UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE FILOSOFÍA

Las reglas como los límites de la inteligencia artificial: Wittgenstein y la imposibilidad de una regla sobre la aplicación de reglas

TESIS

que para obtener el título de

Licenciado en Filosofía

PRESENTA

Juan Salvador Sandoval Romero

Asesor: Dr. Cristian Alejandro Gutiérrez Ramírez

Ciudad Universitaria, Ciudad de México septiembre de 2020

A Gloria, Salvador y Frida.

Agradecimientos

El presente no se dio en las condiciones idílicas, por lo que considero particularmente relevante el apoyo sin el cual no hubiese sido posible su realización. Quisiera agradecer a mi asesor Cristian Alejandro Gutierrez Ramírez, pues ha sido constante en mi formación desde (literalmente) el segundo día en la licenciatura, así como a mis sinodales: Alejandro Vázquez del Mercado Hernández, Norma Ivonne Ortega Zarazúa y José Francisco Barrón Tovar, y particularmente a Renato Huarte Cuéllar, pues a pesar de la premura, sus contribuciones fueron esenciales para que este trabajo fuera legible.

Agradezco también la ayuda del Grupo Integral de Lógica en acción (a todos y cada uno de sus miembros) y la revisión de Jennifer Pérez Dorante, la bibliografía sugerida por Samuel Alejandro Lomelí y el desvelo solicitado a Josué Raziel de la Rosa. Así mismo quisiera extender dicho agradecimiento a mis padres Gloria Guadalupe Romero Galván y Salvador Sandoval de Escurdia, por su apoyo, paciencia y cafeína, de igual forma a Frida Masiel Sandoval Romero, Nancy Gutierrez de la Sancha y Janeth Atwell tapia, quienes (a pesar del tiempo empleado) nunca dudaron de mi capacidad para terminarla.

Índice

Introducción

No es extraño que las raíces de la ciencia nazcan y mueran en las fantasías de la imaginación dormida. Es así como a menudo solemos dirigir los esfuerzos de nuestro ingenio. A lo largo y ancho de la historia, ha habido referencias históricas y culturales de estatuas que cobran vida por un halo divino, de seres creados a los que se les ha dotado de vida o alma, de lenguajes y cálculos capaces de encerrar el funcionamiento del razonamiento humano y máquinas futuristas cuasioraculares que presagian el fin de la humanidad: todas estas unidas por la idea de reproducir cualidades que parecen ser únicamente humanas por medios artificiales.

Pamela McCorduck escribe: “De ida y vuelta entre el mito y la realidad, nuestra imaginación ha suministrado lo que nuestros talleres no han podido, nos hemos dedicado durante mucho tiempo a esta extraña forma de autorreproducción [la inteligencia artificial].” (2004, p. 3). La autora considera al proyecto de la inteligencia artificial (I.A. en lo siguiente) una forma de “autoreproducción”, pues dicho proyecto pretende reproducir las características que suelen ser atribuidas a los seres humanos como bases de su identidad. Si bien es cierto que la búsqueda de lo que hace humanos a los humanos abre una veta filosófica propia, es claro que dichas fantasías implican la reproducción de sus aparentemente inteligentes comportamientos.

Describir los procesos históricos que implican la concepción e investigación de seres artificiales inteligentes como posibilidad sería extenso, tortuoso e innecesario para los fines del presente, pues podemos encontrar referencias tan antiguas e imprecisas como los relatos griegos sobre autómatas de bronce creados por el dios Hefesto, o tan modernas y específicas como las redes neuronales con las que funcionan páginas en internet, así como programas de edición inteligente que son capaces de detectar patrones en imágenes o tendencias en una nube de información. Debido a lo anterior el presente tiene por objetivo explorar los límites del proyecto de la I.A. en función de los límites de la programación de agentes artificiales inteligentes.

Tomando este eje como fundamento será conveniente determinar una división del campo en función de su orientación como primer paso: bien podríamos tomar la inteligencia humana o el comportamiento humano por objetivo, o bien podríamos tomar por objetivo el pensamiento o comportamiento racional. Los enfoques humanos se desarrollan al inicio del primer capítulo en ocasión de dos eventos históricos: las Conferencias Macy (1946) que dieron inicio al proyecto de la cibernética, así como el artículo *Computer Machinery and Intelligence* (1950) de Alan Turing. Los enfoques racionales se resuelven en la segunda parte del mismo. El segundo capítulo da cuenta de los elementos conceptuales que necesitaremos: la definición de agente, inteligencia y programación, así como las Conferencias Dartmouth de 1955 (McCarthy *et al*., 1955) que se mostró como eje articulador del proyecto moderno de la I.A., por lo que eventos como la arquitectura de John von Neumann y el mismo trabajo de Turing serán comprendidos como antecedentes fundadores.

Una vez que describa la mejor definición de I.A. que me sea posible dadas las fuentes revisadas, será necesario señalar los problemas a los que se enfrenta. Para obtener el comportamiento de un agente inteligente será necesario o bien encontrar mecanismos para la resolución de problemas, o bien un aumento considerable de recursos informáticos. Por un lado, tenemos el problema del seguimiento de reglas, que describe que no es suficiente con un conjunto de reglas para determinar su aplicación o, lo que es lo mismo, procesos estocásticos que determinen la relevancia de la información y, por otro lado, la complejidad de problemas de esta naturaleza, que al crecer de manera obscena, requieren de una demanda no realista de recursos. Esto será descrito en el tercer y último capítulo, dando paso a las conclusiones.

1. Inteligencia artificial: origen y planteamiento

Los términos que describen al proyecto de la I.A. no han sido constantes a lo largo de la historia ni mucho menos, así como tampoco lo han sido las prácticas, las ciencias auxiliares o el contexto que comparte con otros campos del conocimiento. No es claro el momento en que se origina dicho campo o las creencias de quienes lo configuran, pero describir la historia general de la I.A. o dar una definición completa de sus objetivos pasados y presentes no sólo es una tarea enciclopédica o una decisión metodológicamente cuestionable, sino también innecesario para comprender los límites que pretendo hacer claros. Sin embargo, parece ser claro que busca la creación de máquinas inteligentes, por lo que es difícil que no concediésemos que busca la construcción de seres análogos a los humanos. El problema aparece ante la pregunta: "¿qué es una máquina inteligente?".

Dado que muchas han sido las respuestas intuitivas a dicha pregunta, he tenido a bien utilizar la división temática expuesta en el primer capítulo de Newell y Simon (1995) sobre los principales esfuerzos de reproducción de agentes artificiales inteligentes, a saber, una distinción primaria entre el comportamiento y los procesos cognitivos internos de la máquina, así como una división secundaria entre la inteligencia humana y el ideal de inteligencia o racionalidad. De esta forma, tendremos cuatro posibles enfoques en función de lo que comprendamos como objetivo de la I.A.: máquinas que se comportan como humanos, que piensen cómo humanos, que piensen de manera racional o que se comporten de manera racional. La primera parte de este capítulo describe los primeros pasos del proyecto de I.A. que pretende una inteligencia humana, mientras que la segunda parte desarrolla los fundamentos de los enfoques basados en un ideal de inteligencia racional.

1.1 Máquinas de carne y hueso

No es fácil dar definiciones completas acerca de nociones tan vagas como podría ser la de inteligencia, por lo que se suele partir de acepciones generales del lenguaje común. Siendo esto así, es necesario hacer un análisis reflexivo desde la propia experiencia humana para poder describir la operación del concepto de inteligencia. Es cierto también que no es posible asumir que otros seres piensan de la misma forma en que parecen hacerlo los humanos, por lo que funge la conducta como el criterio a través del cual poder calificar algo como *inteligente*. Esto suele referir a comportamientos presentes en aspectos en la resolución de problemas que podría sugerir que un agente es inteligente o que existe un proceso inteligente detrás de las decisiones que toma.

A continuación, describiré dos proyectos fundacionales en el campo que, si bien no se agotan en los enfoques descritos, ofrecen un panorama por ellos configurado. Los dos momentos a continuación descritos nos permiten identificar a los humanos como máquinas de carne y hueso, por lo que ambos partirán de procesos y comportamientos humanos. El proyecto cibernético, por ejemplo, pretende descubrir los procesos cognitivos que los humanos llevan a cabo al momento de resolver un problema. En lo que a la postura de Turing respecta, el objetivo cambia en virtud de su metodología. Al ser el comportamiento observable el parámetro más relevante desde su perspectiva, no es extraño que plantee el juego de la imitación como una prueba que determina si es inteligente o no una máquina que pretende serlo.

1.1.1 Cibernética, pensar como humanos

Comprender el proyecto moderno[1] de la I.A. como la reproducción de máquinas que piensen como humanos implica un completo entendimiento de dichos procesos, así como la capacidad de reproducirlos en términos formales aplicables a la programación. Lo que se describe brevemente a continuación es el proyecto de la cibernética de primer orden: se gestó en la primera mitad del siglo XX y sostenía, precisamente, dicha comprensión de procesos cognitivos, así como su posible aplicación en máquinas pensantes. Asimismo, se describirán los trazos generales del cognitivismo y conexionismo, que describen la relación que la cibernética establece entre procesos mentales y operaciones computables.

Un conjunto de destacados pensadores de distintos campos del conocimiento se reunió en las Conferencias Macy con objetivo de crear una ciencia general sobre el conocimiento de la mente humana. Dichas conferencias tuvieron lugar en Nueva York de 1946 a 1953 y fueron publicadas en 2016 (Pias, 2016). Podemos comprender las conferencias como un acercamiento interdisciplinar a los flujos de información, los sistemas de control y reguladores que parecen operar en la mente humana. Este campo actualmente está suscrito en las ciencias cognitivas, pues a pesar de su resistencia a identificarse con la cibernética, existe cierta continuidad entre dichos campos.

Las Conferencias Macy son el primer intento por componer el campo de investigación acerca de la mente humana como es comprendida por la I.A., como un sistema estudiado con el objeto de ser reproducido. Existen diversas definiciones de cibernética y son empleadas de acuerdo con la pertinencia de cada caso particular. Es necesario tener en cuenta que, debido a su naturaleza inter y multidisciplinar (hay incluso quienes dirían “transdisciplinar”), el concepto de cibernética no puede agotarse en un solo marco. Aun así, existen puntos coincidentes como el interés sobre las estructuras formales de sistemas y modelos, el flujo de información y la

[1] Por “proyecto moderno de la I.A.” comprendo el proyecto de I.A. que se trazaba desde los albores de la década de 1940, pues al ser una noción tan amplia, parece imprudente no acotarla. A lo largo del presente será dibujado a través de procesos históricos que tuvieron lugar hasta décadas más recientes.

aplicación de herramientas formales a problemas que no fueron planteados originalmente en estos términos.

Para una breve descripción de los orígenes del proyecto cibernético, seguiré la exposición de Dupuy (2009), pues en la introducción de su obra, señala algunos puntos importantes para el desarrollo del presente. Dicha descripción no pretende ser una explicación completa o última de un proyecto tan general o amplio como es la cibernética. El proyecto inicial de la cibernética es consecuente con dos creencias: (1) el pensamiento es una forma de computación, por lo que se puede describir en términos algorítmicos mecánicos y (2) las leyes físicas observadas dan cuenta de los términos de significado, direccionalidad, intencionalidad y finalidad.[2]

Es importante advertir también que los primeros cibernéticos tenían la firme convicción de que comprender algo implicaba poder replicarlo o, en su defecto, poder hacerlo en condiciones idealizadas. Que esto fuese cierto implica que si comprendiésemos *absolutamente* el funcionamiento detrás de las manos humanas, por ejemplo, cada uno de los tendones, huesos, uniones cartilaginosas y fibras musculares, así como los mecanismos en virtud de los cuales funciona, podrían ser replicadas. Incluso si las limitaciones científicas actuales no lo hicieran posible, con la tecnología adecuada, podría replicarse una mano completamente funcional. La idea parece ser que cualquier cosa se compone únicamente de las partes que la integran y las relaciones funcionales entre dichas partes.

