido a que la sintaxis es una propiedad que depende de los observadores surge el segundo problema, la inclusión de un homúnculo quien hace el papel de observador en las descripciones computacionales, este homúnculo será la única forma en que se puede presentar la sintaxis en los sistemas computacionales. El tercer punto es quizá el más complejo, pero Searle busca con su argumentación demostrar que un programa implementado, es decir una estructura sintáctica, no tendrá poderes causales más allá de los del medio en el que se encuentra y por lo tanto el cognitivismo no permite explicar correctamente las operaciones causales del cerebro; dado que la

esto no significa que sean relativos al observador. Por otra parte se puede encontrar una reconstrucción general de los puntos más importantes del debate en el artículo "Searle, Syntax and Observer Relativity" (Endicott, 1996).

sintaxis no existe en el campo físico, una vez se reconoce la forma en que el medio se comporta específicamente (esto es, el comportamiento neuro-biológico) la explicación computacional se torna irrelevante. Esto introduce el cuarto punto, este busca contrarrestar a aquellos defensores del cognitivismo quienes manifestarían que, a pesar de esto funcionar en todos los demás sistemas, el cerebro tiene como propiedad intrínseca ser un procesador de información y que, por lo tanto, las descripciones y simulaciones computacionales que se hagan de él podrán informar sobre los procesos subyacentes. Pero lo que ocurre en el cerebro y en un computador son situaciones distintas, mientras que el último está programado para producir un resultado específico el cuál puede ser leído, el segundo por su parte genera un "evento consciente concreto y es producido en el cerebro por procesos bilógicos electroquímicos específicos"[11] (Searle J. , 1994, pág. 225). En el cerebro cada evento sensorial no es procesado para obtener un resultado concreto, existe es una experiencia la cual se deriva de la forma en que los eventos físicos estimulan e interactúan con la estructura biológica. De esta manera se considera que las dificultades al momento de establecer un paralelo entre el cerebro humano y un sistema computacional son tales que el proyecto del cognitivismo se encuentra equivocado desde un principio. Ahora bien, ¿qué efectos tiene el que este argumento sea correcto o incorrecto sobre la pregunta inicial?

Si este argumento es correcto, la implicación es la negación del cognitivismo. Esto es, responder con un claro "no" la pregunta que formuló Searle "¿Es el cerebro un computador digital?"; pero esta es una pregunta en la cual sólo se habla de las capacidades del cerebro, no sobre las capacidades de las máquinas. Si la argumentación es incorrecta, puede que el proyecto del cognitivismo sea o no viable; la inviabilidad de la argumentación de Searle no garantiza la posibilidad de la paridad entre la computación y la neurobiología, esta es una etapa del proyecto del cognitivismo el cual aún cargaría con la prueba y necesitaría mostrar efectivamente esta paridad. En general, este es un argumento que sólo funciona para determinar lo correcto de una forma de describir el cerebro, este no es un argumento que pueda funcionar para poder decir algo sobre las máquinas. Además, como veremos en la segunda sección de este texto, es posible construir un sistema en el cual se evalúe la pregunta acerca de las máquinas sin por esto tener que relacionarlas con el cerebro humano.

[11] "Concrete conscious event and is produced in the brain by specific electrochemical biological processes".

1.3. Incapacidades de la máquina en general.

Existen múltiples argumentos que buscan mostrar que las máquinas se encuentran demasiado limitadas y, por lo tanto, el proyecto de la AI fuerte está condenado. Ejemplos que podemos encontrar de argumentos de este tipo son "El problema del marco" (Pylyshyn, 1987) o el teorema de la incompletitud de Gödel aplicado a la inteligencia artificial (Penrose, 1989). El primero es un problema sobre la capacidad de la máquina para tomar decisiones y priorizar la información dado un set de reglas estable. Al momento de tomar una decisión un robot debe poder discernir de la información disponible en su entorno (por ejemplo, el clima, el color de la habitación, los días faltantes para el solsticio) aquella que es relevante y, además también es necesario que cuente con la programación para el caso particular. El segundo manifiesta la incapacidad de una máquina para producir un tipo de conocimiento que un humano sí podría obtener; se supone que algunas de las paradojas en las cuales se sustenta el teorema de la incompletitud de Gödel no podrán ser conocidas ni resueltas por una máquina mientras que una persona podría acercarse a ellas. A pesar de que en la discusión "una variedad de soluciones funcionales al problema lógico del marco han sido desarrolladas y éste ya no es considerado un obstáculo serio, aún para aquellos trabajando un paradigma basado estrictamente en la lógica"[12] (Shanahan, 2009), esto sólo indica que ha sido posible continuar con los ejercicios de programación después del planteamiento del problema; las inquietudes epistemológicas que este argumento despierta son un tema activo de discusión. Por su parte, el argumento de Penrose presenta una discusión viva[13] sobre su adecuación y métodos para la formalización matemática..

Pero aun cuando estos argumentos buscan mostrar cómo las máquinas son incapaces de llevar a cabo un proceso que consideramos inteligente, el criterio mediante el cual se evalúa sigue siendo netamente de comparación directa con los humanos. Al intentar mostrar la incapacidad que tienen las máquinas para llevar a cabo ciertos procesos se está, al mismo tiempo, estableciendo una comparación con la forma en que los humanos pensamos nuestro mundo. Por ejemplo, si se encuentra una limitación en las capacidades lógicas de una máquina, esta solo será una limitación que afectará su status como pensante o no en el caso de la limitación no esté

[12] "a variety of workable solutions to the logical frame problem have been developed, and it is no longer considered a serious obstacle even for those working in a strictly logic-based paradigm (Shanahan 1997; Reiter 2001; Shanahan 2003)." (Shanahan, 2009)

[13] Una referencia de la discusión que ha traído consigo este argumento puede encontrarse en (Megil, 2012), para un ejemplo de una réplica en la cual se explican las matemáticas subyacentes, así como diversas opciones para atacar el problema planteado por Penrose se puede ver (Wang, 2007, págs. 250-253).

presente en los humanos. Así, un argumento de esta clase es relevante sólo en la medida en que se considera que una persona, en condiciones normales, tiene la capacidad de encontrarse por encima de las limitaciones que estos problemas plantean. El problema del marco es relevante en la medida en que una persona al entrar en una habitación donde hay una bomba puede tomar con rapidez una decisión acerca de qué acción seguir. Si, por el contrario, el problema del marco nos mostrase una situación en la cual ni el humano ni el robot pudieran reaccionar con agilidad o una situación en la que el humano no puede actuar de manera efectiva mientras que la máquina lo hace, este no pasaría a ser un argumento en contra -o a favor- de la capacidad de pensamiento de ninguno de los dos. Esto mismo ocurre con el teorema de la incompletitud, solamente estamos dispuestos a aceptarlo como un argumento en contra de la capacidad de pensar de las máquinas en la medida en que nos encontramos en una posición privilegiada. De hecho, una de las contra-argumentaciones que se presentan a Penrose es mostrar cómo, a pesar de que las máquinas se encuentran limitadas por el teorema de la incompletitud, los humanos también nos encontramos limitados por él (Wang, 2007, pág. 252). De esta manera las limitaciones procuran mostrar cómo, a pesar de que los sistemas de AI pueden llegar a desarrollarse hasta puntos altamente funcionales, existirán siempre barreras que, al momento de producirse la comparación con los humanos, dejarán a las máquinas en una posición desfavorecida lo que hará que no puedan ser consideradas como pensantes.

Con esto no pretendo en ningún momento afirmar o negar la validez de los argumentos tratados, este no es un ataque hacia ellos. Tampoco pretendo decir que *toda crítica* funcione de esta manera, simplemente que es una tendencia que se observa con claridad en aquellos que son quizá los argumentos más representativos al momento de acercarse a la discusión filosófica de la AI. Es claro que en la discusión el introducir a los humanos como criterio de evaluación ya sea de una manera directa, como en el caso de Turing y Searle, o de una manera indirecta como cuando se critican las capacidades de las máquinas, es una constante. Y no por esto se está cometiendo un error, claramente la información que tenemos sobre la experiencia de pensar como humanos se encuentra en un lugar privilegiado respecto a la de los demás sujetos. Sin embargo, todas las discusiones que se presenten de esta manera terminarán por incluir el presupuesto no explícito "pensar es pensar tal y como lo hacen los humanos" y no deberíamos olvidar que la pregunta que nos trajo a toda esta discusión era una simple "¿pueden pensar las máquinas?".

2. Introduciendo una nueva vía.

Como he querido mostrar hasta el momento, existe un problema en la justificación de la AI, a saber, se ha transformado la pregunta "¿pueden pensar las máquinas?" en "¿pueden pensar las máquinas tal y como lo hacen los humanos?". No es evidente inmediatamente que esto sea un problema, es posible justificar este cambio en la pregunta en el hecho de que somos humanos llevando a cabo una investigación y, por lo tanto, sólo contamos con nuestra experiencia para establecer un criterio evaluativo. Para ejemplificar cuándo tomar esta actitud podría ser problemático narraré la siguiente historia:

Suponga que existe una especie extraterrestre la cual nos ha observado desde largo tiempo atrás y, gracias a esto, ha logrado construir un mecanismo mediante el cual puede realizar traducciones entre su lenguaje y el nuestro. Un día, al mejor estilo de *La guerra de los mundos*, encontramos en la plaza central un robot gigante ubicado debajo de lo que reconocemos como una nave espacial. Sólo cuando los medios se hacen presentes en la escena el robot emite la frase "¿Atacarán los líderes terrestres?", después de esto hay silencio, cada día lo único que se repite es esta misma pregunta. Finalmente diversas naciones se pronuncian, ellas producen comunicados extensos en donde algunas se comprometen incondicionalmente a conservar la paz, otras plantean algunas condiciones que deberán ser cumplidas para evitar los ataques, Corea del Norte promete aliarse con los extraterrestres en caso de que ellos quieran tomarse el mundo. Cada una de estas promesas se hacen y el robot guarda silencio hasta que la mayor parte de las naciones han hablado; posteriormente se escucha un nuevo comunicado por parte del robot: "No atacaré, ¿Preguntas?".

Los medios presentes bombardean inmediatamente con preguntas, pero toda respuesta que se obtiene es un "Sí", un "No" o, en el mejor de los casos, un verbo. Por ejemplo, ante la pregunta "¿qué hacen acá?" la respuesta es "investigar". Las naciones envían también embajadores para intentar establecer de manera directa contacto con el robot, esto no cambia en modo alguno la forma de responder del aparato; esto despierta en las grandes potencias dudas, a pesar de que toda respuesta había sido congruente, y empiezan a estudiar de formas no invasivas a nuestro visitantes. Los primeros registros muestran que solamente existen comunicaciones entre la nave y el robot; no hay ninguna clase de onda que se esté emitiendo al espacio (y éste no parece estar distorsionado); el robot parece funcionar mediante componentes existentes en la tierra, incluso algunos de los chips se asemejan a arquitecturas desarrolladas con fines militares que aún no se habían hecho públicas; la nave está hecha de un material desconocido pero basado

en arsénico, lo único que se aprecia es actividad lumínica dentro de ella, pero no hay rastros de calor ni parecen existir tripulantes.