Dicho lo anterior, parece claro que el propósito es la comprensión de la mente en su totalidad. Esto resolvería algunos de los problemas que han atormentado desde siempre a la filosofía, específicamente el problema de la relación mente-cuerpo[3]. Es por lo anterior que hay quienes podrían concluir que la cibernética pretende humanizar a la máquina (aquellos que pretenden construir máquinas que tengan habilidades humanas), dotándola de características que denominamos exclusivamente humanas (i.e.: habilidades cognitivas complejas);

[2] Es posible profundizar en el significado de estos términos, pero para fines de la presente exposición, podrán operar con el significado que les asigna el lenguaje común.
[3] Resolvería el problema tomando por eje que no existe algo característico de la consciencia, pues la mente podría ser replicada fuesen encontradas las condiciones de un sistema material que rescate las conexiones abstractas.

pero parecería más bien que los primeros cibernéticos pretendían mecanizar la mente humana, dotándola de características denominadas exclusivamente mecánicas. Siguiendo esta línea de pensamiento, deberíamos operar bajo el supuesto de que somos una suerte de *máquina de carne*, partes del mundo que pueden ser comprendidas bajo los mismos principios que el resto y que no pertenecen a un estatus privilegiado e imposible de comprender en su totalidad (en contra de una posición dualista[4] de la mente). Podría ser útil pensar esto como una versión actualizada de la noción materialista de "máquina de carne".

Si se es consecuente con esta línea de pensamiento, será necesario preguntarse por la manera en que debería ser comprendido el el modo como se lleva a cabo este tipo de cómputo. Ciertamente no como la manipulación de símbolos abstractos, tal como Turing (1950) y algunos otros pensadores lo comprenderán: una cinta con instrucciones que da por resultado largas cadenas de procesos informáticos[5]. De esta manera, sostendríamos que la naturaleza de las computaciones de la mente humana está inserta en las redes causales del mundo. Esto significa que gran parte del peso de los procesos que lleva a cabo la computación reside en dichas relaciones causales y no sólo en las descripciones funcionales de los procesos.

Como he dicho antes, las posturas de la cibernética no pueden ser inscritas meramente en la filosofía tradicionalmente entendida, pues es un campo que nace de una colaboración interdisciplinar. Por ello es la psicología la que daría cuenta de la rama del conocimiento que sostiene esta suerte de funcionamiento del mundo, denominada como cognitivismo, alrededor de la década de 1960. Esta perspectiva inscrita en la psicología acorta la brecha existente entre el mundo físico y el del sentido. Es decir, mantiene una distinción entre lo simbólico y la realización material de lo simbólico, pues sostiene que los procesos sintácticos se materializan en los procesos causales dentro de un ordenador en tanto objeto físico y, a menudo, se

[4] No es útil para el presente hacer una descripción detallada de las teorías dualistas de la mente, baste saber que se comprometen con una ontología dual: la mente no corresponde a la misma sustancia que la realidad material del cuerpo.
[5] Profundizaré sobre el trabajo de Turing en la siguiente sección. Baste saber que esta será una de las nociones más relevantes en los aportes de su trabajo.

explica en términos de la analogía entre el funcionamiento de la mente y del de un computador[6].

Como consecuencia del cognitivismo tenemos teorías que comprenden a la mente de manera análoga a una computadora. Éste es el caso de la teoría computacional de la mente que la describe como: “un sistema computacional similar en aspectos importantes a una máquina de Turing, y los procesos mentales centrales (por ejemplo, razonamiento, toma de decisiones y resolución de problemas) son cálculos similares en aspectos importantes a los cálculos ejecutados por una máquina de Turing.” (Rescorla, 2020). Es decir, un sistema que computa una serie de procesos cognitivos. El computador es fácilmente relacionado a la mente en términos de modelo, pues podríamos llegar a pensar en imitar procesos cognitivos medianamente complejos o desarrollar secuencias de instrucciones (justo uno de los temas centrales de este trabajo es evaluar si esto es así). Es importante anotar que para cuando inicia el proyecto de la cibernética, aún no existían los computadores como los conocemos hoy en día, pues esto es producto del trabajo del matemático húngaro-americano John von Neumann[7].

Ahora bien, probablemente nadie sostendría que, del hecho de que una máquina sea capaz de ejecutar un programa, se puede derivar que la máquina es consciente de lo que hace o que surge en ésta una suerte de consciencia. Pero esto no implica de alguna forma, tampoco, que la consciencia sea sólo abstracta, pues para poder interactuar de manera causal con el mundo, es necesario que siempre esté aplicada en algún tipo de sistema material. En el paradigma cognitivista se ignora (en general) el problema de la conciencia (entendida como un estado fenoménico, apercepción, etc), y se apela únicamente a la intencionalidad (la capacidad de tener representaciones, como creencias, deseos, etc), dando maneras de caracterizar la intencionalidad de formas que no dependen de que exista una experiencia consciente.

---

[6] Utilizo este término de manera intencional para hablar de algo que podría ser una computadora o un dispositivo parecido con la capacidad de computar.

[7] Con esto pretendo señalar la arquitectura computacional de John von Neumann compuesta de: Unidad de Control, Dispositivo de Operación, Memoria y Dispositivo de E/S (Entrada - Salida). Esto posibilitó la implementación de conjuntos de instrucciones variables y la re-programación de la máquina. Originalmente descrita en: von Neumann, John (1945), First Draft of a Report on the EDVAC, Universidad de Pennsylvania, EEUU.

Atendiendo a lo anterior, surge una de las mayores aportaciones de la cibernética temprana: el concepto de red neuronal. Ésta es una herramienta formal que describe un sistema basado en neuronas idealizadas, una suerte de calculadora elemental que procesa unos y ceros en forma de uno o más datos de entrada y da un resultado por salida, es decir, un valor obtenido a través de la iteración de operaciones discretas aleatorias. Una neurona puede conectarse con otras y de esa manera modelar procesos complejos de procesamiento de información. Cada neurona opera como un nodo que modifica los datos de entrada de otras para crear una capa de operaciones que tiene por objetivo alcanzar un procedimiento efectivo que de una solución adecuada a una tarea específica. Lo anterior funciona como un conjunto análogo a las neuronas biológicas y permite formas simples de aprendizaje, ya sea supervisado o no. El conexionismo, sostiene que, podemos comprender a la mente humana como el producto de estas conexiones, al igual que las redes neuronales. Sostener una postura conexionista implica creer que con el diseño adecuado, es posible imitar el comportamiento del cerebro humano a través de un modelo de sus conexiones y comportamiento.

Es debido a esta suerte de comprensión de los fenómenos cognitivos y cerebrales, que la cibernética era en sus inicios, pensada como la teoría de los mecanismos teleológicos. La idea detrás de esto descansaba en el intento de reconciliar dos formas distintas de dar una explicación a los fenómenos en el mundo: la conocida y siempre avalada por la ciencia, explicación mecánica de causas y la teleológica, en función del *telos* o finalidad[8]. La primera de ellas articula una serie causal que relaciona eventos en una explicación de naturaleza mecánica, mientras que la segunda se refiere a la comprensión de los eventos individuales a la luz de una finalidad a la que se dirigen. De la misma forma, los cibernéticos creían que los así llamados mecanismos teleológicos imitan ciertos comportamientos sin apegarse al más riguroso sentido del término intencionales.

[8] La relevancia de las explicaciones teleológicas residen en una interpretación laxa de *explicación teleológica*: refiere a la simulación de las máquinas por aparentar una suerte de *intencionalidad* no necesariamente presente. No parecen decantarse por la explicación claramente contrastante con la ciencia y sus métodos explicativos.

Ahora bien, los hay quienes sostienen que esto requiere de la existencia de propiedades emergentes dentro de un sistema complejo como los descritos. Una propiedad emergente es una que surge sólo en la unión de los elementos de un sistema. Esto es relevante debido a que la consecución de un fin parece dotar a un sistema de características como intencionalidad, autonomía, direccionalidad y autoorganización.[9] Dupuy insiste en que uno de los puntos más polémicos del proyecto cibernético reside en la pregunta "¿son acaso las nociones de finalidad, intencionalidad, significado y direccionalidad una ilusión ocasionada por ciertas propiedades emergentes de los sistemas complejos?"[10], pero incluso si esto así fuera, lo más prudente sería actuar como si de hecho hubiese algo así como la finalidad en el mundo[11].

Podemos encontrar varias razones por las cuales la cibernética dio origen a muchas de las teorías y corrientes que actualmente se desarrollan en las ciencias cognitivas y campos de la I.A., pues dicho proyecto se basa en el problema expresado en la pregunta: ¿es posible traducir la totalidad de la mente en términos algorítmicos ejecutables por alguna máquina? Es necesario anotar que el objetivo de Dupuy es demarcar el origen cibernético de las ciencias cognitivas a pesar de su franca resistencia.[12] Sin definir el pensamiento como la única dimensión de la mente consciente, así como una suerte de computación, no parece posible de primera instancia dicha traducción.

Por último, para comprender adecuadamente este programa, tendríamos que concebir que la I.A. descrita en estos términos es posible y el objetivo del proyecto implica asumir que: i) el cerebro contiene de alguna forma el pensamiento humano, ii) si comprendemos el cerebro como una máquina que produce pensamiento, una

[9] La relevancia de dichas características descansa en que parece que determinan la manera en que se comporta un agente naturalmente inteligente. Son características que describen la manera en que comprendemos nuestro propio comportamiento.

[10] Un sistema se denomina complejo cuando de la unión de sus múltiples elementos genera información que de otra forma sería imposible, pues nace de la conjunción de las partes y no reside en ninguna de ellas de manera particular.

[11] Esto lo expone Immanuel Kant en la *Crítica del Juicio*, donde afirma que la finalidad debe ser supuesta para entender ciertas partes del mundo (en particular lo biológico), sin que ello comprometa a la ciencia con que de hecho exista (como una especie de principio heurístico).

[12] Al principio del texto especifica que, debido a su aparente fracaso a lo largo de la historia y los resultados arrojados, las ciencias cognitivas suelen desmarcarse del pasado que comparten.

I.A. será la emulación de la misma máquina a través de medio artificiales y iii) imitar los mecanismos causales funcionales de manera idéntica tendría que tener por consecuencia imitar el comportamiento de dichos mecanismos.

1.1.2 Máquinas de Turing. Actuar como humanos

Quizá no hay mejor ocasión para hablar del objetivo de la I.A como imitación del comportamiento humano que el artículo "Computer Machinery and Intelligence" (Turing 1950). Dicho artículo comprende el comportamiento humano como un parámetro válido para determinar la inteligencia maquínica, pues al ser imposible dar una respuesta simple debido a la ambigüedad de la pregunta[13], el autor decide reemplazarla por la siguiente: ¿puede una máquina imitar el comportamiento de un ser humano? Esta pregunta acota los términos que aparecen en la misma y elabora una definición matemática formal de máquina, que pretende ser general, pero no última, así como los parámetros de la imitación. Pretende, a través de la pregunta de la imitación, explicar la manera en que es posible una máquina inteligente.

A continuación, procede a describir el juego de la imitación. Esto es un juego en el que un interrogador C intenta, a través de mensajes escritos, determinar cuál de los otros dos jugadores (A,B) es hombre y cuál es mujer. La pregunta es entonces ¿puede ser diseñada una máquina capaz de jugar este juego de manera eficiente, una máquina capaz de hacer fallar el juicio de C tanto como podría hacerlo un ser humano? Si la máquina fuese eficiente en su tarea, debería ser difícil distinguir, no ya entre hombres y mujeres, sino entre máquinas y seres humanos.

Si antes determinamos las condiciones en las cuales se plantea la pregunta, también tenemos que determinar las condiciones del juego de la imitación. Para ello nos da una descripción matemática formal de máquina, una máquina de Turing, es decir, un dispositivo que manipula símbolos en una tira de cinta potencialmente

[13] Es decir, lo que comprendemos por pensar, por máquina y la capacidad que tuviese la segunda de hacer lo primero. Al igual que muchas otras preguntas en el campo, no parece tener una respuesta sencilla y última, pues implica la comprensión de problemas para los cuales parece que no tenemos las herramientas o no parece que las consigamos tener.

ilimitada conforme a un conjunto de reglas y para ello se sirve de una cabeza lectora que posibilita, en caso necesario, imprimir símbolos en la misma.[14] Dado su comportamiento automático, es un mecanismo que se encuentra en un estado determinado de entre un conjunto finito de posibilidades, es decir, una máquina de estados discretos. Una máquina de este tipo pasa de un estado a otro en función de una serie de instrucciones descritas en términos de relaciones causales y éstas funcionan en tanto conjunto de condiciones-consecuencia, siendo denominadas como programa. Una máquina de esta naturaleza claramente es automática, es decir, no depende de la decisión de un operador externo y puede resolver problemas específicos. Turing ejemplifica las máquinas de estados discretos de la siguiente forma: "Como ejemplo de una máquina de estados discretos podríamos considerar una rueda que gira 120 grados una vez por segundo, pero que se puede detener con una palanca externa; además, una lámpara se enciende en una de las posiciones de la rueda." (Turing, 1950, p. 6)

Es importante advertir la naturaleza formal de las máquinas de Turing, pues no pretende referir a un conjunto de máquinas físicas (incluso si podemos construir máquinas de Turing en el mundo físico), sino más bien un conjunto de máquinas posibles, pues existen máquinas de Turing que no serán implementables debido a las limitaciones materiales que presenta el universo. Como hemos dicho antes, dichas máquinas funcionan dando saltos entre estados de manera que se diferencien unos de otros y podamos predecir estados futuros con la suma de estados pasados y reglas con las que cambia entre estos.

Cada regla de transición en el programa tiene la forma de una cuádrupla de la siguiente forma: estado actual, símbolo, estado próximo y acción. En lo anterior, la acción está determinada por el estado actual, el símbolo y el estado próximo, siendo que si consideramos la cinta de la máquina, el estado actual es el número del cuadro, el símbolo su contenido, el estado próximo el resultado de ese conjunto específico de condiciones y la acción el siguiente movimiento del cabezal.

[14] De esta forma Turing se deshace de todos los elementos materiales de la máquina y nos presenta un modelo matemático con el que podemos trabajar a un nivel abstracto.

Ante la complejidad de ofrecer un concepto de máquina de Turing que cumpla con las características mínimas necesarias para jugar, Turing restringe el campo a computadores digitales. Un computador digital tiene tres características: almacenamiento, unidad ejecutiva y control. El primer elemento comprende el almacenamiento de información; el segundo, la capacidad de ejecutar operaciones basadas en la información del almacenamiento y, el tercero, las instrucciones de la máquina, un tablero que garantiza el correcto seguimiento de las instrucciones. Turing escribe: “Si uno quiere hacer que una máquina imite el comportamiento de un computador humano en alguna tarea compleja, se le debe preguntar cómo lo hace, y luego traducir la respuesta a una tabla de instrucciones.” (1950, p. 5) Los computadores digitales tienen la capacidad de ser programados para hacer más de una tarea. No están determinados como los mecanismos cuya configuración fija los obliga a ser rediseñados para ejecutar alguna otra tarea.

La introducción de la máquina de Turing implica la formalización de un proceso computacionalmente efectivo, lo cual es útil, por ejemplo, al respecto del problema de la decisión, pues preguntamos si existe algún proceso efectivo que pueda decidir si lo que describimos es o no teorema de la lógica de primer orden o de cualquier sistema en general. Es necesario anotar que algo es intuitivamente computable si existe una serie de instrucciones mecánicas adecuadas para llevarlo a cabo algo en un tiempo finito. Lo que plantea la tesis Church-Turing, la mencionada más arriba, es que si algo (una función) es intuitivamente computable[15], entonces es Turing computable (algo es Turing computable cuando puede ser procesado por una máquina como las anteriormente descritas).

Ahora bien, si esto es lo que de manera positiva obtenemos del trabajo de Turing, también es cierto que podría sernos útil destacar algunas de las objeciones

[15] Es claro que uno de los grandes problemas aquí es determinar cuándo una función es intuitivamente computable. Podríamos entender de manera formal que una función es intuitivamente computable cuando podríamos resolverla a través de una serie de instrucciones finitas, que puede ser realizada a través de la ejecución mecánica una serie de reglas. El problema se encuentra en la manera en qué es que sea mecánica. En palabras de Kevin Klement “Informally, we could say that a property or relation is “effectively decidable” if there is a purely mechanical or rote way of determining whether or not it holds for a given thing or given relata, something that could be calculated or computed without any special ingenuity or creativity or room for debate, using a process guaranteed to terminate after a finite number of steps. Similarly, we could say that a function is “effectively computable” if there is such a means for determining its value for a given argument or arguments.” (2020, p. 46).

que su trabajo recibió al ser publicado. Comprender la manera en que podríamos criticar esta postura podría arrojar luz sobre su operación. El mismo Turing ofrece, en su artículo de 1950, algunas de las objeciones más usuales a su trabajo[16] y posibles respuestas a ellas.

1. El argumento de la conciencia. La siguiente postura es resumida por la siguiente cita de Jefferson (1949), un reconocido neurocirujano de la universidad de Manchester:

> Hasta que una máquina pueda escribir un soneto o componer un concierto debido a las emociones y pensamientos que haya tenido, y no debido al uso de símbolos al azar, podremos estar de acuerdo que máquina es igual a cerebro, es decir, no sólo que lo escriba, sino saber que lo escribió. (Jefferson, 1949, p. 1110).

   Es claro que lo que comprendamos por inteligente determinará las líneas generales de cualquier proyecto en búsqueda de la I.A. que concibamos. Es también cierto que, si la consciencia resulta necesaria para la inteligencia, difícilmente podríamos dar cuenta de tan elevadas exigencias. No obstante, no parece ser este el caso, pues a diferencia de la inteligencia comprendida en este contexto, en sentido humano, la consciencia no se muestra necesariamente a través del comportamiento. Si fuésemos consecuentes con esta creencia, no podríamos asumir que todos los seres humanos son inteligentes, es decir, conscientes.

2. La objeción matemática. Esta objeción se debe al trabajo del matemático Kurt Gödel, quién, a través de los teoremas de la incompletud, mostró lo siguiente para todo sistema lógico completo: es posible que existan oraciones indecidibles, es decir, que no pueden ser probadas o desaprobadas. Si no existen estas oraciones, el sistema es inconsistente.

   Que esto es un límite posible en una máquina descrita en el juego de la imitación, es cierto. Existen preguntas que posiblemente no pueda responder

[16] Turing escribe en un contexto distinto y no parece necesario describir todas las objeciones que anticipa, argumentos como el de la existencia del alma poco o nada tiene que ver con el presente, por ejemplo. Es por esto que he incluido los que parecerían ofrecer resistencia.

cualquier máquina que concibamos dentro del juego de la imitación, es decir, computadores digitales. No obstante, no podemos perder de vista que no proponemos máquinas que excedan el comportamiento humano, y no hay pruebas tampoco de que estos límites no se apliquen al mismo. En palabras de Turing, "con bastante frecuencia respondemos equivocadamente como para justificar algún tipo de satisfacción por tener evidencia de la falibilidad de las máquinas" (1950, p. 11) Esta objeción, a lo más, podría probar la falibilidad de las máquinas, es decir, la capacidad de equivocarse, y no la humana.

3. La objeción de Lady Lovelace. Nuevamente nos enfrentamos a una objeción que pone en tela de juicio las capacidades productivas de una máquina. Dicho argumento niega que una máquina pueda presentar tendencias para hacer algo. Es decir, no es claro que exista una suerte de agencia independiente en la máquina, su capacidad de producir (cualquier cosa) parece provenir de un controlador externo que determina su conducta.

   Esto podría darse por verdadero, completamente, en tiempos de Lady Lovelace. Sin embargo, sucede que ha habido cambios importantes en la tecnología con la que podríamos contar, por lo que tampoco podemos descartar su imposibilidad. Incluso si esto no fuese el caso, de todas formas, sucede que no tenemos certeza de la posibilidad creativa de las máquinas en el futuro lejano. Es perfectamente posible, debido a la ausencia de pruebas, que los seres humanos tampoco tengan injerencia en las redes causales en las que están inmersos. No está claro tampoco que las máquinas se encuentren en la situación descrita, e incluso si así fuera, no es claro que los seres humanos se encuentren en una posición distinta, por lo que no parece negar la posibilidad de que dichas máquinas sean, de alguna forma, inteligentes (o que puedan llegar a serlo).

4. El argumento de la continuidad del sistema nervioso. Antes hemos caracterizado a una máquina de estados discretos. Hemos inquirido sobre la tajante diferencia entre los cambios analógicos de estado. Como tiene a bien

escribir Turing, "no se puede esperar que sea posible imitar la conducta del sistema nervioso con un sistema de estados discretos." (1950, p. 15). Esto se da debido a que las neuronas no son análogas, una diferencia mínima en la carga o frecuencia entre una neurona u otra implica una diferencia considerable en su operación.

Es claro, también, que dicha crítica presupone algunos de las líneas generales del proyecto cibernético, pues Turing no pretende simular las estructuras materiales que posibilitan el funcionamiento de la mente humana, sino su comportamiento general.

5. El argumento de la formalidad de la conducta. He dejado de manera deliberada a este como último porque tiene cierta semejanza a los problemas límites analizados en el tercer capítulo del presente trabajo de investigación. Dicha objeción sostiene que es imposible que una máquina se comporte como un ser humano, debido a que, si esto fuera posible, podríamos dar cuenta de las reglas de comportamiento de un ser humano. Es decir, podríamos caracterizar al ser humano de manera maquínica, pero este no es el caso.

   Si bien esto último es cierto y no podemos dar una lista acabada de reglas fijas que configuren nuestro comportamiento, sí que podemos hablar de leyes de comportamiento. Esto señala las tendencias independientes a nosotros, reacciones automáticas que sí parecen ser modeladas por una serie de reglas. Si bien éstas no posibilitan tampoco al proyecto, si resultan tener una injerencia fundamental en los términos en que se plantea el problema en el campo de la I.A.

1.2 La racionalidad de la I.A.

Los enfoques anteriormente descritos parten de un modelo humano imitable a través de diversos e interesantes métodos. No obstante, de manera independiente a

sus objeciones particulares, se gesta la siguiente pregunta: ¿de qué manera nos es útil una inteligencia *completamente* humana en un agente artificial? Es claro que, en otros campos, un esfuerzo de tal magnitud quizá podría ser valioso por sí mismo, pero no parece realmente este el caso, pues incluso de ser exitosa nuestra empresa, no es claro que alcance objetivos fundamentales que no obtengamos de manera más simple o cualitativa[17]. Con la intención de argumentar contra la versión más estable del proyecto de la I.A. describiré brevemente a continuación una noción orientada e ideal de racionalidad y los dos posibles enfoques que se dan en ocasión de ésta: agentes que piensen de manera racional y agentes que se comporten de manera racional.

Si bien la noción de racionalidad que aquí opera no pretende satisfacer una discusión filosófica sobre la naturaleza de la razón, cumple con una importante función metodológica: proyectar la inteligencia como una función no necesariamente inherente al cerebro, replicable sin el carácter de la naturaleza humana y accesible por medios ajenos a este. En este punto será posible deshebrar dos hilos de los que tirar: el pensamiento y el comportamiento racional. Mientras que el primero se compromete con la forma del pensamiento en general, el segundo se compromete con la proyección de la que antes hablábamos. Por ello, se concebirá como racional cualquier comportamiento que, dada la misma entrada, arroje la misma salida de datos, en este caso, percepciones.

1.2.1 Las leyes del pensamiento. Pensar de manera racional

No han escaseado nunca los intentos de describir la totalidad de nuestros procesos cognitivos como un conjunto de reglas, es fácil notar patrones emergentes aún en razonamientos sencillos. Podríamos asumir incluso que el razonamiento no es más que una manipulación lingüística de símbolos respecto a reglas, análogo a un idioma en dónde las palabras se ordenan en función de su gramática. Lo descrito

[17] Incluso sin recurrir a ejemplos hiperbólicos como un ordenador que se cuestiona por la naturaleza de la electricidad o una caminadora falta de motivación, dotar a un agente de pensamiento o de conducta humana podría ser incluso contraproducente, pues formas desarrolladas del pensamiento podrían por consecuencia dar lugar a inconvenientes afectos o aversiones: no es claro que podamos dibujar una línea entre los procesos estrictamente cognitivos y aquellos impulsos que consideramos intuitivos y poco útiles.

ahora es un esquema general histórico a través de los esfuerzos más relevantes para traducir el pensamiento a reglas generales abstraíbles.

Sin duda, el primer registro que ha llegado hasta nosotros de manera medianamente completa es el de Aristóteles y su construcción silogística de la lógica, que aunque no es completamente reducible a reglas, compone el primer sistema de razonamiento deductivo. Esto describe la relación lógica entre enunciados y por ello garantiza la transmisión de la verdad de las premisas a la conclusión.[18] Asimismo, Raymon Llull, un filósofo mallorquín, desarrolló en el *Ars magna generalis* una serie de descripciones profundas de la racionalidad y la lógica. Incluso describió ciertos dispositivos mecánicos que, implementando el resultado de sus consideraciones, superasen las barreras lingüísticas y permitiesen la discusión filosófica o teológica en función de las meras reglas lógicas del razonamiento. En consonancia con lo anterior, advertimos más tarde una exploración parecida por la idea de la *characteristica universalis* del filósofo germano Gottfried Leibniz, que consiste en un lenguaje universal diseñado para conceptos de naturaleza matemática y metafísica, cuya finalidad era parecida a la de las máquinas de Llull: atender a los procesos racionales del pensamiento y superar las divisiones lingüísticas, ideológicas o conceptuales. Todas estas posturas son compatibles (no se excluyen las unas a las otras o divergen de manera parcial) y comprenden el pensamiento racional como una suerte de manipulación de símbolos a través de reglas específicas que dan por resultado procesos lógico-racionales.[19]

Estas ideas, similares entre ellas, han permeado en distintos campos y han tenido por resultado corrientes como el logicismo, que podemos rastrear hasta el

[18] Es imposible presentar una descripción sencilla de su propuesta de racionalidad, pues ésta está inmersa en una teoría metafísica, una propuesta lógica y psicológica que dan cuenta de la mente humana en su totalidad, inserta en un contexto político específico, así como de ciertos campos del conocimiento. Incluso las leyes de la lógica, por ejemplo, el principio de no contradicción, se presentaban en una dimensión lógica, una psicológica y una metafísica. Así que no está claro si este principio rige a la racionalidad humana por alguna característica propia de esta clase de seres o porque describe la estructura de la realidad misma.