Intrigadas, las naciones empiezan a hacer preguntas del tipo "¿Algún miembro de su especie se encuentra en este planeta?", a lo cual siempre se responde afirmativamente; ante las solicitudes de observar a dicho alienígena el robot responde "hecho". La tensión crece, los gobernantes sienten que esta última respuesta muestra que los alienígenas se encuentran dispuestos a mentir de una manera abierta. Se pierde la confianza en las respuestas anteriores, en particular, se cree que es probable que al declararse como inofensivos se está simplemente ganando tiempo para estudiar un plan de ofensiva contra la tierra; ¿Qué otra razón podría motivar a una especie a enviar solamente máquinas de reconocimiento y no permitir establecer alguna forma de contacto directo con ellos? El gobierno americano decide entonces presionar el botón rojo, tan pronto como se produce el ataque y justo antes de que el robot y la nave sean destruidas se observa por primera vez una señal que de ellos se dirige hacia el espacio.

Los estudios sobre los residuos del ataque no muestran señales de la existencia de algo más allá de lo mostrado por los análisis previos. Unos años después más naves aparecen, todas con ligeras variaciones en sus formas, el ataque comienza, nuestro sistema automático de defensa es usado como la primera forma de ataque contra la humanidad. La guerra dura un par de días y la vida en el planeta se ve fuertemente afectada, pero sólo los homo-sapiens-sapiens son exterminados. Una nueva forma de vida se encuentra en el planeta, un habitante de otra galaxia que ha encontrado un nuevo hogar sólo se comunica en estados discretos: solo se plantean preposiciones sencillas con valores de verdad determinados, los puntos intermedios y los condicionales no hacen parte de su forma de habla –a pesar de ser parte de su proceder cotidiano- y no tiene la capacidad de comprender más allá que los contenidos explícitos dentro de las comunicaciones. No conoce el concepto de la mentira, no tiene la capacidad siquiera de imaginar que las verdaderas intenciones puedan ocultarse al momento de comunicarse. Mediante su tecnología el mundo retorna a su estado previo, solo que ahora los humanos no somos parte de él. Una vez concluido este proceso la mayoría de robots y naves pasan a la inactividad, sólo uno de ellos continúa moviéndose y observando un planeta cuya especie tecnológicamente más avanzada fue incapaz de convivir con él y por lo tanto él debió actuar por su cuenta para poder observar aquél mundo que le agradaba.

En este ejemplo lo que ha ocurrido es un conflicto entre formas de inteligencia: por una parte nos encontrábamos los humanos y nuestra forma característica de pensar; por la otra una

especie que solamente llevaba a cabo sus interpretaciones de acuerdo con acciones, descripciones simples y dos valores de verdad, una especie que, además, está compuesta por un único individuo (todas las naves atacantes eran parte de una misma mente). Lo aquí presentado se encuentra dentro del campo de lo posible, ninguna ley natural se ve rota en el ejemplo planteado. Además, ante la pregunta "¿Puede pensar esta especie?" la respuesta, creo, siempre sería afirmativa. Varias razones se pueden encontrar para esto: la especie puede manejar nuestra tecnología, puede comunicarse efectivamente con nosotros, tiene capacidades de viajar en el espacio, afecta planetas enteros mediante sus acciones y establece planes de acción organizados hacia objetivos. Sin embargo se hace evidente que esta especie no posee una consciencia del mismo modo en que la poseen los humanos, las experiencias que él pueda tener serán completamente diferentes a las humanas al tener acceso a más de una expresión física de manera simultánea; tampoco se hará presente una equivalencia ni en la forma en que se organiza la información ni en la forma en que se exterioriza, a pesar de que no podamos acceder a lo que sucede en la mente del alien, la narración permite vislumbrar lo que allí ocurre. De esta manera tanto los estados internos como los externos del alienígena serán altamente diferentes de la experiencia humana, una diferencia lo suficientemente grande como para impedir que, de entrada, podamos comprenderlo. En otras teorías (en especial, como vimos en la sección anterior, la de Searle) estas razones serían suficientes para considerar que esta especie carece de consciencia y que los procesos que llevan a cabo no están dentro de aquello que llamamos "pensar": esta especie no piensa al no hacerlo como lo hacen los humanos.

Pero negar desde la propia definición que puedan existir otras formas de pensar diferentes a la humana puede traer inconvenientes: al asumir que toda forma de inteligencia funciona de la misma manera que la humana consideraremos inmediatamente que todo otro sujeto pensante lo hará a nuestro modo; además al asumir esto no se genera el espacio para intentar comprender otra mente pensante, se presupone que ella piensa y que por lo tanto ya tenemos la información necesaria –la información privilegiada de la perspectiva humana- desde la cual establecer cómo lo está haciendo. Nos enfrentamos de esta manera a dos problemas, por una parte juzgar que todo lo que piense tiene que hacerlo en la forma que los humanos lo hacemos y, por la otra, la inexistencia de las herramientas desde las cuáles sea posible comprender otras formas de pensamiento; estos son problemas relacionados, en la medida en que se presente una solución a uno de ellos se observará una salida para el otro, las herramientas llevarían a considerar otras formas de pensar y el deseo de tener en cuenta estas formas lleva a la construcción de las

herramientas. Se imponen las condiciones contingentes en las cuales se presentó la evolución humana como única posibilidad al momento de definir un concepto como pensamiento y, al estar seguros de que esto es así, parece dejarse a un lado la investigación de otras posibilidades[14] y del alcance que puede llegar a tener. Puede que en un estudio posterior, o en una narración como la hecha acá, se pueda comprender[15] la forma en que este individuo piensa, pero si mantenemos como presupuesto el criterio de evaluación humano, entonces nunca se dará la posibilidad de estudiarle. En el ejemplo es el fin de la especie -no por negar la inteligencia de esta otra especie, sino por proyectar en ella formas de pensar humanas- pero, siendo menos dramáticos, es posible encontrar oportunidades perdidas debido a esta restricción. En las siguientes secciones procuraré mostrar una forma en que se podría construir una definición de "pensar" la cual no esté fundamentada exclusivamente en la forma en que los humanos pensamos, y que permita abrir el camino para construir herramientas para comprender mentes no humanas. Después de esto mostraré algunas de las ventajas que puede tener el adoptar esta postura.

2.1. El Equilibrio refractivo[16] de Goodman.

Nelson Goodman en *Fact, Fiction and Forecast* propone una forma de justificar las inferencias, tanto inductivas como deductivas. En general, la propuesta de Goodman consiste en que se puede extraer de la práctica información suficiente como para justificar las reglas y que, al mismo tiempo, las reglas de la inferencia modifican la práctica a medida que limitan los procesos que se consideran correctos.

> He dicho que las inferencias deductivas se encuentran justificadas por su conformidad a reglas válidas generales, y que las reglas se encuentran justificadas por su conformidad a las inferencias válidas. Pero este es un círculo virtuoso. El punto es que las reglas y las inferencias particulares se encuentran ambas justificadas si se llevan a un acuerdo entre ellas. *Una regla se enmienda si trae consigo una inferencia que no estemos dispuestos a aceptar; una inferencia se rechaza si ella viola una regla que no estemos dispuestos a enmendar*. El proceso delicado de justificación es aquel de hacer ajustes

[14] Por ejemplo Searle afirma que "Mi propia perspectiva es que *solamente* una máquina puede pensar, y efecto solo tipos bastante especiales de máquinas pueden hacerlo, es decir cerebros y máquinas que tienen los mismos poderes causales que los cerebros" (Searle J. , 1980, pág. 424).

[15] Pueden existir mentes radicalmente distintas de la humana las cuales no podamos comprender, pero estas tampoco las podré plantear inicialmente dado de que soy un humano intentando imaginarlas.

[16] Si bien la traducción tradicional de "Reflective Equilibrium" es "Equilibrio Reflexivo", considero que es más adecuado referirse a él como refractivo. Esta última palabra trae a la mente la imagen de superficies que se reflejan, como lo hacen las partes en equilibrio. Por la otra, considero que la palabra "reflexión" trae a la mente la idea de que cada una de las partes de equilibrio pueden auto-regularse y modificarse libremente. En sí, la ventaja que veo de la refracción sobre la reflexión es que la primera invita a ir de un lugar a otro, mientras que la segunda puede quedarse en un solo lugar.

> mutuos entre las reglas y las inferencias aceptadas; y en el acuerdo logrado yace la única justificación que se necesita para ambos. (Goodman, 1955, pág. 64)

Así, para Goodman, las reglas de la inferencia y las inferencias válidas se encuentran justificadas mutuamente entre ellas. Existen inferencias las cuales son evidentemente correctas, una regla que no permita que una inferencia de este tipo sea considerada como válida será una regla que esté en contra del sentido común y, por lo tanto, deberá ser reformulada. Esto aplica también sobre las reglas: existen reglas las cuales se encontrarán tan arraigadas o validadas por la experiencia que no estaremos dispuestos a negarlas por una inferencia contradictoria, en este caso se negará la inferencia y no se modificará la regla. Es de suma importancia observar que las reglas como tales no deben ser rechazadas o eliminadas, en lugar de esto deben ser enmendadas. Para llevar a cabo un ejercicio de equilibrio refractivo deben existir tanto reglas como inferencias dadas; se presupone que si una regla ya se encuentra presente es debido a que ella da cuenta de diversas inferencias válidas y que existen inferencias las cuales se encuentran presentes por su uso previo. Por lo tanto, dado que las reglas existen en este entorno, al encontrar que una regla produce una inferencia indeseada, se deberá modificar la regla de tal modo que esta no se produzca pero que, al mismo tiempo, siga dando cuenta de las inferencias válidas a las cuales se podía llegar mediante la regla antes de ser enmendada. Este proceso mediante el cual reglas e inferencias se modifican mutuamente es conocido como “equilibrio refractivo”.

Cabe aclarar que el método de justificación mediante el equilibrio refractivo no se encuentra libre de críticas. Pero para entender la forma en que estas pueden o no adecuarse a la presente aplicación del método, primero es necesario entender cómo será aplicado. Por lo tanto, aplazaré por ahora la explicación de estas críticas (la cual será tratada en la sección 2.4) y comenzaré por la explicación de cómo será aplicado el método. Sin embargo la utilidad de esta forma de justificación se observa en tanto no sólo ha sido aplicado en el campo de la justificación de las leyes matemáticas; su uso más conocido (Daniels, 2011) se encuentra en el campo de la teoría ética y política. Los planteamientos de John Rawls en *A Theroy of Justice* usan el equilibrio refractivo planteado por Goodman como parte fundamental de su justificación.

En este caso nos enfocaremos en buscar la forma en que se puede, mediante este equilibrio, trabajar el concepto “Pensar”. Al aplicarlo al esclarecimiento conceptual, el proceso se encontraría determinado mediante una definición (que sería el equivalente a las reglas en Goodman) y mediante la extensión del concepto (que serían los casos en que el concepto es aplicado -las instancias en que se presenta el concepto, equivalente a la práctica en el proceso

inicial-). Para que el concepto de pensamiento pueda ser definido de esta manera existen dos pasos fundamentales que se deben tener en cuenta antes de iniciar el proceso: Definir el concepto base desde el cual partir, garantizando que se encuentre delimitado pero que al mismo tiempo no legisle definitivamente qué clase de objetos pueden ser incluidos[17]; y, en segundo lugar, garantizar que, a medida que el equilibrio entre el concepto y los objetos es afinado, sea posible enmendar las reglas dadas sin que por eso el espíritu[18] de ésta se modifique con cada cambio. La primera de estas características se justifica en el problema planteado con anterioridad; si el concepto es demasiado estrecho, la posibilidad de aplicarlo a otras mentes desaparece. Podría parecer que de esta manera se está solicitando de manera inmediata que las otras mentes deban ser incluidas dentro de aquello que se considera pensar, pero esto no es cierto. Se solicita que no sea *esencial* al concepto que este sea aplicado *exclusivamente* a las mentes humanas. Si bien en un principio la posibilidad de la presencia de otras mentes es necesaria, a medida que se produce el equilibrio es posible delimitarlo de tal manera que todas las demás mentes queden excluidas. Esta exclusión no puede ser una premisa, debe ser un resultado que se da mediante la investigación. Además, existirán objetos que inicialmente sabremos que no caben en la definición de "sujetos pensantes", por ejemplo una roca, los cuales deben ser dejados de lado por este criterio inicial; esta clase de criterios provienen de los usos cotidianos del lenguaje y no estaríamos dispuestos, en ningún momento del equilibrio, a dejarlos a un lado. La segunda solicitud garantiza que el proceso sea fructífero, que en realidad se esté trabajando de manera juiciosa sobre un concepto y que no se están construyendo de manera caprichosa definiciones a medida que se presentan los objetos de los cuales se está predicando sin tener que reiniciar el proceso ante cada error o conclusión deseada.

[17] Con esto no quiero decir que el concepto de pensamiento planteado lleve a que *todo* objeto sea un objeto pensante. Este requisito tiene la función principal de impedir que la realización material no sea aquello que delimite sobre qué puede predicarse "piensa". Así, a pesar de que el concepto que se establezca inicialmente de "pensamiento" se aplique a un grupo reducido de objetos, el criterio de selección no será arbitrario.

[18] Palabra usada en el mismo sentido que tiene en la disyuntiva entre "la ley escrita" y "el espíritu de la ley", mientras el primero se refiere a la formulación específica, el segundo lo hace a las intenciones con las cuales la norma ha sido planteada. En este caso, las reglas de selección se modificarían en escritura conservando el espíritu.

2.2. Construyendo las bases del equilibrio.

La perspectiva humana aún deberá seguir existiendo en el concepto que se construya de "pensar", ya que finalmente se pretende dar cuenta de un evento observable por parte de los humanos; el error recae en imponer la forma de pensamiento humano como único criterio posible para la interpretación del pensamiento en los diferentes objetos. Creo, sin embargo, que es posible crear una definición que acuda a la experiencia humana sin que por esto se limite a ella.

En general, eventos sobre los cuales la principal fuente de información disponible es la de la primera persona deberán ser tratados con cuidado. Al considerar un evento desde la primera persona "se crea la ilusión de un estado de las cosas ontológicamente especial. Después de todo, en nuestras cabezas estamos con nosotros mismos, y las acciones funcionales de nuestros propios estados psicológicos"[19] (Lycan, 1996, pág. 67) por lo que, si el deseo es establecer un concepto que no se encuentre limitado por las imposiciones de la perspectiva humana, se deberá tomar un paso atrás al momento de establecer los conceptos ya que no contamos con una garantía de que se presenten exclusivamente en la forma en que los humanos lo experimentamos. El que el concepto no se encuentre limitado por la perspectiva de la primera persona, buscando expandir su aplicación a otras manifestaciones, nos solicita considerar que la información que adquirimos de nuestra experiencia es una entre muchas posibilidades y que no debe imponerse; sin por esto negarlo como una fuente de información ya que sigue siendo nuestra perspectiva de investigación. Si nos limitamos a establecer los conceptos mediante la información adquirida a través de la experiencia subjetiva de la primera persona, nuestra experiencia como investigadores, tendremos una descripción del evento típico en un humano. Sin embargo, a pesar de que debamos tener este cuidado en mente al momento de establecer lo que consideraremos por "pensar", no podemos olvidarnos que nuestra propia experiencia es aquella con la que contamos y, por lo tanto, nuestra experiencia es la principal guía para conducir una investigación en el tema.

Para que se puedan cumplir estas características al momento de formular la noción de "pensar", es necesario encontrar una manera en que esta pueda ser aplicada a cualquier sujeto sin importar su realización material. En este caso se está entendiendo "sujeto" como cualquier sustantivo posible, no se usa el término asignando ninguna propiedad específica. Además, el que

[19] "And the fact [that "No one else human, bat, or bat-human could know the same facts by being in the same functional state"] quite naturally creates the illusion of an ontologically special kind of state of affairs. After all, we are in our heads with ourselves, and our own psychological states' functional doings are of paramount importance."

la noción de pensar no se encuentre limitada por las condiciones materiales del objeto no quiere decir que todo objeto, sin importar sus condiciones materiales, piense; la primera afirmación es aceptar que pueden existir múltiples facetas físicas en las cuales se puede manifestar el pensamiento, lo segundo es simplemente un absurdo. Es absurdo porque entonces sería un concepto el cual no aportaría ninguna clase de información sobre los objetos, si la mera existencia de algo lo califica como pensante, entonces pensar se transformaría en un sinónimo de existir; es claro que de esta manera no se usa el término en el lenguaje cotidiano, al decir que algo piensa se está diferenciando de un (extenso) conjunto de objetos que no lo hacen. Pero a primera vista parecen existir demasiadas posibilidades de entidades físicas, justo como se mencionó en el ejemplo de los aliens invasores, de las cuales no sería posible dar cuenta de manera eficiente en una definición (menos aún si se pretende dar una definición extensional). Una manera en que es posible solucionar este inconveniente es mediante la formulación realizada por Vedral Vlatko en *Decoding Reality: The Universe as Quantum Information*:

> Este libro argumentará que la información (y no la materia, energía o el amor) es el fundamento en el cual todo se construye. La información es mucho más fundamental que la materia o la energía debido a que se puede aplicar exitosamente tanto a las interacciones macroscópicas, tales como la economía y los fenómenos sociales, y, como argumentaré, la información también puede ser usada para explicar el origen y comportamiento de interacciones microscópicas tales como la energía y la materia. (Vlatko, 2010, p. 10)

La información pasará a ser el bloque fundamental bajo el cual todo puede ser explicado. A pesar de que el objetivo del libro es explicar cómo la información, con toda su carga conceptual, es la forma en que se puede predicar sobre todo lo existente no será necesario, para los propósitos de la investigación, aceptar este presupuesto. El término será usado debido a que los alcances que pretende tener la teoría, como se puede observar en la cita anterior, son similares a aquellos requeridos por un concepto que dé cuenta de todo sin limitarse de acuerdo con las realizaciones materiales. A continuación se buscará *construir* una forma de comprender la información la cual la despeje de sus presupuestos teóricos y, al mismo tiempo, siga cumpliendo su función como término descriptivo básico de todo lo existente; cuando algo es "información" en el modo en que se usará en este texto no quiere decir que "informe algo a un observador", solo quiere decir que existe. Si el lector lo prefiere, esta palabra la podría cambiar por "Sivak", o "Existente" o cualquier palabra con una carga teórica mucho menor y la que escogiese tendría la misma relación con la *teoría de la información*.

Comprender de esta manera el universo facilita la definición inicial al proporcionarnos una manera de referirnos a todo lo que se encuentra en el universo sin por ello asignarle ningún tipo

de propiedad[20] y establece un primer criterio mediante el cual trabajar. Aceptando esta teoría, el computador donde estoy escribiendo, el desayuno que consumí esta mañana, el planeta en el que habitamos, los neutrinos que atraviesan el espacio, un rayo laser, el hambre, el amor y en general todo, será información. En este entorno se sabe que "Pensar" (y en realidad, cualquier otro proceso que se pueda llevar a cabo) será un predicado que se aplicará sobre la información sin, por esto, excluir ningún objeto en principio. Si el concepto al que lleguemos de "pensar" empieza su enunciación de la forma "Es todo aquello que" o "Es cualquier cosa que" la estructura de formulación parecería de inmediato predicar sobre cuales objetos se puede aplicar el concepto. Una forma en que se limitaría es que estas formulaciones parecen indicar de antemano que a lo que nos referimos es a un objeto físico, dejando de lado la posibilidad de aplicarlo, por ejemplo, a las relaciones sociales. En cambio si se llega a un punto descriptivo básico, en el cual simplemente tener la posibilidad de hablar o identificar algo es criterio suficiente para su inclusión, queda claro que el concepto al que lleguemos finalmente podrá aplicarse de formas como "El estado mental *x* piensa" o "los genes piensan". Esto es deseable debido a las múltiples manifestaciones que pueden existir de una experiencia si se quiere introducir una perspectiva no humana, en el caso del concepto de pensar podríamos hablar de una inteligencia evolutiva o social. Si al llevar a cabo la investigación encontramos que los procesos genéticos cuentan con las características establecidas para pensar, no existirá una razón para que se dé la exclusión simplemente porque la manifestación del pensamiento se da en un medio diferente al de las especies. También se podría presentar a nivel macro, considerando como un conjunto social piensa (por ejemplo un panal de abejas) sin que por ello se diga que cada uno de los miembros de la sociedad lo hacen. En este caso el uso de la palabra pensamiento para describir estos procesos funcionará no solamente para establecer una analogía (como cuando se habla del "gen egoísta" o la "inteligencia social") sino que será un componente a tener en cuenta al momento de construir el concepto de pensamiento. Si posteriormente se considera que algunas de estas formas de expresión van en contra de aquello sobre lo cual se podrá decir que "piensa", será labor de las reglas y los criterios que se establezcan su exclusión.

[20] Esto no ocurriría si hablase de algo como materia o energía, lo cual ya tiene una carga de contenido en el mundo de la física. Personalmente me vería inclinado a sugerir "Quark" y Leptón" como las palabras adecuadas para ser usadas en este punto, pero esto implicaría aceptar como verdadera en el campo de la filosofía la física de partículas.

Sin embargo, dos dificultades se hacen evidentes al iniciar este proceso: parece que, con el propósito de la inclusión, se ha perdido la barrera que diferencia un unicornio de la luna y, además, aún parecen colarse presupuestos relevantes de la teoría de la información al momento de realizar la aplicación.