[19] Aquí una de las cuestiones centrales es poder decidir o clarificar en qué sentido un proceso debe ser considerado un proceso lógico racional. Tradicionalmente la racionalidad se ha asociado con el seguimiento de las reglas de la lógica, pero eso puede llegar a ser medianamente confuso, especialmente cuando queremos definir cuestiones relacionadas con la inteligencia en la visión tradicional. Entonces se considerará a un proceso lógico-racional si se realiza conforme con las reglas de la lógica clásica y de alguna u otra teoría formal que nos permita procesar información, por ejemplo la teoría de la probabilidad y, mucho más recientemente, la teoría de la elección racional.

atomismo lógico de Bertrand Russell. Dicha corriente del pensamiento sostiene que todo conocimiento (matemático) puede ser descrito por la lógica y que en última instancia se corresponde con *inputs* (entradas) de información del mundo. Pero es posible, en todas sus expresiones, sostener la creencia de que los comportamientos racionales son suertes de procesos de información, pues podemos determinar de manera clara lo que significa ser racional. El problema con esta postura es que, como veremos, nos compromete con creencias sólidas acerca de la naturaleza del pensamiento que son difíciles de sostener y requieren una investigación ulterior que excede los límites del campo de la I.A.

Hay al menos dos buenas razones por las cuales podría ser problemático decantarnos por este enfoque: 1) seríamos incapaces de recuperar todas las características del comportamiento que consideramos inteligentes, pues no todas las decisiones tomadas de manera racional implican una concienzuda ponderación de posibilidades derivado del razonamiento como seguimiento de reglas, ***i.e.*** retirar la mano del fuego, difícilmente diríamos que alguien es inteligente si pusiese la mano sobre una estufa encendida y deliberase si es mejor o no retirarla del fuego en lugar de sólo retirarla y 2) nuestro concepto de racionalidad no parece ser del todo claro. Un agente inteligente que simplemente logra deducir toda la información posible de un conjunto de datos (dados por el mundo, por ejemplo), difícilmente podría deliberar de manera adecuada un curso de acción posible sin saturar fácilmente sus recursos computacionales.[20]

Si, por otro lado, asumimos una descripción medianamente arbitraria de racionalidad, como la tradicional concebida en términos de la lógica, nos comprometemos con una postura que tiene consecuencias indeseables en dos sentidos: i) limitaría el posible comportamiento de nuestros agentes[21] a parangones muy estrechos y ii) permite clasificar a los mismos seres humanos como seres más o menos (o nada) racionales. Es decir, si consiguiésemos que un agente A se

[20] La necesidad de incluir procesos automáticos en la racionalidad ha sido planteada por los llamados teóricos de los sistemas duales, quienes *grosso modo*, sostienen que existen dos sistemas en la mente humana: el primero encargado de las respuestas automáticas y el segundo encargado de los procesos reflexivos irracionales. Esto será brevemente tratado en la siguiente sección.

[21] En la primera parte del segundo capítulo ofrezco una descripción técnica del término agente. Basta por ahora aclarar que podría decir en su lugar 'seres naturales', pues me refiero al conjunto de humanos y animales.

comportase de manera racional como consecuencia de pensar de manera racional ante un escenario hipotético y luego consiguiésemos que un agente B hiciese lo mismo, ¿no se comportaría exactamente igual un agente respecto al otro? No obstante, si no lo hiciese, no podríamos decir que ambos son del todo racionales. El *quid* de la cuestión reside en la gama de soluciones posibles dentro del pensamiento racional. En lo que a las personas respecta, permitir determinar su nivel de racionalidad sólo tendría por consecuencia desagradables posturas políticas, que podrían ensalzar por ejemplo a un conjunto de personas, europeas por ejemplo, frente a otras, digamos, no europeas, y generar teorías que lo justifiquen en virtud de su supuesta racionalidad.

Estas posturas tienen en común buscar un marco conceptual y completamente objetivo que les permitiese establecer discusiones y tener un método completamente confiable para poder resolverlas, un método que finalmente no debería depender de una postura ideológica particular. Si bien este sueño parece deseable, no es claro que reconozca la diversidad de pensamientos, ni que reconozca la posibilidad de que dos sujetos completamente racionales y bien intencionados puedan estar en desacuerdo genuinamente. De alguna forma parece que la postura de estos autores sugiere la existencia de una única y adecuada forma de razonar. El problema principal con esta postura será entonces que, si este sueño no es posible, parece poco probable que un proyecto de inteligencia artificial que trate de recuperar sus ideas pueda tener éxito. En otras palabras, si existen posturas encontradas pero igualmente racionales, no parece claro que nosotros podamos decidir entre ellas por métodos puramente lógicos. Si tal posibilidad es real, entonces parece que la Inteligencia artificial basada en lógica no contará con las herramientas suficientes para garantizar que su actuar es correcto en toda ocasión posible.

Dados los compromisos con los que nos atamos, sería sumamente complicado, sino imposible, mantener la postura aquí descrita. Incluso si encontrásemos un modo de hacer análogo el proceso de pensamiento a los procesos informativos de un programa de computadora, no es claro que podamos capturar el funcionamiento global de la mente humana. No obstante, no parece que

fuera un desacierto reemplazar nuestro concepto de inteligencia de parámetros humanos a parámetros racionales. Podríamos, de manera análoga a la perspectiva expuesta en la primera sección de este primer capítulo, intentar describir los procesos racionales como lo que posibilita un comportamiento racional. A continuación, describiré dicho enfoque.

1.2.2 Actuar de manera racional

De conformidad con lo dicho en la primera sección de este primer capítulo, hemos de descartar la posibilidad de un modelo de racionalidad ideal. No podemos determinar criterios de racionalidad fijos y unívocos que nos permitan tomar un (y sólo un) curso de acción posible. Lo anterior niega la posibilidad de crear mecanismos que ejecuten procesos racionales y obtener de esa forma un agente artificialmente inteligente, pero no niega la posibilidad de *emular* dichos procesos racionales. Será necesario partir del significado de actuar de manera racional. Como bien escriben Russell y Norvig, “un agente racional es el que actúa con la intención de alcanzar el mejor resultado, o cuando hay incertidumbre, el mejor resultado esperado.” (2004, p. 5) Tenemos entonces que la noción de inteligencia ya no parte del comportamiento humano, sino de una suerte de ideal de racionalidad, un parámetro con el cual podemos saber que un agente actúa de la mejor manera posible, dadas las circunstancias en las que se desenvuelve.

De adoptar esta postura, el objetivo de la I.A. será el crear un agente racional, es decir, un agente capaz de obtener información de su entorno, construir un modelo de dicho entorno y, a través de un cálculo de consecuencias posibles, determinar el mejor curso de acción para la consecución de una meta u objetivo. A diferencia del programa descrito en la sección anterior, no busca imitar los procesos involucrados en la cognición humana y tampoco busca sólo ejecutar procesos que operen de manera racional. Este enfoque del proyecto de la I.A. podrá valerse de mecanismos heterogéneos siempre y cuando puedan operar de manera coordinada para la satisfacción de los criterios de racionalidad que describiremos en lo siguiente.

Esta postura supone que en toda ocasión posible existe una mejor manera de actuar y que puede ser determinada por medios puramente racionales, incluso tal vez introspectivos. Sin embargo, ¿no requiere esto un examen ulterior? No podemos perder de vista que, para justificar esta opinión, es asumido un conjunto de ideas no explícitas y claramente no obvias. Esto nos lleva a una serie de problemas que podrían presentarse en el futuro del presente como posibles objeciones. No es claro que podamos encontrar una solución óptima en situaciones lo suficientemente complejas. Más aún, parece que el calificativo 'óptimo' sólo puede quedar establecido una vez que se establezcan una serie de parámetros bien específicos, como la solución óptima en un sentido concreto y bien determinado. En ese caso sólo podremos denominar a la opción más racional de manera subordinada a un objetivo. Podría ilustrar de manera adecuada el inicio de la película "Yo, robot", en la cual el detective Spooner, interpretado por Will Smith, encuentra inadmisible de la decisión tomada por un androide de salvarlo a él y no a una niña pequeña, bajo el argumento de que era más probable que él sobreviviera.

En este caso parece que uno de los grandes problemas, es que no es claro cuál sería la solución óptima en caso de no poder salvar a ambos. Por un lado, parece una solución óptima que la mayor cantidad de individuos tuviesen la mayor probabilidad de sobrevivir. Sin embargo, no parece que la supervivencia de la niña no fuese mejor para la comunidad que la del detective. ¿Qué parámetros deberíamos considerar en qué consideración y con qué relevancia? ¿Debería influir su ocupación, edad, estado de salud? En lo que a la niña respecta, ¿su valor debería considerar que era una persona joven y que probablemente tendría una vida más larga, tal vez más incluso más productiva que el hombre de mediana edad?

Para encontrar una solución óptima es necesario establecer los resultados más deseados, pues, si no se tiene claro cuáles son los fines, entonces no es posible establecer cuál es el mejor curso de acción a seguir. En el tercer capítulo discutiremos a profundidad las razones por las cuales esta información no se encuentra como un hecho en el mundo. No podemos determinar un curso de acción

razonable sin haber considerado nuestros objetivos y los problemas que le atraviesan. Así, parte de lo que está en juego es el sistema de valores bajo el cual se pueden establecer los resultados más deseables y calificar como óptimo un curso de acción que nos permitan obtenerlos al menor costo posible. Es claro que dicho sistema de valores no se puede establecer por medios meramente lógicos, además de que es muy probable que no exista un único sistema adecuado y que en realidad sea necesaria una serie de sistemas normativos para diferentes situaciones posibles.

Un segundo problema es que no es claro que un sujeto cognoscente sea consciente de todas las consecuencias posibles en su deliberación del mejor camino posible.[22] Además, es posible que, si los factores relevantes son muchos, se requiera de una enorme cantidad de datos sobre el entorno para poder hacer los cálculos pertinentes. Es posible incluso que lo que consideramos un comportamiento racional, sea mediado por factores que parecen ser independientes, como el orden en que se nos presenta la información o la manera en que se expresan, incluso de naturaleza cultural o social. Es plausible que nuestra elección de los factores relevantes para resolver un problema esté determinada más por una serie de prácticas aprendidas que por una reflexión puramente lógica o racional. Pero si esto es así y dado que hemos abandonado un modelo en donde las máquinas pretenden emular el funcionamiento de la mente humana, no es claro qué mecanismos podrían seguir nuestros programas de inteligencia artificial basado en lógica que puedan resolver el problema determinar las variables relevantes. Hasta aquí no hemos afirmado que esto no sea posible. Simplemente será un problema que el teórico de la Inteligencia artificial tendrá que enfrentar.

Incluso si pudiésemos resolver los dos problemas anteriormente descritos, aún tendríamos que enfrentarnos a un problema propio de la manera en que pretendemos ofrecer soluciones que ostenten un comportamiento racional: el tiempo de reacción. Imaginemos una situación en la que un agente considerado racional tiene que ofrecer una respuesta en un tiempo limitado y de no hacerlo tendrá un perjuicio grande. Imaginemos además que las limitaciones de tiempo le impiden

[22] Esto se analizará más detenidamente en el tercer capítulo.

llevar a cabo los cálculos pertinentes y decide no hacer nada. Difícilmente podríamos considerar racional su comportamiento. Esto representa un problema que exige solución inmediata y está relacionado con los recursos cognitivos de los individuos, así como el entorno que les rodea. Parece, entonces, que un ser racional debería ser capaz de considerar la relevancia de tomar una decisión dado un contexto particularmente apremiante. En este caso, el resultado debería aproximarse lo más posible al obtenido en una situación ideal, pese a la falta de información, la falta de recursos o la ausencia de parámetros claros. En lo que a elementos como tiempo, recursos y condiciones no idóneas, se refiere, no parece claro que la I.A. logre, ya no emular a un ser humano, sino simple y sencillamente existir en el mundo.