La primera dificultad se presenta porque, si bien se plantea la reducción a información, aún no se ha presentado la forma en que, a partir de ella, se podrán describir los diferentes sujetos gramaticales. Nuevamente tomaré dos conceptos que provienen de la teoría de la información para continuar con la construcción y marcar divisiones: Bits y piezas[21]. El concepto de bit representará la unidad mínima de información, por su parte las piezas serán grupos de estas unidades las cuales componen un complejo definido. Estas divisiones estarán marcadas por el uso del lenguaje natural, la forma en que la perspectiva humana entra a jugar dentro de la construcción: por una parte se tendrá el lenguaje de la información que nos permitirá reconocer a todos los objetos en un mismo nivel mientras que, por el otro, se tendrá el lenguaje natural que aportará los objetos ya reconocidos y diferenciados. Más adelante (infra, sección 2.3, pág 32.) se mostrará la forma en que se produce la traducción desde el lenguaje natural al de la información. Sin embargo, inicialmente, podemos decir que los bits no tendrán un contenido determinado, simplemente será todo aquello que pueda considerarse como un componente mínimo y que las piezas serán todo aquello que podamos nombrar y se pueda descomponer; las piezas las identificaremos y utilizaremos mediante sus nombres en el lenguaje natural. Así, nuestro lenguaje natural nos permite diferenciar entre los objetos y, al mismo tiempo, poder trabajar con ellos de tal modo que tengan las mismas posibilidades de ser tenidos en cuenta; estos nombres representan una estructura de bits subyacente en la cual se justifica la diferencia. En algunos momentos del proceso de traducción podrá parecer que la diferencia entre dos objetos se pierde, al describir dos objetos diferentes en el lenguaje de la información estos podrán ser descritos del mismo modo; esto se presenta ya que el lenguaje descriptivo básico que se plantea busca, precisamente, cumplir esta función: no eliminar la posibilidad de evaluar un objeto como "pensante" simplemente por la carga que posee previamente en el lenguaje natural. Así el proceso de traducción permite eliminar y mantener la diferencia entre los objetos de una manera deseable: por una parte, son descritos de tal manera que todos tienen la misma posibilidad de ser tenidos en cuenta al momento de establecerlos como pensantes, por la otra, conservamos formas

[21] Se recurre al concepto de pieza en lugar del más familiar concepto "conjunto" debido a que el último puede indicar toda la "teoría de conjuntos".

de identificarlos para presentar ejemplos específicos que cumplen con la definición. La traducción permite identificar aquello que hacen para ser considerados como pensantes, la forma en que operan: el lenguaje natural permite diferenciar quiénes lo hacen mediante la conservación de los nombres. Si bien en el lenguaje de la información no hay algo que diferencie a "la luna del unicornio", los nombres que diferencian a estos dos objetos se conservarán durante todo el proceso de equilibrio; de esta manera a pesar de que en la traducción no haya forma de identificarlos, siempre se podrá observar como los objetos son incluidos o excluidos de acurdo a su nombre común para reconocer si algo está ocurriendo que no estemos dispuestos a aceptar.

Parece que la segunda objeción se hace cada vez más fuerte: se han introducido aún más conceptos los cuales parecen estar altamente cargados. Sobre todo, hablar de Bits e información parecería implicar que el universo está compuesto de datos binarios. Esto podría ser un inconveniente en la medida en que, si las partículas mínimas son binarias, siempre nos encontraríamos con piezas binarias de información y, como tales, podrían ser procesadas por una máquina universal de Turing. Si esto ocurre, entonces todo proceso, sin importar su clase, se podría representar en una máquina y la única diferencia entre uno y otro sería, necesariamente, el grado de complejidad del algoritmo; cuando todo componente del universo puede ser procesado por una máquina universal de Turing, esto es tanto las partículas como las posibles relaciones entre ellas, entonces todo resultado de las interacciones en él se podrán seguir procesando. Además, dado que uno de los motivos para la construcción del criterio es poder dar una respuesta al interrogante "¿Pueden las máquinas pensar?", se estaría cayendo en un círculo vicioso donde "las máquinas pueden pensar porque todo es información binaria y todo es información binaria porque necesitamos construir un criterio de pensamiento". Sin embargo, al solicitar simplemente la existencia de una unidad mínima de información desde la cual se compone, no es necesario que su representación se dé en un sistema de dos valores o, inclusive, de estados discretos. Cualquier concepción de la información que fuese distinta de, e irreductible al, binario permitiría configuraciones las cuales no necesariamente podrían reducirse a una máquina de Turing. Incluir un tercer (o cuarto, quinto, etc...) valor adicional sobre la información que no pueda reducirse ni expresarse en un sistema binario hará que, por definición, los algoritmos elaborados en este sistema de información no puedan ser representados en una máquina universal de Turing. Además, la expansión de la cantidad posible de valores no implica de manera alguna que este bit no sea la unidad mínima de expresión de la información, si la unidad básica de descripción tiene tres estados, ninguno de los cuales se puede representar como alguno de los otros dos, entonces

estos serán las formas mínimas de expresión. Si bien esto no garantiza nada sobre *qué* sea la información que se está tratando, esta incertidumbre permite evitar caer en el círculo. Si a pesar de no comprometerse con una estructura básica en la cual se expresa la información se puede construir un sistema para evaluar el concepto pensamiento, entonces las respuestas no estarán sesgadas por la descripción inicial; los términos que se usan acá son simplemente operativos y no por ello incorporan aquello que significan normalmente en la teoría de la información, no se está solicitando un universo reducible a términos binarios o de *información (con toda su carga conceptual)* se está dando un lenguaje en el cual describir las cosas y con suficiente maleabilidad para que se pueda aplicar a diferentes concepciones de lo que son estas cosas.

Si bien ahora tenemos el concepto de piezas, aún no parece ser un concepto útil: hasta el momento solo indica que ellas son "complejos de bits de información", pero parecería que no estamos con esto en ninguna medida mucho más cerca del concepto de "pensar". Sin embargo de esta manera ya se tiene un modo de decir sobre "qué se piensa" (información) sin estar limitados por cuestiones materiales. En esta perspectiva no se busca predicar sobre el mundo una característica del pensar, no se está diciendo si este es un proceso biológico, químico, eléctrico o espiritual; sólo se solicita el aceptar que es un proceso el cual existe[22]. El siguiente paso será construir el concepto de "pensar" mediante estas ideas básicas porque, si así lo hacemos, tendremos la certeza de que otras mentes, sin importar lo que sean, estarán incluidas.

2.3. Proceder a realizar una evaluación.

En el momento de empezar a ejecutar la construcción del concepto es cuando se debe recordar una condición del proceso: somos humanos quienes lo realizamos. De esta manera, los objetos sobre los cuales se trabaje deben ser aquellos que tienen la posibilidad de ser identificados por un humano; también la definición inicial debe corresponder con aquellos casos mínimos en los cuales reconocemos que se ha dado el evento "pensar". Tomando como base la construcción previa, el siguiente paso será extraer de algunos ejemplos el concepto base desde el cual puede partir el equilibrio. Para intentar que este concepto no incluya demasiados presupuestos (de aquello que es "pensar" en la forma netamente humana) se buscarán extraer las

[22] Se solicita que existe porque la investigación comienza para definir un concepto el cual usamos y poseemos en el lenguaje; es un presupuesto tácito que la conclusión de la investigación no será llegar a un concepto vació ya que esto sería igual de ridículo que llegar a un concepto que aplicase a todos los sujetos.

características de algunas situaciones ejemplares, las cuales podríamos aceptar como ejemplos de pensamiento sin importar si quien las ejecuta es, necesariamente, un humano.

Situaciones sencillas como "Frank abre la caja con una palanca", "Rodrigo habla con Frank" y "él hizo una suma" son ejemplos de eventos en los cuales se aceptaría, con facilidad, que se ha pensado. Por otra parte eventos como "El árbol hizo fotosíntesis", "Frank digirió una pizza", "la mesa está en la sala" o "El volcán lanzó una roca con sus gases" son algunos de los casos en los cuales consideramos que no se presenta pensamiento. Si bien existe una extensa variedad de ejemplos, ya de estos es posible rastrear características que permitirán construir inicialmente un criterio específico de inclusión y exclusión y, además, al expresar estas acciones de acuerdo con el lenguaje que se ha planteado en términos de información, se observa de manera directa que estas acciones cubren un espectro amplio de eventos. Para poder llevar a cabo las traducciones, las cuales se presentan como una herramienta para la construcción del concepto mas no como una descripción o reducción de las propiedades de los objetos, se podrán aplicar los siguientes lineamientos:

a. Todo sustantivo podrá traducirse directamente al lenguaje que hemos establecido.
b. Si sabe que el referente del sustantivo puede descomponerse de alguna forma, entonces este será una pieza de información.
c. Si se considera que el referente del sustantivo *puede* ser una unidad mínima, entonces será un bit.
d. Si no es posible determinar sobre un sustantivo (b) o (c), entonces será información. Sólo se deberá recurrir a esta regla para los sujetos pasivos de la oración.
e. Todo verbo deberá traducirse, de acuerdo a aquello que exprese, en "procesar", "estar" o "ser"[23].
f. Cuando el verbo implica el uso de, o por parte de, un objeto, como por ejemplo "pintar", este último debe hacerse explícito y aplicar (a).[24]

[23] Dado que el verbo predica sobre el sujeto, este o bien indicará una propiedad o una acción del sujeto. Si indica una propiedad, entonces se puede expresar como "ser" (y sus conjugaciones), de señalar una posición "estar", y en caso de ser una acción, esta será "procesar". Se selecciona "procesar" debido a que al ser una acción quiere decir que la información *x* interactúo y produjo algún cambio en un bit o un set *y*, el procesar es la forma de expresar que *y* ha sido alterado por *x*.

[24] Así, por ejemplo, al traducir "el martillo clava la puntilla" se sabe que "clava" en este caso implica el uso por parte de un sujeto del martillo por lo que la traducción no será "una pieza procesa otra pieza"

El lenguaje de la información que se ha construido permite dar cuenta de cualquier cosa que podamos nombrar, por esto es que cualquier sustantivo podrá ser traducido directamente al lenguaje. Los criterios (b) y (c) introducen por primera vez en el proceso de manera activa a la persona que lleva a cabo la evaluación y, de acuerdo a con su juicio, podrá determinar el tipo de información del que se habla; (d) hace que no tenga que forzarse una traducción en ausencia de evidencia, los sujetos pasivos podrán no determinarse en tanto ellos no son aquellos sobre los cuales se está evaluando la existencia de un pensamiento. Sin embargo no se deberá recurrir a (d) en el caso de que se esté traduciendo el sujeto activo de una oración, en estos casos es necesario reconocer si se está hablando de un bit o una pieza para así reconocer la clase de objeto que está produciendo el pensamiento. Reconocerlo es importante porque nos puede dar información acerca de la teoría que estemos construyendo, por ejemplo determinar si en ella los componentes mínimos e indivisibles pueden pensar, por esto solo se podrá evaluar para un sujeto pasivo una vez se ha determinado si este es una pieza o un bit. Dado que el verbo predica sobre el sujeto, el criterio (e) se construye para explicar cómo este indica una propiedad, posición (espacio-temporal) o una acción del sujeto. Si indica una propiedad, entonces se puede expresar como “ser” (y sus conjugaciones), de señalar una posición “estar”, y en caso de ser una acción, esta será “procesar”. Se selecciona “procesar” debido a que al ser una acción quiere decir que la información x interactúo y produjo algún cambio en un bit o un set y, el procesar es la forma de expresar que y ha sido alterado por x. “Procesar” es acá una forma de decir que ha ocurrido un cambio y que éste se puede atribuir a alguno de los sustantivos. Finalmente (f) es importante debido a que en una conversación existen verbos los cuales incluyen información la cual no está explícita pero de la cual se debe dar cuenta al momento de traducir. Siguiendo estos lineamientos una persona podrá traducir al lenguaje de la información diversidad de frases, en todas ellas su criterio como evaluador humano y usuario competente del lenguaje común le permiten llevarlas a cabo. Bajo estos lineamientos, a pesar de que se pueda llegar a más de una traducción posible para cada frase, la formulación permitirá identificar de manera neutral lo que ocurre en cada situación descrita para, así, extraer de ellas información que permita establecer los criterios para determinar cuándo ocurre el pensamiento; a pesar de que existen diversas maneras en que un evento pueda ser traducido al lenguaje de la información, bastará con que solo una de ellas apruebe el criterio de pensamiento. Aplicando estos criterios tenemos que “Frank abre la caja con

---

sino “una pieza procesa otra pieza' mediante otra pieza'”” donde “pieza” es el usuario del martillo, “pieza'” es la puntilla y “pieza'”” es el martillo.

una palanca" se lee "una pieza procesa una pieza mediante otra pieza", "Rodrigo habla con Frank" puede leerse como "una pieza procesa bits en otra pieza"[25] [26] y "él hizo una suma" será "una pieza procesó información[27]". Por la otra parte "El árbol hizo fotosíntesis" puede leerse como "una pieza procesó información", "Frank digirió una pizza" puede ser "Una pieza procesó otra pieza de información", "la mesa está en la sala" será "hay una pieza de información[28]" y "El volcán lanzó una roca con sus gases" se leerá "Una pieza modifica una pieza mediante otra pieza".

En estas traducciones se pueden identificar dos problemas iniciales: el criterio para determinar si algo es información, una pieza o un bit parece demasiado blando y existe información implícita que no se resalta en la traducción. El criterio parece ser demasiado blando por la posibilidad de que no se determine el tipo de información del que estamos hablando, y debido a esta posibilidad parecería que los conceptos de pieza y de bit son innecesarios. A pesar de esto, la utilidad de determinar la clase de información que se está presentando, sea esta un bit o una pieza, permite identificar propiedades del contenido mental relevante. Si se permite la no determinación de la información al momento inicial de la traducción es para permitir que el proceso avance sin que se hayan construido por completo las categorías para entender el mundo en términos de información, sin embargo, a esta cláusula se recorre solo en los casos en que no se puede establecer en primer lugar el tipo de información del que se habla. La segunda dificultad aparece al considerar que hay verbos los cuales indican procesos que implican otros procesos, pero que sin embargo no se hacen explícitos. Al pensar en procedimientos como la fotosíntesis o la digestión se considerará que los sistemas que llevan a cabo estos procesos no se encuentran mencionados de manera explícita y que, por (f), deberían aparecer en la traducción. Sin embargo esto no ocurre debido a que cuándo se dice "Frank digirió una pizza" parte del sujeto "Frank" que se está traduciendo es su sistema digestivo; esta será una información que está claramente disponible para quien lleva a cabo la traducción. Hay que recordar que no se intenta reducir ni

[25] En este caso se utiliza la palabra bits debido a que, a pesar de que el lenguaje puede ser un acto complejo, este puede comprenderse como la emisión y recepción de una idea y la idea *podría* considerarse como una unidad mínima.

[26] Si bien podría entenderse que el acto de emitir información hacia otra persona podría considerarse como "un set emite bits que son leídos por otro set", de esta manera no se estaría hablando, la participación del segundo sujeto sería simplemente adicional y no estaría considerada en la forma en que se interpreta la frase.

[27] No es necesario determinar si los números son bits o piezas de información.

[28] A pesar de que en la oración se indican dos objetos, en tanto ninguno de los dos está llevando a cabo acción alguna o existe otro sistema con el cual relacionarlos, podemos considerarlos como pertenecientes a una misma pieza de información.

explicar en la traducción todos los procesos relevantes al momento de llevar a cabo un proceso, esta traducción sirve para identificar *quiénes*, en tanto sujetos identificados por un investigador, están haciendo *qué*, entendido como la clase de operación sobre la información. Es necesario recordar que estas traducciones surgen como una asistencia al momento de construir un concepto de pensamiento en el que no se discrimine a los objetos por sus características materiales, pero esto no quiere decir que la traducción represente una reducción y que ella se transforme en la única información relevante desde la cual trabajar.

Los ejemplos se han planteado intencionalmente de tal manera en que se tuvieran casos en los cuales la lectura en el lenguaje de la información fuera idéntica y, aún así se siguiese diciendo, desde el análisis del sentido común, que en algunos casos se piensa mientras que en otros no. Esta no es una razón por la cual se deba abandonar la traducción, dado que a pesar de que en un primer momento se presenta este inconveniente, es posible seguir extrayendo de ella datos sobre aquello que debemos entender por pensamiento. Lo primero que se puede observar es cómo la pasividad de la información no puede ser considerada como un pensamiento. El que un bit o una pieza de información permanezcan estables indica que no están llevando a cabo un pensamiento ya que pensar es hacer un proceso; la ausencia de un procesamiento de información niega la existencia de un pensamiento. Hay que aclarar que en este nivel descriptivo no se está negando la existencia de procesos de pensamiento los cuales no tengan una manifestación física si se recuerda que los monólogos internos, relaciones sociales, sentimientos o demás son considerados también acá como formas de información. Cada vez que se procede activamente sobre ellos o sobre objetos físicos se está negando la inactividad, cada vez que estos eventos mentales son llevados a cabo existe una modificación en la pieza a la que corresponden.

Además los pensamientos se asignan a una pieza de información particular o, desde la perspectiva humana, a un sujeto dentro de una oración. Al evaluar “Frank abre la caja con una palanca” se considera que Frank es quien pensó para que la acción se produjese a pesar de ser la palanca la cual actuó de manera directa sobre la caja. Al observar una situación cualquiera las responsabilidades se pueden asignar con facilidad, en el ejemplo inicial será evidente, una vez observemos las respuestas que el robot da, que hay alguien comunicándose con nosotros, ya sea “una especie alienígena”, “un robot inteligente” o inclusive “una persona haciéndose pasar por otra cosa”, debido a la diversidad de preguntas que él puede responder (y la congruencia existente entre ellas). Sin embargo asignar la tercera de las posibilidades es incorporar información que no está disponible de manera explícita o implícita en las formulaciones sobre aquello que está

ocurriendo y esto sería un error. Para que podamos decir que algo piensa tendremos que poder asignarle la capacidad de llevar a cabo alguna manipulación o acción sobre la información, en general, debe ser capaz de procesarla. Con esto podemos tenemos una primera característica para decir que se presenta pensamiento:

(i) Debe poder adjudicársele a una pieza el procesamiento de información[29]

Si bien esta es una característica que nos indica sobre qué estaremos hablando al momento de asignar el juicio y con la cual se excluye de entrada eventos que no podrían ser considerados pensamientos, este filtro no es suficiente. Como se observa en los ejemplos mencionados, existen diferentes casos en los cuales se presenta procesamiento de información pero aún así nadie diría desde el sentido común que esto es pensar. La digestión o fotosíntesis son casos en los cuales claramente existe procesamiento de información pero aún así no se piensa, sin embargo esto no es motivo para de inmediato abandonar una construcción en términos de información. En este punto la perspectiva del observador humano se hace nuevamente relevante, pero no para imponer en el concepto lo que él hace al pensar, sino para encontrar aquello que encuentra erróneo al momento de declarar estas acciones como pensamientos. Si se hace lo primero, el observador simplemente resalta algo que él puede hacer mientras que el objeto no pensante es incapaz, si se hace lo segundo, se evalúa lo que ocurre en el objeto para encontrar algo que allí ocurra que no se presente en los demás casos que aún se encuentran incluidos (y que se presenta en aquellos que son de la misma clase que él, es decir, que objetaríamos por las mismas razones).

Una primera observación que se puede hacer es que aquellas situaciones que son estrictamente necesarias para la supervivencia del objeto no son tenidas en cuenta como eventos donde se presente pensamiento. Situaciones complejas, como por ejemplo un animal regresando a su lugar de nacimiento para poder llevar a cabo el acto reproductivo, no son consideradas como un acto de pensamiento; estas son usualmente desechadas como meros “instintos naturales”. Lo mismo pasa con la digestión o la fotosíntesis, estas situaciones por complejas que sean y, por más que puedan ser descritas en términos de procesamiento de información, nadie dirá que al llevarse a cabo se está pensando.

[29] Aquí es de suma importancia recordar la regla (f) y la nota al pie 11.

Esta supervivencia se muestra útil al momento de hablar de organismos vivientes, pero si cambiamos la escala y nos referimos a los sistemas específicos encargados de la función (por ejemplo el sistema digestivo) no es claro en qué manera se pueda referir a la supervivencia. Además, introducir este concepto tampoco parece ayudar en lo más mínimo a salvar casos como el de los volcanes haciendo erupción. Pero si extendemos la idea de supervivencia a aquello que denominaré "supervivencia *conceptual*" considero se puede crear un criterio suficiente de exclusión. Un evento es de supervivencia conceptual cuando, de este no presentarse, aquello de lo que estamos hablando dejaría de ser para nosotros lo mismo. Tomemos el caso del sistema digestivo, para este el acto de digerir es un acto de supervivencia conceptual ya que en su ausencia no podríamos seguir considerando que este sea un "sistema digestivo". En el caso del volcán que lanza rocas con su gases, una característica por la cual identificamos que un volcán es un volcán es el hecho de hacer erupción, y es consecuencia de hacer erupción que haga volar rocas con sus gases; por lo tanto si el volcán dejase de enviar rocas con su lava este dejaría de hacer erupción y entonces dejaría de ser un volcán para nosotros. En este caso la perspectiva de observadores humanos es altamente importante, dado que depende de aquello que entendamos por la cosa de la que estamos hablando lo que para ella será un acto de pensamiento.

Si bien de esta manera podemos dejar de lado una gran cantidad de eventos que desearíamos excluir, es necesario evaluar si este procedimiento elimina más de lo debido. Podría pensarse que para la supervivencia conceptual del humano una característica definitoria es que esté vivo, y que como consecuencia directa de esto se sigue que debe alimentarse y por lo tanto debe cazar o consumir vegetales. Parecería que de esta manera un patrón de caza organizado o desarrollar la agricultura no podrían considerarse como ejemplos de pensamiento ya que son una consecuencia de mantenerse con vida. Sin embargo no es consecuencia de cazar el cazar en grupos de manera organizada ni es consecuencia de consumir vegetales el desarrollar una agricultura, estas acciones se pueden llevar a cabo de formas como la recolección y la caza individual (estas formas mínimas sí pueden considerarse como casos pertenecientes a la supervivencia conceptual) y, por lo tanto, su negación no afecta directamente la supervivencia del humano. De la manera en que la vida se incluye en este ejemplo, también podrían incluirse todas las características que implican la supervivencia biológica y fisiológica de un organismo (por ejemplo la capacidad de replicación de un virus).

Por otra parte el identificar la supervivencia conceptual trae consigo la ventaja de permitir actos que parecerían iguales inclusive en el lenguaje cotidiano. "Frank hace una suma" será

diferente de “la calculadora hace una suma” ya que, aunque ambas oraciones tienen la misma estructura básica y traducción al lenguaje de información, si una calculadora no puede realizar una suma quiere decir que esta no lleva a cabo los cálculos y, por lo tanto, deja de ser calculadora. Por la otra parte, si Frank carece de educación y por lo tanto no es capaz de llevar a cabo una suma, de esto no se sigue que se viole una de sus características constitutivas. De esta manera el criterio ayuda a asignar el pensamiento a aquello que en realidad deseábamos asignárselo, Frank pensará pero la calculadora no.