Podríamos incluso, desarrollando sobre nuestro análisis de las objeciones posibles, encontrar un cuarto problema. Sucede que diferenciar entre comportamiento racional y pensamiento racional, no parece posible para un observador externo. No parece obvio que pudiésemos determinar si un agente que se comporta de manera racional imita los procesos racionales de toma de decisiones o resolución de problemas, o si más bien lleva a cabo dichos procesos, pues sólo podemos acceder a su comportamiento, no a la suerte de estados mentales que lo configuran.

En este capítulo se presentó un desarrollo de corte histórico de los programas de I.A. basados en lógica en función de los objetivos de cada programa, así como los elementos a tomar en cuenta al momento de determinar los medios por los cuales cumplirlos. La investigación partió del origen de la cibernética en el contexto de las Conferencias Macy y desarrolló algunos problemas del proyecto de I.A. basada en lógica. Asimismo, se describieron cuatro importantes problemas del enfoque que la presente tesis sigue, y serán analizados con más detalle en los capítulos posteriores. Se estableció una conexión entre racionalidad y el seguimiento de reglas (lógicas) del pensamiento. Dicha relación fue criticada y puso

sobre la mesa una serie de objeciones, relacionadas con el tipo de recursos cognitivos que los seres humanos tenemos y que de alguna manera los agentes artificiales tendrían que lograr emular.

Uno de los problemas presentados más relevantes fue el limitado repertorio de herramientas que tiene el programa de Inteligencia artificial basado en lógica y se le plantearon una serie de retos que tendría que responder, apelando únicamente estas reglas. Por otro lado, tenemos problemas determinando las variables relevantes del contexto para analizar el mejor curso de acción posible. También fue planteada la necesidad de considerar el tiempo disponible para realizar o tomar una decisión. Fue expresado que los cursos de acción óptimos sólo podían establecerse a partir de un sistema de valores, que no iba a estar determinado por elementos puramente lógicos, cuestión que nos permitirá evaluar cuál sería el mejor curso de acción. En los siguientes capítulos se tratarán estos temas con un poco más de atención.

2. Entre la ciencia y la ficción: I.A. y programación

Se ha acotado un período histórico específico y han sido descritas características particulares del proyecto moderno de la I.A., por lo que ahora podemos establecer los parámetros de éxito o fracaso de los agentes racionales en entornos específicos. Una vez determinemos lo que consideramos un agente, será necesario describir las características con las cuales serán calificados como generalmente inteligentes en términos de comportamiento racional.

En este segundo capítulo mostraré la relación entre los términos agente artificial, máquina programable, inteligencia general y programación para poder afirmar lo siguiente: los límites de la programación respecto a una inteligencia general son los límites de un agente artificial generalmente inteligente y eso determina los límites del proyecto racional la I.A. Dado que la presente tesis se expone a partir del proyecto de I.A. basada en lógica, ésta será una propiedad posiblemente atribuible sólo a un agente artificial. Toda conducta será resultado de un programa en tanto secuencia de instrucciones automáticas de un agente programable. Tengo por objetivo mostrar en este capítulo que, si lo anterior es cierto, sucede una de dos cosas en el proceso para obtener la conducta más racional posible: o bien el programa asume todas las acciones posibles en su marco de acción así como sus consecuencias de manera indeterminada o bien determina un criterio de relevancia entre las opciones posibles en dicho marco para poder establecer una meta y los momentos relativos a esta.

2.1 Agentes, programas y funciones

Es difícil determinar el momento en que adquiere sentido preguntar sobre la inteligencia de una máquina. Es claro que existen máquinas que claramente sólo se mueven en virtud de las leyes de la física o el sistema de reglas del entorno en el que están inmersas, como un viejo reloj de péndulo o incluso un complejo ingenio autómata.[23] Pero, conforme las máquinas adquieren complejidad y presentan, de alguna forma, comportamientos observables, se convierten en posibles sujetos de estudio. Asimismo, dicha conducta es posible debido a una serie consecutiva de instrucciones que otorgan cierta independencia a nuestro objeto de estudio y adquiere la categoría de agente cuya serie de instrucciones denominaremos programa.

De manera análoga a prácticamente todos los términos discutidos en filosofía, agente, así como agencia, pertenecen a una compleja y ambigua discusión que involucra a los filósofos griegos y a los escolásticos. No obstante, la reconstrucción de dicha discusión abarcaría más tiempo del que sería prudente dedicar a estas alturas de la investigación. Si buscásemos una definición medianamente estándar, podríamos recurrir a un diccionario de filosofía “quién toma la iniciativa de una acción o aquel de quién emana o resulta la acción, en contraposición a *paciente* que es quién la sufre” (Abbagnano, 1993, p. 27) y definir *agencia* en función de esto como la capacidad de tomar la iniciativa de una acción. Este es un punto de partida para lo siguiente, pero resulta muy general para nuestros propósitos. Por ello, sin perderlo de vista, será necesario ahondar en ello. Describiré a continuación el término de agente, el de programa y los tipos en virtud de su programación.

[23] Los autómatas fueron muñecos humanoides de movimiento automático construidos y estudiados desde la antigua grecia, pero sumamente relevantes a inicios de la Edad Media, presentes en las cortes de la realeza europea y de diversos usos: desde el entretenimiento fugaz de los y las nobles de la corte, hasta la estafa sistemática de jugadores de ajedrez.

2.1.1 Agente: definición funcional

Turing (1950) apunta a un problema importante cuando intenta delimitar el conjunto de máquinas que deberían ser consideradas en el juego de la imitación y describe tres características que deberían cumplir:

> Es natural que queramos permitir el uso de cualquier técnica de ingeniería en nuestras máquinas. También queremos admitir la posibilidad de que un ingeniero o un grupo de ingenieros pueda construir una máquina que funcione, pero cuya forma de operar no puede ser descrita satisfactoriamente por sus constructores debido a que ellos usan un método que fuera experimental en gran medida. Finalmente, queremos excluir de las máquinas a los hombres nacidos de manera normal. (p. 437)

El problema es que no nos es posible dar un concepto que abarque estas tres características sin describir un tipo de máquinas en particular. Por esto, como hemos visto, da una descripción funcional de máquina (*máquina de Turing*), pero dar un listado de las máquinas sobre las cuales trabajaremos parece más bien tedioso y fácilmente superable a través del paso del tiempo. La importancia de la conceptualización, tal como es ofrecida por Turing, es que cualquier entidad que cumpla funcionalmente con la descripción pertenece al conjunto que pretendemos demarcar. Esto le permite abarcar no sólo las máquinas hasta ese momento existentes, sino también las máquinas posibles.

Podríamos con este objeto acudir a la definición general de la cual partimos, pero tendremos que *ser capaces de hacer algo* es un parámetro sumamente general y se presta fácilmente a inconvenientes ambigüedades. Por lo anterior, lo definiremos funcionalmente como una entidad que sea: “capaz de percibir su medioambiente con la ayuda de sensores y actuar en ese medio utilizando efectores” (Russell y Norvig, 2004, p. 37.). Por resultado tendremos un conjunto de entidades vivas y artificiales dentro de nuestro concepto de agente. Sin embargo, dada la naturaleza de nuestra investigación, seleccionaremos a los agentes no naturales bajo la clasificación ‘artificial’. Por lo tanto, el agente artificial al que pretendemos dotar de comportamiento racional será el objeto de estudio en el campo de la I.A.

Una vez claro esto, tendríamos que comprender a todo ingenio sensor que posibilite el flujo de información del entorno al agente como perceptor. En el caso de agentes humanos, por ejemplo, los oídos, los ojos o las manos podrían fungir como perceptores, porque perciben la información del mundo a nuestro alrededor, en forma de percepciones. Es decir, es a través de estos ingenios, que podemos obtener percepciones a través de la información que les es provista. De acuerdo con los autores que defienden esta postura, en el caso de las máquinas tenemos cámaras, sensores de luz, movimiento o calor, sonares y radares, entre otros. En lo que a efectores corresponde, habremos de hacer análogas las extremidades o la boca del ser humano, así como las extremidades, antenas o alas en otros seres vivos, con los efectores del agente artificial, como bandas eléctricas, tenazas o turbinas.[24] Siguiendo este hilo de pensamiento, toda percepción puede ser almacenada en un historial que determinará la forma en que el agente decide que hacer a continuación y llegados a este punto podríamos considerar dos escenarios. O bien esta información desencadena una serie de reacciones automáticas o bien se une a una memoria más compleja, pero opera de manera menos inmediata. Ambas serán determinadas por la naturaleza del programa.

Si somos consecuentes con lo anterior, diremos que un agente presenta, o no, inteligencia en virtud de su comportamiento, pero para ello es necesario que presente comportamiento que pueda llevar a cabo acciones de manera automática en virtud de sus percepciones. Diremos, también, que un agente es autónomo cuando no depende únicamente de la información que le es dada para tomar decisiones, sino que de alguna forma *decide* una serie de acciones que llevará a cabo en el futuro cercano. Si bien es cierto que no esperamos una autonomía absoluta, es cierto que debería poder ser capaz de tomar decisiones que permitan evaluar su comportamiento, a través del análisis constante de sus conductas. Un agente es capaz de presentar conductas y eventualmente un comportamiento (un conjunto de conductas) en función de su programación. En lo siguiente será

[24] Es importante resaltar que los críticos de estas posturas consideran que hay una distinción clave entre unidades que te permitan obtener datos del mundo y unidades que te permitan obtener representaciones del mismo. La diferencia puede parecer sutil. Sin embargo, es de la mayor relevancia. En un principio podría parecer que las cámaras y los ojos realizan la misma función. Sin embargo, los ojos, junto con nuestro sistema cognitivo completo, nos ofrecen representaciones del mundo muy probablemente mediada por nuestro aparato cognitivo o nuestro sistema conceptual, mientras que las cámaras nos ofrecen datos que aún no han sido procesados.

necesario prestar atención a la manera en que operan dichas conductas en virtud de los programas que las determinan, pues esto nos permitirá clasificarlos.[25]

Por lo anteriormente expuesto, será necesario indagar en el comportamiento del agente, es decir, su función y su programa, con el objeto de concretar los estándares de los agentes artificiales referidos. La función del agente es la descripción matemática que relaciona la información que le otorgan las percepciones recibidas y las acciones que les corresponden, por lo que suele representarse en forma de una tabla de casos y de esta manera explica cómo es que toma las decisiones. Mientras que en una tabla listamos todos los posibles estados de los que las percepciones *informan*, en otro tenemos una columna de reacciones relacionada. Ciertamente, mientras más percepciones considere la tabla, más grande será y dichas tablas suelen por lo general volverse sumamente complejas ante una adición relativamente pequeña de casos.[26]

Por otro lado, la función del agente (es decir la manera en la que opera en el mundo) describe el comportamiento del agente a través de una abstracción matemática de dicho comportamiento, por lo que es importante no confundir la función del agente con su programa. El programa del agente es el conjunto de la ejecución automática de reglas, por lo que, como bien advierten Russell y Norvig (2004), el primero es una descripción matemática del comportamiento del mismo, mientras que el segundo es el conjunto de reglas aplicadas al comportamiento que la función describe.[27]

[25] Es importante advertir que el comportamiento, en el paradigma que está siendo expuesto, deberá ser descrito en términos funcionales.

[26] Hay que destacar, como ya se hizo en la nota anterior, que la descripción ofrecida del comportamiento del agente artificial es completamente funcional y por lo tanto es descrito por una tabla de estados. Esta tabla de estados también puede presentarse como un programa en el contexto de las máquinas de Turing.

[27] Nótese una semejanza notoria entre una máquina de Turing programable y un agente artificial programable. También es importante notar que la distinción, nos permite tener diferentes programas que den como resultado los mismos comportamientos. Un ejemplo de esto podrían ser los emuladores de consolas que, mediante un programa distinto, logran correr el mismo videojuego (o cualquier otro programa) en otros soportes físicos.

2.1.2 Entornos y tipos de agente

Una vez descrito el agente como una pieza fundamental en el estudio de la I.A. lógica, podemos discutir los mecanismos que posibilitan su comportamiento. Estos mecanismos serán determinantes a la hora de clasificar a los agentes en distintos tipos: podríamos partir de las características del entorno en el que se desenvuelven o de la naturaleza del programa que los configura. Asimismo, será necesario describir nociones fundamentales para el presente, como "programación" o "entorno", para después describir las relaciones por las cuales están unidos de manera intrínseca. Para conseguir esto me serviré del segundo capítulo de Russell y Norvig (2004), pues no he encontrado una descripción más adecuada para los fines del presente en la bibliografía especializada.