Así la segunda característica inicial que planteo para la definición de pensamiento es:

(ii) Es un procesamiento de información que no pertenece a la supervivencia conceptual de la pieza.

La inclusión del concepto de supervivencia conceptual tiene otra ventaja que es necesario resaltar: permite hablar de conjuntos de piezas, no solamente de ejemplos particulares. Los conceptos, sin entrar en las minucias de la discusión filosófica al respecto, son generalmente considerados como abstracciones las cuales no hacen referencia a un objeto en particular. Así, al llevar a cabo una investigación en la cual se considere la supervivencia conceptual de un objeto, la investigación no se remite a la pieza en particular sino a todas aquellas que comparten las características para encontrarse abarcadas por el concepto dado. Las investigaciones no serán acerca de Juan, Alfred, Laki o Mrs. Chippy, estas tratarán sobre humanos, máquinas, volcanes o gatos.

Lo que hasta el momento he hecho es mostrar un punto desde dónde comenzar iniciar el proceso de equilibrio, mostrando las consideraciones que se deben tener en cuenta al momento de introducir un nuevo criterio. Puede que los criterios enunciados hasta el momento asignen como pensantes objetos que no aceptamos como tales, pero para esto se establece el proceso de equilibrio. Una vez se cuenta con un criterio se observan sus consecuencias y se analiza cuáles de ellas estamos dispuestos a aceptar. Cuando se encuentra en los resultados cosas indeseadas, será necesario observar si es necesario adicionar criterios o afinar los ya establecidos. Con el criterio establecido sería el tiempo de evaluar que objetos han entrado, observar sus características comunes, rastrear consecuencias indeseadas y, en general, continuar con el proceso.

La definición de “pensar” que se ha presentado considero que es lo suficientemente sencilla y sólida como para garantizar que, ante diferentes retornos por parte de la experiencia, no se reemplazará sino que se afinará. Este proceso para delimitar de manera cada vez más certera el

pensamiento puede o no encontrar su final, pero la riqueza y la información valiosa no se encuentra en la respuesta final sino en el propio proceso. Seguir con el equilibrio refractivo conlleva a investigar a fondo los procesos que llevan a cabo cada uno de los sujetos que estamos considerando explorando la manera como ellos procesan la información. Las exploraciones sobre "cómo lo hacen" serán las que permitirán afinar el equilibrio y, al mismo tiempo, traerán consigo los criterios necesarios para la exclusión de los sujetos no pertinentes. Si el equilibrio encuentra su punto final, el concepto de "pensar" podrá aplicarse a un número cualquiera de sujetos, pero en este punto habremos encontrado una gran cantidad de información valiosa de investigación sobre los sujetos que se han excluido dado que el proceso solicita que si se presenta su exclusión, esta se dé por algo que ellos hacen.

La forma en que se ha planteado el lenguaje de la información permite que la formulación de los criterios sea lo suficientemente general como para que ellos no estén excluyendo en su formulación más de lo que hacen de manera explícita, de esta manera es poco probable que sea necesario reformular los criterios por exclusiones no previstas (exclusiones que se deben a las palabras utilizadas). Además, el criterio de la supervivencia conceptual ha permitido excluir situaciones las cuales, en el lenguaje básico usado para la descripción, podrían llegar a tener una misma descripción sin por esto imponer una característica propia del pensar humano. Quizá algunas situaciones indeseadas puedan pasar el filtro de la supervivencia conceptual, aún así él no excluye algo que se desee considerar como "pensamiento" y su utilidad al momento de diferenciar es lo suficientemente grande como para querer conservarlo. Sin embargo, teniendo en cuenta características ya mencionadas de los criterios planteados, creo que se ofrece una base sólida sobre la cual edificar un concepto de pensamiento el cuál no se limite a imponer la experiencia individual humana como la única información válida.

2.4. El lugar de las críticas al equilibrio refractivo.

En este momento es pertinente preguntarse qué se ha logrado hasta el momento con la aplicación del equilibrio refractivo y hasta dónde sus críticas pueden llegar a desmantelar esta aplicación. Sin embargo la mayoría de estas críticas[30] se enfocan en dos aspectos: el equilibrio refractivo no da cuenta de la manera en que se producen las inferencias y, en segundo lugar, no es

[30] Algunas fuentes de referencia sobre la crítica y el desarrollo del problema del equilibrio refractivo son (Kelly & McGrath, 2010), (Siegel, 1992)y (Stich, 1990). Todas estas son críticas que se aplican a la noción presentada por Goodman del equilibrio refractivo, las críticas que surgen a la aplicación de Rawls no han sido tenidas en cuenta en esta investigación.

un método adecuado de justificación. El primer tipo de críticas buscan mostrar cómo del equilibrio refractivo no se llega a las leyes de inferencia que en efecto aplicamos ni que éste es un reflejo de la forma en que las reglas se producen convirtiéndolo, por ende, en un método inadecuado para dar cuenta de las leyes de inferencia. Esta clase de críticas claramente no afectará la aplicación del método para el caso planteado, si el equilibrio es adecuado o no para dar cuenta de los procesos de inferencia no tiene nada que ver con el hecho de que este pueda o no ser un proceso efectivo para una situación diferente.

Ahora bien, el segundo grupo de críticas de aplicarse podría dejar sin justificación el uso del método para producir la definición de un concepto. Si el equilibrio refractivo no es suficiente para otorgar por sí mismo justificaciones entonces no podrá derivarse de él conocimiento debido a que no existiría una justificación de la definición planteada. Dado que estamos usando el método para definir un concepto, sería deseable que, como resultado de la investigación, se alcanzara una formulación justificada; si este es o no conocimiento dependerá de la verdad del enunciado al que se llegue, y esta es una cuestión que puede variar mediante el resultado de cualquier investigación futura, sin embargo la existencia de una justificación fortalece las probabilidades de que la definición planteada sea la adecuada. Mostrar que el equilibrio refractivo es una forma adecuada de justificación es un problema independiente al desarrollado en esta monografía, el cual será aún más complejo si se identifica que:

> Hemos notado como una acusación común entre los detractores del método del equilibrio refractivo es que el método es extremadamente *débil*: esto es, ellos dicen que, aún si se ejecuta el método de manera impecable, aún así se podría llegar a puntos de vista carecientes de diversas características [features] deseables. Un hecho notable que ha emergido de nuestro estudio de Goodman, Rawls, y Scanlon es hasta dónde parece que comparten esta posición. Esto es, es notable lo modestos que son acerca del estatus que se puede adjudicar a las resultados que arroja el método. Para Goodman, aún cuando el método se aplica impecablemente para llegar a creencias verdaderas sobre el futuro, estas creencias se quedarán inevitablemente cortas de ser conocimiento. Rawls era agnóstico (en su momento más optimista) y escéptico (en sus escritos tardíos) acerca de que diversos individuos quienes aplicaran el método pudieran converger en un único equilibrio refractivo amplio; algo que él consideraba como una condición necesaria para la existencia de "verdades morales objetivas" (y entonces, presumiblemente, para el conocimiento de dichas verdades). Scanlon reconoce que la "fuerza justificadora" de una aplicación del método depende de la credibilidad de sus puntos de partida. Así, Scanlon parece estar de acuerdo con que aún si una persona empieza desde todos sus juicios considerados y solamente estos, y desde ellos alcanza efectivamente el equilibrio refractivo amplio, sus apreciaciones pueden sin embargo no estar justificadas si los juicios considerados desde los cuáles parte carecen de la suficiente credibilidad racional. (Kelly & McGrath, 2010, pág. 38)

Observamos así cómo inclusive los principales proponentes del equilibrio refractivo como un método efectivo desde el cuál establecer una justificación son bastante parcos al momento de establecer la certeza del conocimiento al cual se puede llegar mediante él. El que el equilibrio refractivo pueda o no llegar a presentar conocimiento como resultado de su aplicación depende,

entre otras, del marco de creencias epistemológicas en el cual se aplique[31] o de la noción de justificación que se considere.

Pero de la misma forma en que esta incertidumbre sobre el conocimiento alcanzado por la aplicación del equilibrio no es aspecto que impidió la presentación de los sistemas manejados por Goodman, Rawls o Scanlon, considero que la presente propuesta cuenta con suficientes elementos para ser considerada como valiosa al momento de estudiar el concepto de "pensar". En primer lugar, esta es una perspectiva planteada como una alternativa de estudio en que se permita la entrada de las otras mentes al momento de la definición. La existencia de formas de pensamiento que no se encuentran representadas por la forma de pensar humano es un presupuesto desde el cual se parte para la formulación de las bases del equilibrio, pero esto no se encuentra ajeno a una justificación.

Por una parte, si no se considera que esta sea una perspectiva valiosa, los estudios previos sobre la mente y los criterios de evaluación para establecer que una máquina piense serán los adecuados; pero aún así desde las bases presentadas para edificar el presente equilibrio será posible llegar en una etapa posterior a una conclusión similar. Lo que esto quiere decir es que es posible considerar que las otras mentes son finalmente irrelevantes para establecer lo que significa pensar, lo que haría que los criterios planteados en la sección 1 fueran los adecuados para responder la pregunta acerca del pensamiento en las máquinas. Sin embargo esto no eliminaría la utilidad de conducir un estudio mediante el equilibrio refractivo, al proceder con él podríamos encontrar las razones y las características que convierten a las otras mentes en elementos irrelevantes. No solamente se establecería la experiencia humana como criterio de comparación y evaluación para el pensamiento, existiría una exploración de los procesos existentes en cada clase de mente y las razones por las que estos procesos no son suficientes para calificar como pensamientos.

Por la otra, al reconocer las otras mentes como un campo valioso para de investigación, la incapacidad de reconocer en ellas sus los procesos internos trae consigo la necesidad de establecer un criterio el cual pueda ser susceptible a modificaciones presentadas por la nueva información que podamos adquirir sobre ellas. Este último punto se ve fortalecido en el equilibrio,

[31] Serán muy diferentes las dificultades que un fundacionalista encuentre al aplicar el equilibrio refractivo que aquellas que un coherentista pueda ver, por ejemplo. El primero observará cómo no existen garantías sobre la relevancia de y la verdad de aquello que afecta el equilibrio, el segundo podría objetar que el equilibrio sólo requiere coherencia interna y no con toda la estructura del conocimiento.

de la manera en que se ha presentado acá, debido a la forma en que permite establecer la paridad entre diversos tipos de sistemas, permitiendo que tengan una oportunidad más equitativa de ser tenidos en cuenta dentro de las conclusiones a las cuales se llegue por el procedimiento. Así, las exploraciones podrán centrarse en diversos tipos de piezas pensantes y explorar en ellas características generales que permitirán la eliminación o validación de múltiples sujetos de investigación.

Finalmente, una diferencia clave entre investigar el concepto de "pensar" mediante el equilibrio refractivo con otras aplicaciones del método es la influencia posible de la investigación científica. Las reglas de la inferencia y las creencias morales son campos en los cuales los resultados de hecho presentados por una investigación científica se encuentran tradicionalmente distantes. Lo que se considera como una creencia pertinente, o un juicio moral considerado, al momento de ejecutar el método se encontrará vinculado principalmente a las creencias de quien lo lleva a cabo; esto es lo que ocurre al momento de trabajar en la moral o las inferencias. Sin embargo, al momento de definir mediante el equilibrio el concepto de pensar, toda información que se tenga sobre el funcionamiento de las otras mentes, esto es, toda evidencia que pueda resultar de una investigación científica, tendrá la oportunidad de ser parte del resultado final. Con esto se permite la entrada de elementos los cuales serán constantes a través de las diferentes instancias del equilibrio refractivo. La incorporación explícita de elementos pertenecientes a las ciencias, los cuales no se modifican de acuerdo a las perspectivas individuales establece una diferencia directa entre la justificación de usar el equilibrio para definir las reglas de inferencia y reglas morales y la justificación de hacerlo para definir el concepto "pensar". Al definir estas reglas siempre se acude a una experiencia individual, lo que trae consigo dificultades al momento de encontrar uniformidad en los resultados que puede arrojar el método; por otra parte el papel privilegiado de la información científica al momento de definir el concepto hará que la diversidad de posibles conclusiones sea menor. Esto, en sí mismo, representa una variación en el poder justificador del método en cada caso ya que, por ejemplo, eliminaría dudas acerca del acuerdo al que se puede llegar entre las diferentes aplicaciones del método.

Puede que con esto no se llegue a una conclusión sobre la justificación del equilibrio refractivo, sin embargo quedan claras dos cosas: si el proceso del equilibrio se encuentra justificado, entonces el ejercicio que se ha llevado a cabo lo está y, en caso contrario, las características particulares de este ejercicio invitarían a evaluar si la inclusión de información no-

variable[32] hace que sí lo esté. Puede que finalmente la respuesta sobre la justificación sea negativa, pero aún en este caso el ejercicio sería útil en la medida en que podría hacer explícito aquello que se encuentra en juego al momento de considerar algo como pensante, aún cuando no sea por este método que se llegue a la definición justificada.

2.5. Ventajas de esta definición.

Hasta el momento se ha hablado de cómo la construcción de este concepto de pensamiento posibilita la inclusión en la discusión de las mentes no-humanas. Sin embargo esta no es la única ventaja que puede traer consigo abordar esta perspectiva, ni tampoco es el campo en el que se piensa aplicar el único en el que puede ser útil.

La forma en que se ha buscado construir el concepto tiene como consecuencia que éste sea flexible a nivel ontológico. Con esto quiero decir que, debido a la concepción de información que se está manejando, es posible adaptar el criterio y los contenidos para que coincidan con diferentes concepciones ontológicas. Un dualista podría encontrar una forma de expresión en múltiples valores o clases de bits, un reduccionista podría manifestarlo todo en bits o un dualista de propiedades podría hablar de las cualidades emergentes en las piezas. Todo esto se puede construir a partir de la definición planteada de manera inicial. Los dos criterios se plantean desde niveles los cuales pueden ser aceptados por diferentes clases de teorías, (i) es flexible debido a la forma en que se ha utilizado el concepto de información y por las razones que previamente mencionadas, (ii) solamente solicita reconocer que somos humanos realizando el estudio y que, como tales, identificamos los objetos que estamos tratando. De hecho este rodeo que se ha presentado para definir el pensamiento busca precisamente que la perspectiva humana no se transforme en un factor desde el cual se definirá el concepto sin que por ello esta deje de arrojar información y guiar el proceso de la investigación. Las diferentes teorías que se pueden presentar sobre el funcionamiento de la mente humana, los trabajos en filosofía de la mente, son una fuente posterior de información para avanzar más en la definición, pero estos deberán ser considerados en conjunto con los avances en estudios de aquellas mentes no humanas que puedan ser encontradas.

[32] Los contenidos de investigación científica. A pesar de que los postulados científicos puedan variar, cuando lo hagan, será un cambio que deberá tenerse en cuenta en cualquier aplicación del equilibrio. De esta manera los postulados sobre los cuales exista acuerdo en la comunidad científica serán no-variables en las distintas instancias del equilibrio a pesar de que la formulación o el acuerdo pueda cambiar.

Es necesario aclarar que la metodología del equilibrio refractivo nos permite declarar la definición que se ha dado previamente de pensamiento como un punto de partida de la investigación. La definición actual pretende dar cuenta de unas condiciones que se considerarían necesarias para dar cuenta de un concepto de pensamiento en el cual exista posibilidad para la existencia de mentes no humanas. No por esto se está afirmando que estás sean *las* condiciones necesarias ni que tampoco estas sean *suficientes*, sólo se afirma que pueden funcionar como un punto de partida; al encontrarse la definición dada en un proceso de equilibrio estas dos características (necesidad y suficiencia) no pueden ser asignadas. Esto ocurre porque cada condición que se haya manifestado está sujeta a ser enmendada o modificada y, además, existe la posibilidad que al establecer el equilibrio nuevas condiciones se hagan presentes. Al momento de establecer el equilibrio el investigador siempre deberá considerar que toda especificación que lleve a cabo sobre el concepto siga cumpliendo un requisito de necesidad, sin embargo la suficiencia no es una aspiración que se deba perseguir.

3. Conclusiones.

En esta sección pretendo señalar las consecuencias que considero más importantes de llevar a cabo la investigación sobre el pensamiento de las mentes no-humanas de acuerdo con el procedimiento que he señalado con anterioridad. En particular se buscará estudiar cómo las máquinas deben ser tenidas en cuenta en el estudio, lo que significa para la pregunta general de la AI y algunos lineamientos de cómo podría continuar la investigación después de este trabajo.

3.1. ¿Y las máquinas?

Considero que mediante los lineamientos iniciales que he presentado no es posible afirmar inicialmente que las máquinas son objetos no pensantes. En particular evaluaré por qué esta afirmación es viable para los computadores personales modernos (PC), los cuales son considerados como un ejemplo característico de máquina (de Turing)[33].

Sólo uno de los dos requisitos que se han asignado hasta el momento para el concepto de pensamiento puede resultar problemático ya que para nadie existe la menor duda de que un PC procesa información. Por lo tanto el punto que podría resultar problemático es determinar si el PC lleva a cabo procesos los cuales no son parte de su supervivencia conceptual. Para construir un caso en el que los PC no puedan ser considerados como pensantes, al menos bajo estos requerimientos iniciales, sería necesario construir una manera de entender los PC mediante la cual cada uno de los procesos que llevan a cabo fuesen considerados como parte de su supervivencia conceptual.

Encontrar los eventos que, de no presentarse, harán que un PC deje de ser un PC para nosotros puede no ser una tarea muy compleja. En efecto, existirán elementos básicos como la existencia de un mecanismo mediante el cual pueda interactuar con los humanos (inputs y outputs), su capacidad de ser programado para múltiples tareas o el intercambio libre de componentes que compartan una arquitectura; todos estos son eventos que pueden encontrarse en común en todos los PC. Aún así la dificultad se presenta al momento de intentar crear la lista de

[33] Si bien esta es una afirmación tradicional, argumentos como el de (Wang, 2007) buscan mostrar como el entender el PC como una máquina de Turing puede ser un error. Aún si este argumento es correcto, los PC siguen siendo un ejemplo de una máquina para cualquier uso y, de determinarse que algunos de sus procesos son parte de lo que se llama inteligencia, serían un ejemplo de AI.

tal modo que dé cuenta de todas las acciones posibles que pueda llevar a cabo un PC. Para ejemplificar dónde puede encontrarse la dificultad supongamos el siguiente caso:

Inicialmente tengo dos computadores exactamente iguales, ambos son relativamente modernos y llevan a cabo todas las funciones que normalmente esperaría de un computador. Un día tomo la decisión de querer probar en alguno de ellos un juego, pero descubro que la tarjeta de video incluida en la configuración inicial no es lo suficientemente potente como para ejecutarlo y decido insertar una nueva tarjeta de video en mi computador. La actualización no requiere ningún otro cambio en el hardware ni en el software del equipo, lo único necesario es la instalación del driver correspondiente y, desde ese momento, puedo ejecutar a la perfección mi nuevo juego. Con esto tengo dos piezas que en su mayoría son iguales pero, aún así, cada una tiene capacidades diferentes.

¿Qué ocurre entonces con la supervivencia conceptual al momento de ejecutar este juego? Tres posibilidades se presentan: ejecutar el juego no es parte de la supervivencia conceptual, lo es, o debido al cambio de hardware estamos hablando de dos conceptos diferentes. Si tenemos la primera de las opciones, entonces el segundo computador está claramente llevando a cabo una acción que se presenta, en los términos que hemos dado hasta ahora, como de pensamiento. En el segundo caso, tendríamos entonces que los computadores que no han sido equipados con esta tarjeta de video no entran entre aquello que consideramos un PC; esta vía sería seriamente problemática, no creo que exista alguien dispuesto a considerar que el computador que no puede ejecutar este juego no es un PC. Finalmente tendríamos la opción de decir que cada uno de estos computadores son ejemplos de cosas diferentes, que cada uno debe ser tenido en cuenta de manera particular y, por ende, deben tener criterios de supervivencia conceptual independientes.

Esta última opción es problemática por dos razones principales: en primer lugar atenta contra el sentido común y, en segundo, imposibilitaría la evaluación de especies en general. Considero que atenta contra el sentido común en tanto cualquier persona que no supiera del cambio al usar los computadores estaría dispuesto a decir que son lo mismo, en general el concepto que manejamos de PC tiene espacio para una gran variedad de máquinas con especificaciones diferentes. El segundo problema se da debido a que, de aplicar este criterio, tendríamos complicaciones generales al definir como, por ejemplo, una persona con un desorden emocional, alguien con incapacidad de movimiento en sus miembros y un autista pueden ser considerados todos como ejemplos de humanos. En la escala que se está trabajando al decir “Los

homo-sapiens-sapiens piensan", "los PC piensan" o "las abejas piensan" se está haciendo una afirmación sobre los procesos que ocurren generalmente en la especie, no sobre cada uno de sus miembros y sus particularidades. Si bajo cada cambio de componentes de hardware dentro de un computador estuviésemos hablando de cosas distintas, esto sería equivalente a hablar de humanos distintos en cada configuración neuronal o fisiológica posible.