Antes hemos dicho que un agente será definido en relación con el entorno en que se desenvuelve, pues de dicho entorno determinan las capacidades esperadas del mismo.[28] Normalmente suele ser asumido que el entorno es una versión de la prueba de Turing, pero es claro que conforme se reconozca la variedad de entornos posibles, comprenderemos objetivos distintos a los de actuar como humanos. Por ello, es importante capturar y clasificar la naturaleza de las características del agente en función de las cuales clasificaremos su comportamiento, así que distinguiremos cuatro puntos: las percepciones (los datos ordenados que percibe el agente del mundo), las acciones (que ejecutará el agente por medio de sus efectores), los objetivos particulares (lo que queremos que ejecute el agente, que a su vez configura nuestros criterios de racionalidad), así como el entorno en el cual se desenvuelve.

Para ejemplificar lo anterior, consideremos un agente que tenga por objeto encontrar patrones potencialmente peligrosos en radiografías o tomografías. En este caso, las percepciones probablemente sean los píxeles de las imágenes del estudio clínico, los objetivos sean reconocer patrones o figuras que coinciden con los ejemplos usados para entrenar al agente, los objetivos serían separar y clasificar

[28] En la siguiente sección de este primer capítulo ahondaremos en las razones por las cuales no podemos determinar criterios absolutos de racionalidad, por lo que los objetivos que miden el éxito de los agentes son relativos al entorno, de esta manera determina sus capacidades.

estas imágenes y asignar un índice de riesgo a cada una y el entorno sería una suerte de biblioteca con cientos de miles o millones de imágenes pertenecientes a estudios clínicos, asociados a nombres y descripciones de posibles pacientes.

Asimismo, podemos clasificar la naturaleza de los factores del entorno en cuatro rubros que inciden en nuestro juicio acerca de su éxito o efectividad: las medidas de rendimiento (es decir, los parámetros con los cuales evaluaremos el progreso de sus acciones, asociadas irremediablemente al objetivo de la creación del agente), las características del entorno, los efectores y los sensores del agente. Trataremos estas características por medio del acrónimo REES (rendimiento, entorno, efectores[29] y sensores) y denominaremos como "descripción REES" a la descripción que da cuenta del entorno de trabajo que determina su diseño. Será necesario prestar atención a las características posibles de los agentes artificiales de los cuales partimos, pues, como veremos en lo siguiente, podemos tener varios tipos de agente distintos de acuerdo con las vicisitudes de cada problema.

Dada la descripción funcional ofrecida de un agente artificial programable, será de suma importancia el entorno en el que se desarrolla, pues éste determinará las características del diseño del agente. Si el entorno nos ofrece una mayor dificultad, nos veremos obligados a aumentar la complejidad de los factores con los que interactúa el agente. Muchos de los problemas con los que lidiamos respecto a la complejidad de nuestras acciones, son producto de la complejidad del mundo en que nos desenvolvemos y los factores que deberíamos tener en cuenta para resolverlo. Los programas, entonces, se pueden dividir en cuatro: reactivos simples, reactivos basados en modelos, basados en objetivos explícitos y en utilidad. En ese orden de acuerdo con su complejidad. Un agente reactivo simple se basa en reglas condicionales de la suerte de los condicionales materiales lógicos: establecen un requisito o suma de requisitos que, de ser cumplidos, tienen una acción como respuesta. Un agente programado de esta forma tendrá reacciones parecidas a los actos reflejos humanos. Los sensores captan al mundo, lo unen a las reglas de condición-acción y devuelven una respuesta.

[29] La traducción original del texto escribe 'REAS' debido a que utiliza el término 'actuadores' en lugar de 'efectores'.

Podemos añadir un modelo interno al agente. En este caso, los agentes reactivos basados en modelos añaden una representación interna del mundo que se actualiza de acuerdo con las partes a las que puede acceder, pues no tienen una visión absoluta. Así, la información almacenada debe dar cuenta lo mejor posible, de las partes del mundo a las que no tiene acceso. Asimismo, tiene que haber mecanismos que puedan actualizar los cambios de la representación interna del mundo y mecanismos que predigan los resultados, que son consecuencia de las acciones llevadas a cabo por el mismo agente. La diferencia entre ambos tipos de agente reside en que el primero no tiene forma de aprovechar este historial para la actualización del estado del mundo y el segundo implica una mejora porque nos proporciona dos fuentes de información: 1) las situaciones independientes al agente y 2) las consecuencias que el agente lleva a cabo, motivado por la primera fuente.

No obstante, no siempre es simplemente el estado del mundo suficiente para saber qué hacer. Como he dejado claro antes, no es posible que el programa determine sus propios objetivos[30] y tenemos que determinar los parámetros con los cuales lo determinará quien determine el programa del agente. Los agentes del tercer tipo, agentes basados en objetivos explícitos, harán el cálculo necesario para la toma de decisiones en función de un resultado asociado al curso de acción, Podríamos pensar en un agente encargado de distribuir de manera equitativa el consumo de agua en un sistema de viviendas. Parece que, dado el objetivo del programa (distribuir esta agua de manera equitativa), acotar el uso que cada usuario tiene en una casa en particular es una buena decisión. Si pudiésemos determinar que, de mantener un riego constante y duradero, obtenemos un resultado positivo, el agente seguiría este curso de acción. No obstante, podría suceder que una de las vías centrales del sistema sufriera una avería repentina, por lo que, en atención al resultado asociado, lo más racional sería suspender el servicio en todas las casas, pues es la única forma en que reciban una cantidad equitativa de agua, ya que no habría forma en que el ala de la vía averiada pudiese suministrar más que un valor de 0 unidades de agua.

[30] No se han enlistado las razones por las cuales es imposible que podamos programar a un agente con dicha característica, pero no es necesario, en el tercer capítulo describimos con ese fin el argumento del lenguaje privado (ALP, por sus siglas).

Los agentes basados en utilidad siguen un camino parecido al descrito en los agentes que atienden a objetivos específicos asociando un valor a los resultados de los cursos de acción posibles. La utilidad podría ser descrita en este contexto, como un cálculo de las consecuencias posibles en cada curso de acción. Una vez hecho dicho cálculo, podría determinar no sólo un resultado posible, sino además determinar cuál es *el mejor* camino posible para cumplir el objetivo. En nuestro ejemplo anterior, el agente encargado de distribuir el agua probablemente pueda calcular que de reparar la avería (o notificar de la necesidad de su reparación) podría obtener un mejor resultado que de conseguir una distribución de 0 unidades de agua.

Es importante advertir que, aunque la sucesión de estos posibles agentes pareciera una progresión lineal o una suerte de evolución maquínica, no necesariamente lo sería. Es decir, es cierto que podrían subsumirse los unos a los otros en atención a la complejidad de los elementos que les integran, pero eso no implica que siempre fuera necesario conseguir los agentes más complejos para la resolución de todos los entornos posibles. Esto se vuelve más claro cuando consideramos una noción de recursos: consumir recursos innecesarios es cuestionable en términos de eficiencia. Dicho lo anterior, el siguiente apartado dará cuenta de las nociones de inteligencia general y programación y describirá la configuración de su relación.

2.2 Inteligencia general y programación

En la sección anterior de este segundo capítulo, hemos descrito la manera en que comprendemos a un agente, la manera en que se encuentra determinado por el entorno en que opera, así como la injerencia que tiene el programa en su comportamiento. Así mismo, se ha enfatizado la manera en que procederemos con vistas al tercer capítulo: la presente presupone un enfoque racional de la I.A. lógica, por lo que la manera en que se configuran los agentes resulta esencial. No obstante, hay un término que atraviesa la discusión de manera transversal y que no se ha especificado de manera debida: inteligencia. No sería conveniente considerar “inteligencia” de todas las maneras posibles, tendremos que utilizar una noción que

describa adecuadamente el contraste que pretendo hacer notar. Es por ello que utilizaré una noción particular: la *inteligencia general*. Esto es debido a que resalta de manera eficiente los elementos presentes en una idea intuitiva de inteligencia, que por alguna razón u otra, argumentaré, escapan de las capacidades de los agentes descritos mediante la I.A. lógica.

Si bien el término de "inteligencia" atraviesa de manera transversal todo proyecto posible de I.A., el de "programación" atraviesa todo problema vinculado con las capacidades maquínicas de un agente (en tanto que sea programable). En el contexto de la discusión sobre el proyecto de I.A. lógica, resulta especialmente relevante, puesto que pretendemos *programar* un comportamiento inteligente. En la segunda parte de esta sección describiré el concepto de programación operante y algunas de las características a tomar en cuenta, cómo la noción de computabilidad. Con esto describiré un primer nivel en el cual el proyecto se encuentra limitado, puesto que el capítulo siguiente describirá a detalle y con ejemplos, un segundo nivel, en el cual también nos constriñen los límites propios de la naturaleza de la programación.

2.2.1 Inteligencia general

Si configurásemos un agente que se comportase como lo haría un perro, por ejemplo, ¿sería considerado inteligente? ¿Quizá sería acotadamente inteligente? Es claro que no sólo reconoceremos que es un ser animado, pues también identificamos un molino como un ser que presenta movimiento, sin ser un ser viviente. Parece ser que podríamos otorgar a un animal cierto nivel de inteligencia por sobre el de un objeto inanimado o una planta. A pesar de que no existe un desarrollo histórico lineal (una secuencia de eventos históricos que nos permita rastrear el término de manera unívoca) del término "inteligencia", conservamos una suerte de nube de características con las cuales podríamos identificar una conducta inteligente y compararlas con otras, como en el caso de los animales y las plantas. Pero, ¿es esto suficiente para elaborar un concepto normativo de inteligencia? Es decir, un concepto que nos permita abstraer las características de un ser inteligente.

Muy probablemente no baste con esta idea intuitiva, pero podría ser un buen sitio desde el cual comenzar.[31]

He obviado hasta este punto el término de inteligencia, pues no parece operar en los agentes (artificiales) descritos como lo hace en el lenguaje coloquial y esto requeriría una investigación ulterior (claramente acotada en la presente sección, pues es posible desarrollar toda una tesis al respecto de dicho concepto). A continuación, describiré brevemente la noción de inteligencia general, contrapuesta a una versión más común, quizá más intuitiva, de inteligencia. Quisiera enfatizar de manera adecuada la importancia de la generalidad de la noción expuesta, pues determina en un espacio negativo el uso que he hecho antes del término "racionalidad". La inteligencia general, a diferencia de otras nociones más intuitivas, atribuye inteligencia a los agentes que presentan la capacidad de adaptarse a problemas que nunca antes se les habían presentado.

Mientras otras nociones de inteligencia privilegian la maestría del uso de una herramienta (ya sea una herramienta cognitiva o una habilidad en particular, típicamente herramientas de naturaleza lógico-matemáticas) para enfrentarse a problemas específicos, la inteligencia general privilegia la capacidad de usar adecuadamente diversas herramientas a problemas originales que no se habían presentado antes o para los cuales no se habían empleado dichas herramientas. En muchas ocasiones se evalúa a un individuo como inteligente si es increíblemente hábil en la aplicación de un programa o un tipo de solución particular, ante un tipo de problemas igualmente particulares.[32] No obstante, dicha maestría no será rasgo de inteligencia, puesto que no es medido el dominio, sino la adaptación a diferentes contextos, así como la elección de las mejores herramientas, para enfrentar los problemas característico de cada uno de ellos. Incluso si lográsemos contemplar

[31] Ofrecer una caracterización adecuada de la noción de inteligencia, es un trabajo que teóricos como Turing han considerado increíblemente complicados. Recordemos que en el primer capítulo se presentó la propuesta de Turing, y se hizo explícito que él prefirió aplicar un modelo de imitación del comportamiento humano, antes de comprometerse con una noción particular de inteligencia. En este sentido, debo aclarar que reconozco las limitaciones (incluso propias) que tiene este trabajo, por lo que sólo pretendo ofrecer algunas características que me permitan hablar de inteligencia general y no proponer una solución al mismo, ni mucho menos.

[32] Usualmente se favorecen contextos en los que la mayoría de la población tiene dificultades para desempeñarse, influenciados naturalmente por factores sociales o culturales. En una sociedad moderna, típicamente se atribuye una mayor inteligencia a aquellos que tienen mayores habilidades de razonamiento lógico-matemático.

todos los casos posibles, no podríamos garantizar que el agente pudiese adaptarse de mejor manera[33], sino que habría incrementado la biblioteca de recursos con la que cuenta.