Existen formas más complejas de plantear la supervivencia conceptual para un computador las cuales no requieren de este nivel de especificidad. Por ejemplo, una característica de un PC es que "tiene la capacidad de ejecutar todo programa que es escrito para él, dado el software y hardware necesarios". Una formulación como esta evita varios problemas, por ejemplo, la imposibilidad de ejecutar programas para sistemas operativos o arquitecturas diferentes (no poder ejecutar un programa escrito para Mac en Windows, o ejecutar un juego que requiera una tarjeta de video con *Hardware Tessellation* en un computador que no la posea), así como la imposibilidad de ejecutar un programa debido a que entra en conflicto con otro previamente instalado. Un criterio como este da cuenta de todo programa que es posible instalar en un computador particular y, sin embargo, puede ser aplicado a cada PC en particular. Pero esto no garantiza que el criterio planteado sea uno de supervivencia conceptual ya que, a pesar de ser aplicable a todos los PC, no por esto es una condición necesaria para que los consideremos computadores. Es posible suponer un caso en el cual se tenga el software y el hardware adecuado para la ejecución de un programa y, aún así, PC no lo ejecute; si bien este comportamiento sería considerado netamente anómalo, no por esto diríamos que la máquina que tenemos frente a nosotros ha dejado de ser un PC. Si acabo de comprar un programa para mi computador, revisando con cuidado que mi PC cumpla las especificaciones allí listadas, y al instalarlo este no funciona (por ejemplo, existe un error de compatibilidad con algún otro programa instalado) mi primera reacción será buscar una forma en cómo hacer funcionar el programa, no evaluar si aquello que tengo en mi escritorio es realmente un PC. Además esta formulación del criterio de supervivencia conceptual se encuentra en contra de la forma en que la supervivencia conceptual es parte de la definición. Sobre un objeto no es necesario predicar de manera explícita y positiva aquello que permite su supervivencia, esta propiedad es evaluada para cada acción relevante observando la pregunta "si dicha acción no se presentase ¿seguiría siendo el mismo concepto para nosotros?". Por esto, mas que presentar las condiciones de supervivencia, el evaluar si existen procesos tempranos de "pensamiento" es considerar si el computador lleva a cabo acciones las cuales de no ser ejecutadas este seguiría siendo un PC.

Con estos argumentos no estoy diciendo que los PC piensen, tampoco estoy negando la posibilidad de construir un criterio mediante el cual pueda mostrarse que, en efecto, todas las acciones llevadas a cabo por un computador son las necesarias para que este sea considerado como tal. Hasta el momento estoy planteando que, estableciendo los criterios ya presentados como un punto de partida para establecer una noción de pensamiento en el que se vean incluidas las mentes no-humanas, los computadores deben ser uno de los elementos que merecen ser estudiados. El que en una primera etapa de la investigación mediante el equilibrio refractivo permita incluirlos no garantiza que, en una etapa posterior, sea necesario conservarlos como una evidencia relevante de pensamiento.

3.2. ¿Pueden o no pensar las máquinas?

El aplicar el método del equilibrio refractivo a la investigación sobre la AI no busca como tal dar una respuesta directa a la pregunta, lo que pretende es cambiar el enfoque de la pregunta adecuada acerca del pensamiento en las máquinas. Al continuar trabajando en el equilibrio es posible establecer si, en ese momento de la investigación, se considera que las máquinas piensan; a pesar de esto, esta no es la pregunta que guía la investigación. Para poder afinar el equilibrio más allá de los criterios iniciales es necesario investigar los procesos que se presentan en las diferentes piezas que se consideran como "pensantes", solamente esta investigación puede llevar a perfeccionar el conjunto de reglas.

Así, tendremos que preguntas del tipo "¿Cómo piensan las abejas?" o "¿Cómo piensan las máquinas?" pasarán a tener una relevancia filosófica mayor. El enfoque central de la ciencia cognitiva y las preguntas sobre la inteligencia artificial se dan de acuerdo con la información que tenemos en respuesta a la pregunta "¿Cómo piensan los humanos?" la cual parece dar cuenta de los requisitos generales para comprender el pensamiento. Al llevar a cabo la investigación en el marco que se ha planteado en este trabajo esta última pregunta no deja de ser sumamente importante, la diferencia es que ahora las primeras se encuentran a un nivel similar de relevancia.

Si mediante esta investigación se responde a la pregunta "¿Pueden pensar las máquinas?" pasará a ser una cuestión de la evolución de la investigación; si lo hacen o no será una cuestión que se derive de la exploración de la forma en que ellas lo hacen. Aún así, El preguntarse por la manera en que las máquinas piensan no quiere decir que se esté presuponiendo que lo hacen. Hay que recordar que el concepto de pensamiento se ha reducido, al menos en esta etapa de la investigación, a la simple manipulación de la información y la supervivencia conceptual del objeto.

En esta medida, al preguntarse sobre "¿Cómo piensan las máquinas?" en esta primera etapa de la investigación se está preguntando por "¿Cómo procesan las máquinas información más allá de la requerida para que las sigamos considerando como máquinas?". Esta última pregunta es mucho menos sospechosa como mecanismo para, a través de ella, determinar si las máquinas piensan. Sin embargo, en esta etapa de la investigación, ambas preguntas son funcionalmente equivalentes y, a medida que avance la investigación, el contenido de la segunda pregunta mutará constantemente mientras que el de la primera seguirá siendo igual.

Este cambio en la pregunta y la metodología de investigación también tendría consecuencias sobre las críticas que se han planteado de manera tradicional al problema de la AI. En particular, la mayoría de críticas deberían ser nuevamente formuladas si consideran que aún tienen cabida dentro de la discusión. Si una crítica ha pretendido mostrar la inviabilidad de la inteligencia artificial debido a su imposibilidad de replicar una experiencia del pensamiento humano, esta debería reconstruirse de tal manera que demuestre por qué esta característica de la forma humana de pensar es necesaria para todo pensar. El orden de la investigación cambiará, si procedemos por este medio no debemos encontrar en primer lugar las características relevantes en nuestro pensar humano para después extrapolarlas a aquellas mentes no-humanas; en lugar de esto se procederá a encontrar requisitos que permitan la inclusión de los diferentes sujetos que reconocemos como pensantes sin tener prejuicios sobre las características centrales que una mente debe tener.

El que la investigación cognitivista (entendida como la capacidad de replicar cerebro humanos en medios artificiales) pueda o no llevarse a cabo pasaría a ser irrelevante para la investigación sobre las capacidades de pensamiento en las máquinas. Esta sería una pregunta acerca de las capacidades computacionales, o de la habilidad de ingeniería de los programadores, en ningún momento una pregunta sobre la posibilidad de que la máquina piense. Si el cerebro humano no puede ser replicado en un medio artificial, esto simplemente sería una demostración de que la mente humana y el procesador de una máquina son estructuras diferentes, pero aún así la posibilidad de que la segunda lo haga de una forma diferente sigue abierta. Los estudios del cognitivismo tendrán valor al momento de estudiar el cerebro humano, no al momento de investigar las capacidades de las máquinas como piezas pensantes.

Este cambio en la manera de enfocar la investigación sobre la AI también trae consigo la consecuencia de realzar la importancia de los estudios sobre la AI débil. Esta forma de AI se ha dejado de lado al momento de producir estudios, observándola más como una herramienta de

trabajo que una forma de “Inteligencia”. Aún así en los niveles iniciales hay tantos criterios para eliminar de la investigación formas de AI débil como fuerte, y sin lugar a dudas la pregunta sobre “¿Cómo piensan las máquinas?” también se extiende sobre la AI débil. Por lo tanto, el que una máquina no se haya planteado con la intención de ampliar los conocimientos sobre los estados mentales, no será un limitante para extraer de ella información sobre lo que significa pensar.

3.3. Perspectivas de investigación.

Considero que la investigación que queda abierta por este trabajo puede dividirse en diversos campos:

- Para quienes mantienen la convicción de que el estudio del pensamiento es el estudio de la mente humana, queda pendiente la tarea de encontrar como aplicando los criterios exclusivos de la mente humana se puede conservar la posibilidad de la inclusión de las otras mentes (radicalmente distintas).
- La investigación sobre “¿Cómo piensan las máquinas?” puede derivar en investigaciones sobre “¿Cómo piensa el software?” y “¿Cómo lo hace el hardware?”. Estas son dos preguntas a las que se puede llegar al momento de estudiar los PC y las cuales pueden tener consecuencias diferentes en la investigación. Explorar estos aspectos de la investigación en computación puede llegar a traer consecuencias inesperadas sobre las formas en que puede ocurrir un pensamiento o sobre lo que significa procesar la información en diferentes niveles.
- El uso y la adecuación del equilibrio refractivo como método para establecer definiciones de conceptos en general puede mostrarse como un campo fructífero de investigación (así como algunas aplicaciones). Además, reconocer los efectos que la inclusión de información científica en el proceso de equilibrio pueda tener en la justificación de los resultados del método puede ser otro campo que se encuentra abierto para el estudio.

4. Bibliografía

Blass, A., & Yuri, G. (2006). Algorythms: A Quest For Absolute Definitions. In A. Olszewski, J. Woleński, & R. Janusz (Eds.), *Church's Thesis After 70 Years* (pp. 24-58). Ontos Verlag.

Daniels, N. (2011). Reflective Equilibrium. (E. N. (ed.), Ed.) *The Standford Encyclopedia of Philosophy (Spring 2011 Edition)*, URL = <http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium/>.

Dennett, D. (1991). *Consciousness Explained.* Boston: Little, Brown and Co.

Endicott, R. P. (1996). Searle, syntax, and observer-relativity. *Canadian Journal of Philosophy*, 101-122.

Gandy, R. (1980). Church's Thesis and Principles for Mechanisms. *The Kleene Symposium*, 123-148.

Goodman, N. (1955). *Fact, Fiction and Forecast.* Cambridge, Massachusetts: Harvard University Press.

Kelly, T., & McGrath, S. (2010, Spring). *Is Reflective Equilibrium Enough?* Retrieved Febrero 26, 2012, from Princeton University: http://www.princeton.edu/~tkelly/iree.pdf

Lycan, W. (1996). *Consciousness and Experience.* MIT Press.

Megil, J. (2012, Enero 19). *The Lucas-Penrose Rgument abour Gödel's Theorem.* Retrieved from Internet Encyclopedia of Philosophy: http://www.iep.utm.edu/lp-argue/

Moural, J. (2003). The Chinesse Room Argument. In B. Smith (Ed.), *John Searle: Contemporary Philosophy in Focus* (pp. 214-261). Cambridge, United Kingdom: Cambridge University Press.

Nute, D. (2011). A Logical Hole in the Chinesse Room. *Minds and Machines*, 431-433.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Pylyshyn, Z. (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence.* Ontario: Ablex Publishing.

Rey, G. (2002). Searle's Misunderstandings of Functionalism and Strong AI. In *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence* (pp. 201-225). Oxford: Oxford University Press.

Searle, J. (1980). Minds, Brains and Programs. *The Behavioral and Brain Sciences*(3), 417-457.

Searle, J. (1994). *The Rediscovery of Mind* (9th Print [2002] ed.). Cambridge, Massachusetts: MIT Press.

Shanahan, M. (2009). The Frame Problem. (E. N. Zalta, Ed.) *The Standford Encyclopedia of Philosophy (Winter 2009 Edition)*, URL = <http://plato.stanford.edu/archives/win2009/entries/frame-problem/>.

Siegel, H. (1992). Justification by Balance. *Philosophy and Phenomenological Research, 52*(1), 27-46.

Stich, S. (1990). Reflective Equilibrium and Analytic Philosphy. In *The Fragmentation of Reason* (pp. 75-100). Cambridge, Massachusetts: The MIT Press.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 433-460.

Vlatko, V. (2010). *Decoding Reality: The Universe as Quantum Information.* Oxford: Oxford University Press.

Wang, P. (2007, Septiembre). Three fundamental misconceptions of Artificial Intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, 249-268.