Cuando añadimos el adjetivo “general” al concepto de inteligencia, abrimos los criterios por medio de los cuales podríamos determinar a un agente como inteligente.[34] Esperaríamos de un agente generalmente inteligente un comportamiento general análogo al de los seres humanos, quienes pueden aplicar soluciones previas en atención a la resolución de problemas nuevos planteados de formas en las que no se habían presentado antes. Si antes se han hecho explícitos los criterios del entorno para determinar la racionalidad del comportamiento, ahora la adaptación a este entorno opera como criterio para determinar un comportamiento inteligente. La eficiencia con que logran la implementación de soluciones previas, la categorización de los problemas nuevos y el grado de eficiencia de dichas adaptaciones, serán factores a tener en cuenta para determinar su éxito o la posible inteligencia que ostenta su comportamiento, en este caso.

En algún sentido parece que abandonamos ciertas intuiciones sobre lo que es ser inteligente, pero no es el caso. Lo que sucede es que, en ocasiones, cuando es nombrado inteligente un individuo, lo es en el contexto particular en el que se desenvuelve. A pesar de que dicho comportamiento es nombrado inteligente, no es suficiente (probablemente ni siquiera necesario) para atribuir inteligencia general a un individuo. Podríamos, en contraposición, calificar lo anterior como “inteligencia particular”.[35] La apuesta por una noción de inteligencia general es trascender un contexto particular y tratar de hacer una evaluación, probablemente no muy precisa,

[33] En 1957 (H. Simon, *et al*.,1959), con el objetivo de resolver problemas para los que no había sido programado previamente Herbert Simon, John Clifford y Allen Newell crearon el programa GPS-I (General Problem-Solver I). El GPS-I no pretendía desarrollar procesos creativos que le permitieran adaptarse a los programas, sino almacenar las soluciones posibles y aplicarlas en función de las características del problema. El programa buscaba la manera de reducir la diferencia entre el estado inicial y el estado-solución del problema (es decir, determinar cuál era el estado inicial y describir el estado-solución, para después aplicar soluciones posibles que tuviesen por objetivo alcanzar el estado ideal o al menos reducir la diferencia entre ambos).

[34] En la segunda parte de la primera parte del segundo capítulo del presente.

[35] Medir la inteligencia es increíblemente complicado y es muy plausible que en ciertos contextos podamos considerar a alguien increíblemente inteligente, pero que esa misma persona sea considerada brutalmente inepta en otros contextos, debido a que no ha elegido las herramientas correctas para desenvolverse en ese contexto.

del comportamiento general de un individuo. La generalidad de la inteligencia en este contexto describe la capacidad de adaptación sobre la capacidad de resolver problemas.[36]

Reconocer un problema y encontrar soluciones de problemas anteriores que fueran útiles en virtud del tipo de problema para luego aplicarlo de manera total o parcial es una tarea que requiere una gran cantidad de habilidades. Dichas habilidades no pueden ser descritas en términos funcionales, pues implican oscuros procesos heurísticos[37], creativos y aleatorios que se resisten a operar de manera regular en entornos no controlados. Es de esta manera que la noción de inteligencia general hace aparecer un espacio vacío en virtud del cual describimos a los agentes programables racionales, propios del proyecto de I.A. lógica. La generalidad de la inteligencia atañe a la noción operativa de inteligencia. Ahora procederé a explicar los problemas que atañen a la noción operativa de "programa".

2.2.2 Programación e I.A. lógica

Es claro que no es posible comprender la I.A. como la he descrito sin comprender cómo es que funciona un programa. La programación parece encontrar sus límites en la distinción entre los problemas que son (o no) computables, por lo que será necesario revisar la operación de dicha distinción. Es precisamente en virtud de la codificación del problema a resolver (el cual será o no computable), así como el procedimiento de su respuesta que podemos hablar de la posibilidad de agentes racionalmente inteligentes. De esta manera, los límites de la programación serán los límites de mi investigación.[38]

[36] Esto es claro cuando tratamos de establecer las reglas de un juego, ¿de qué manera podríamos *ganar* un juego cuya finalidad es, por ejemplo, botar la pelota mientras sea divertido? Un agente artificial (o un agente de inteligencia no general) no podrá sino dar procedimientos satisfactorios para juegos de finalidad clara, pues son problemas análogos a ocasiones previa**s.**

[37] Un proceso heurístico es un procedimiento no algorítmico flexible que opera de manera particular en cada problema presentado. No es esencial desarrollar más su definición en el presente.

[38] Esto podría parecer una alusión wittgensteiniana a la proposición 5.6 del *Tractatus Logico-Philosophicus* (originalmente involuntaria), pero dicho así parece tener sentido, pues el lenguaje de programación determina de qué cosas puede hablar (con sentido) la máquina. En este paradigma, parece ser que los límites del lenguaje (de programación) serán los límites del mundo (de la computadora).

Un programa está compuesto de instrucciones secuenciales y mecánicas[39] que ejecuta una máquina capaz de decodificarlo. Existen dos niveles en los cuales esto condiciona los límites de nuestra investigación: 1) limita los problemas que podría resolver una hipotética I.A. que cumpliese correcta y cabalmente todos los criterios con los cuales intentemos probar su inteligencia, a través de las categorías "computable" y "no computable"[40] y 2) opera como condición de posibilidad de los problemas dados en virtud de la naturaleza de sus objetivos. Es decir, problemas que ponen en evidencia las características del problema, de la I.A. en este caso, que se resisten a un proceso efectivo para su resolución. Lo anterior me empuja a trazar las líneas de estos límites que, si bien no son fundamentales para los límites explorados en el presente, sí lo son para determinar el campo en el que estas teorías se desarrollan. Con este fin, ofreceré una breve caracterización de los límites de la programación en términos de computabilidad.

Un programa, en un sentido medianamente restringido, consiste en la iteración de una serie de funciones específicas de manera matemática. Antes he descrito brevemente la operación de una máquina de Turing, pero es conveniente también en este punto de la exposición. Dada la naturaleza de dicha máquina, es posible cargar en ella una serie de instrucciones (describir reglas en la cinta en virtud de la cual opera) y estas instrucciones simples (serie de funciones específicas) pueden unirse (en una serie iteradas de manera matemática) unas con otras para integrar un programa. Todos los problemas susceptibles de formulación e implementación en una máquina de Turing serán Turing-computables.

Turing mostró la imposibilidad de un método efectivo para decidir[41] toda oración matemática adecuadamente formalizada, pues hay problemas que son (Turing) computables y no son decidibles (problema conocido como

[39] Aquí pretendo recuperar una noción preteórica no formal de procedimiento efectivo.
[40] Este problema, como su nombre lo indica, abarca incluso un mayor radio: determina qué problemas pueden o no ser *calculados*, pero no parece ser necesario para el presente seguirlo hasta sus últimas consecuencias. Basta con tener en cuenta que es incluso más grande que la lente con la cual lo analizamos en este momento y que estas son limitaciones para un programa, pero no necesariamente para un agente, ya que el agente puede estar hecho de diversos programas trabajando paralelamente y algunos de ellos pueden consistir en heurísticas. Es decir, nuestra mente sería una especie de enjambre de abejas independientes, trabajando en conjunto, lo cual iría en línea con la hipótesis de modularidad masiva.
[41] Aquí "decisión" es un término técnico que alude a las *funciones de decisión*, que asocian a cada relación 1 o 0 en virtud de si es el caso o no.

Entscheidungsproblem[42]). De esta demostración nace la *Tesis Church-Turing* (física), que indica lo siguiente. Si una función *F* es una función computable física o intuitivamente, entonces es Turing-computable (es decir, es implementable en una máquina de Turing universal). Es medianamente claro que una tesis de esta forma es un condicional material cuantificado universalmente (en el lenguaje de la lógica cuantificacional), por lo que su prueba consistirá en mostrar que, no existe una función física o intuitivamente computable *F* y que no sea Turing computable. Si bien podríamos decir que una función Turing computable es física o intuitivamente computable, no es claro qué tipo de evidencia podría mostrar que, si una función es física o intuitivamente computable, entonces es Turing computable. Si bien antes dijimos que, de acuerdo con algunos enfoques, podríamos ser considerados *máquinas de carne*, es decir, que pudiésemos ser copiados en términos funcionales y operar de la misma forma, no es claro para todos que podamos ser así considerados. Hay personas que por razones extrañas creen ser objetos con un alma, sea lo que sea que eso signifique, hecha por algún ser superior o algo similar. Pero incluso si este no fuese el caso, no es obvio que todo lo que sea computable es de hecho Turing computable. El trabajo de encontrar una solución a este problema, si es que la hubiese, excede, naturalmente, el presente.

En este capítulo se han descrito las nociones fundamentales del proyecto de I.A. lógica, los agentes sobre los que se plantea, los programas que posibilitan su comportamiento y las funciones que lo describen. Asimismo, se ha descrito la manera en que operan nociones como inteligencia y programación y que, dada la manera en que hemos descrito el proyecto, resultan ser fundamentales. Pretendo que, de esta manera, sea medianamente clara nuestra concepción de inteligencia, así como las acotaciones prudentes.En el siguiente capítulo expondré los problemas límites a los que se enfrenta el proyecto que me he esforzado por presentar claramente. Serán expuestos dos problemas, el problema del marco y el argumento del lenguaje privado, con la finalidad de determinar de manera adecuada los límites propuestos.

[42] Este problema propuesto por David Hilbert en la década de 1930 es sumamente interesante, también motivó las investigaciones de Kurt Gödel y el teorema de la incompletud (del cual se hace mención en las objeciones que Turing (1950) responde), pero extendería de manera innecesaria la explicación ofrecida en este apartado. De la misma forma, la operación y la relación de las máquinas de Turing y el cálculo Lambda de Alonzo Church.

3. Problemas límite de la I.A. lógica

A estas alturas del partido, es claro que la I.A. basada en lógica ejerce presión constante en el problema de la formalización. El *quid* de la cuestión parece descansar sobre la capacidad de traducir de manera adecuada, problemas postulados en un lenguaje natural a un lenguaje formal, pues hemos de decidir entre la complejidad de nuestra representación o la sencillez con la que podemos tratarlo. Si optamos por la primera, los problemas son tan o más difíciles de resolver que los que planteamos en términos naturales. No obstante, la complejidad creciente nos permite manipular de manera más simple conceptos altamente abstractos. Si por otro lado hemos apostado a la sencillez de nuestro lenguaje formal, deberemos enfrentarnos a la poca capacidad explicativa del mismo. De esta misma forma, hemos de resolver el problema del comportamiento racional de un agente artificial: o bien hacemos una descripción completa a costa de la resolución del problema o simplificamos el problema a costa de nuestro poder explicativo de los matices en la situación.

En lo siguiente analizaremos el enfrentamiento directo del agente que hemos descrito a lo largo del presente y acudiremos a dos problemas propios de la filosofía, como son el problema del marco y el argumento del lenguaje privado, que se desarrolla en ocasión del primero y pretende negar una posible solución del mismo. Por último, será analizado el problema P = NP de la teoría de la complejidad[43], el cual opera como una instancia parcial del problema del marco, pues si encontrásemos una solución satisfactoria, podríamos dar una respuesta concluyente (que no favorable necesariamente) a algunas de las consecuencias de la manera en que planteamos los otros dos problemas que le preceden.

[43] Descrito por primera vez por el reconocido científico de la computación Stephen Cook en Cook, Stephen (1971) *"The Complexity of Theorem Proving Procedures"* Proceedings of the third annual ACM symposium on Theory of computing, Toronto, Cánada. Págs. 151–158.

3.1 El problema del marco

Daniel Dennett describe en el artículo ***"Cognitive Wheels"*** (1984) una situación hipotética en que R1, un robot artificialmente inteligente, es programado para obtener una batería de una habitación con una bomba. R1 entra a la habitación y encuentra la batería en un vagón, entonces ejecuta la función E(v,f)[44] donde E(a,b) significa extraer a de b, v es la variable *vagón* y f es la variable *fuera de la habitación*: extraer el vagón fuera de la habitación. Pero en el vagón también va la bomba y R1 explota. Tiempo después crean a R1D1 (robot deductor), un robot capaz de deducir que la bomba saldrá con la batería si ejecuta E(v,f), justo cuando termina de evaluar el curso de acción en que la bomba explota cuando la extrae de la habitación, comienza a evaluar si cambiar el color de la bomba impediría que explote, cuando la bomba explota. El nuevo robot, R2D1 (robot deductor de relevancia) será entonces capaz, no sólo de inferir las consecuencias de los cursos de acción posibles, sino que además será capaz de separar los que resultan relevantes de los que no. R2D1 es llevado ante la habitación y se sienta a discriminar una gran cantidad de implicaciones no relevantes, hasta que se da cuenta de que esa tarea misma consiste en información no relevante, por lo que deduce que deberá incluirla en la lista y eventualmente ignorarla, tarea que jamás llevaría a cabo, por cierto, pues la bomba había estallado mucho antes de que se levantara. Dennett declara que, para que R2D2 tenga éxito donde sus predecesores fallaron, deberá resolver el seguimiento de las implicaciones de cada curso de acción posible, problema al cual denomina el *problema del marco* y que será revisado a continuación.

Cuando un agente naturalmente inteligente (entiéndase humano) establece un problema a resolver, parte de una situación de conceptos que operan como objetos y un conjunto de reglas que al menos de manera intuitiva reconoce. No obstante, es necesario advertir que un agente artificial no puede asumir este tipo de relaciones ya que tienen que ser explícitas. En el caso que antes describimos, quizá asumiendo una descripción parcial de las leyes físicas que de manera tan simple

[44] En el texto original la fórmula es ligeramente distinta, pero la aquí expuesta es más simple y opera de la misma manera, por lo que, en términos funcionales, son equivalentes.

percibimos, el agente fácilmente hubiese inferido que la bomba se movería con el vagón. El problema entonces se origina en la descripción de dichas leyes. Si aprendiese a reconocer e inferir este tipo de información contextual, entonces tendríamos que describir explícitamente las consecuencias de cada curso de acción elegido, pues, como veremos en lo siguiente, no hay manera de explicitar la relevancia de cada uno de estos hechos, sin descargar la autonomía sobre la agencia de la persona que haya programado a nuestro agente.[45]

Con esto no pretende decir que no hay otra forma de programar el robot para que resolviese su labor de manera satisfactoria, sino más bien que, para describir todas las reglas del entorno en que opera el agente, tendríamos que describir todas las decisiones posibles. Denett quiere explicar el problema del marco de manera intuitiva, por lo que la descripción que hemos expuesto podría considerarse simple y claramente no se presenta formalizada. La idea que subyace a este problema es que, al tener que ofrecer una descripción formal del mundo, ésta tendrá que ser completa o presentará estos fallos comunes.[46]

Es cierto también que no es Daniel Denett el primer autor en divisar este tema ni mucho menos. El problema del marco se encuentra publicado en "*Some Philosophical Problems from the Standpoint of Artificial Intelligence"* de McCarthy y Hayes (1969), donde también se inquieren sobre muchos de los problemas de I.A. basada en lógica aún no resueltos de manera satisfactoria. Empero, la manera en que es abordado el problema del marco, a pesar de que se puede dibujar a trazo ancho a través de todo el ensayo, privilegia un lenguaje formal que implica la comprensión de algunos otros procesos no esenciales, al mismo tiempo que parece restar atención a la relación entre el problema y el proyecto de I.A. y esto parece ser relevante en el presente.

[45] Con esto quisiera apuntar al *argumento del lenguaje privado* que será expuesto en breve. En dicho argumento se pretende dejar claro el motivo por el cual no es posible deducir reglas, que justifiquen la aplicación de reglas de un mero conjunto de *instrucciones*.

[46] Ciertamente no parece ser *a priori* imposible construir un modelo del mundo, pero esperaríamos que un agente inteligente pudiese desarrollar un comportamiento racional, sin la necesidad de conocer todas las aristas posibles del medio en que se desenvuelve -que parece ser la manera en que operan los humanos y otros animales. Por "fallos comunes" me refiero a los problemas en donde carece de una relación que podríamos considerar como sentido común, incluso.

Hemos dicho antes que un hipotético agente artificial inteligente estaría completamente determinado por un programa de naturaleza algorítmica, es decir, un conjunto de instrucciones secuenciales y mecánicas que resuelven problemas para cuya resolución fueron programados y en el que se esperaría que dicho agente sea capaz de actuar de una forma adecuada (sino de la mejor forma posible) de acuerdo con el problema al que se enfrenta. Es posible, no obstante, que muchas de las situaciones a las que se enfrente nuestro agente sean sumamente complejas, pues basta con intentar describir una situación sencilla en un entorno no controlado, para advertir lo rápido que aumentan las posibilidades creciendo de manera exponencial.

Podríamos argumentar en este punto que el problema del marco surge cuando pretendemos resolver todas las posibles conductas ante un evento indeterminado. Quizá sería conveniente incluir en la programación de nuestro agente la manera en que debería asignar relevancia a las consecuencias de los cursos de acción posibles. Con este propósito, deberemos cuestionar el carácter del programa y de qué manera o en qué nivel deberíamos añadir esta suerte de instrucciones al agente. Así, determinaremos qué características debería tener un programa que permita discriminar información no relevante y bordear el problema del marco.

3.2 El argumento del lenguaje privado (ALP)

Un programa es, en términos más o menos simples, un conjunto de reglas. Un programa computacional se compone de un conjunto de instrucciones secuenciales que se siguen de manera algorítmica. Ahora bien, cuando hablamos de una serie de instrucciones que determinan el nivel de relevancia de cierta información, hablamos de alguna suerte de meta-operación que determina la manera, fuera de dicho programa y el modo en que debería nuestro agente ejecutar las reglas que lo integran: seleccionando y ejecutando las que parecen más convenientes para fines concretos. Esto implica que si encontramos la manera de formalizar los hechos que determinan el adecuado seguimiento de una regla, podemos programar agentes que puedan evitar el problema del marco a través de la manera en que siguen reglas.

Esta posibilidad no es original ni mucho menos fue expuesta en las *Investigaciones filosóficas* (Wittgenstein, 2009) de manera fragmentada y en función de algunos de los temas más estudiados de Wittgenstein. Para el presente he tenido a bien seguir la exposición de Saúl Kripke en *A propósito de reglas y lenguaje privado* (2006), pues incluso, si tuviesen razón quienes sostienen que es una lectura agresiva con el autor y tendenciosa por lo general[47], sería un enfoque particularmente útil al leer bajo la luz del problema de *seguir una regla*. El argumento niega la posibilidad de un lenguaje privado, postura equivalente a negar la posibilidad del seguimiento privado de una regla.

La manera en que Kripke presenta al argumento es a través de la siguiente situación hipotética descrita al inicio de su obra (2006, pp. 22-24). Pretendo calcular el resultado de una suma que nunca antes había hecho, 68 + 57 y como era de esperarse, obtengo 125 por resultado. De pronto aparece un escéptico que cuestiona la certeza que parezco tener para decir que el resultado es 125. Sostiene que es posible que haya confundido mi uso pasado de la función de adición y haya aludido a una función distinta, la "cuadición" que es idéntica a la suma, pero que, al encontrarse con números mayores o iguales a 57, el resultado deberá ser invariablemente 5. Un problema, quizá, debido a un descuido, un error o una alucinación, pero un error, a fin de cuentas.

Ahora, Kripke mismo sostiene que nadie en el pleno uso de sus facultades mentales podría sostener la postura del escéptico seriamente, pero "si es falsa [la postura del escéptico], debe haber algún hecho acerca de mi uso pasado que pueda citarse para refutarla. Pues, aunque la hipótesis sea descabellada, no parece que sea *a priori* imposible" (1982, p. 24). No hace falta un análisis muy profundo para advertir que, al ser las reglas descritas únicamente de manera funcional (es decir, en virtud de la manera en que *funcionan*), la posibilidad de confundir mi uso pasado de la regla se debe a que la evidencia disponible sobre mi uso de adición y de *cuadición*[48] se resuelve de manera idéntica hasta el número 57. Nuevamente, el

[47] Kripke se deslinda de verter opiniones propias intencionales en la exposición del argumento en el prefacio "Si esta obra tiene una tesis principal propia es que el problema y el argumento escépticos de Wittgenstein son importantes, merecedores de investigación seria." (Kripke 2006, p. 13)

[48] Término imaginario con el objetivo de ilustrar el parecido entre las funciones. Su uso es reiterado más abajo en el presente.

argumento parece apuntar al hecho de que una descripción funcional necesita de un hecho que dé cuenta de mi uso de la regla y la figura del escéptico funciona como la personificación de una exigencia de prueba imposible de satisfacer.

Al igual que muchos de los problemas planteados en filosofía, el ALP está inmerso en un contexto y otros autores han trabajado sobre problemas similares, cuya explicación ahora nos resulta iluminadora. Éste es el caso de Willard V.O. Quine y Goodman[49], quienes exploran problemas similares en sus bases y divergentes en sus formulaciones. Quine describe el problema de *indeterminación de la traducción* por ejemplo, que delata la inescrutabilidad de la referencia de dos términos que podríamos hacer análogos en dos idiomas distintos. Esto es, sostiene que no existe un hecho en el mundo que sirva de garante para decir que un término y otro, en dos idiomas distintos, refieren al mismo objeto.

Pone por ejemplo el caso de un lingüista que pretende traducir el término 'conejo' a una nueva lengua de la cual no existen estudios previos, por lo que tendría que confiar en que una definición ostensiva (señalar al objeto, por ejemplo, y asociarlo con la expresión oral del término) debería ser suficiente, pero sin conseguir nunca la certeza de si eso refiere a 'conejo' o 'conejidad'. Este caso es similar debido a que aparece sólo ante cierta interacción social, una suerte de común acuerdo que se configura ante las necesidades que resuelve cuando se presenta. Difiere (y eso es importante) en tanto que el problema sobre el que ejecuta Quine no admite una evidencia introspectiva (es decir, de la cual no podemos tener idea, pues no presenta una conducta observable por un tercero), como en el caso de Wittgenstein, sino conductista (es decir, sólo admite la evidencia que podemos deducir del comportamiento observado). Es por esto que el papel de una *práctica* no adquiere el peso que en nuestro caso adquiere, pues Wittgenstein siempre apunta al peso que tiene una práctica para poder determinar el seguimiento de una regla, mientras que la respuesta de Quine descansa sobre una posible respuesta de corte disposicionalista (la cual respondería al reto escéptico

[49] Resulta útil para el presente recuperar *grosso modo* cómo es que se valen de un argumento escéptico, pero no pretendo comprometerme con una descripción completa del mismo, una explicación de su resolución o un desarrollo de los problemas a los que apuntan de manera particular.

asegurando que podemos determinar el uso de una regla por la disposición de quien la ejecuta).

En el caso de Goodman seguiremos un razonamiento escéptico más cercano al de Wittgenstein que el de Quine, pues postula[50] la existencia de un color denominado como 'verdul' que nace de la mezcla del azul y el verde. Goodman plantea una situación donde no es posible predecir el color que tendrá una esmeralda pues es posible preguntarse: ¿cómo es que sé que no he confundido mi uso pasado de verde con que calificó a las esmeraldas y en realidad he querido decir verdul? Las semejanzas entre ambos casos es evidente, pues en ambos la pregunta inquiere sobre el uso pasado de lo que parece ser una instrucción simple, pues el determinar el color de una esmeralda podría ser comprendido como la aplicación de una regla. Es claro que Goodman refiere al problema que implica la inducción y no el significado, pero la analogía es útil en tanto que remarca que el ALP puede ser aplicado en más ejemplos y que la decisión de exponer una regla matemática reconocible y descriptible en su totalidad, atiende más a razones relacionadas a la concepción que normalmente solemos tener de las normas matemáticas, como se ha remarcado antes.

A la luz de lo anteriormente dicho, teniendo en cuenta el objetivo del ejercicio mental del escéptico en el ejemplo de Kripke, tenemos que no se trata de una pregunta escéptica de naturaleza aritmética, pues de lo contrario podríamos responder con una simple prueba formal. El argumento apunta, más bien, a la carencia de un hecho capaz de diferenciar exitosamente entre una regla y otra, pues una regla descrita de forma operacional no parece ser distinguible de otra, hasta que existe un punto de divergencia. Al no haber un hecho de esta naturaleza, no es posible desarrollar un método formal de proveer de esta capacidad a nuestro hipotético programa. Conforme avancemos, parecerá cada vez más claro que no podemos resolver el ALP sin ceder algo de razón a la duda escéptica sobre nuestro uso pasado de la regla.

[50] Expuesto por el autor en: Goodman, Nelson, (1956). *Fact, Fiction and Forecast*, Harvard University Press, Massachusetts, EEUU.

Siguiendo este mismo hilo de pensamiento, parece un error el haber supuesto que éramos consistentes con nosotros mismos, pues, al seguir una regla asumimos que estamos siguiendo indicaciones asignadas desde un principio por nosotros mismos. De esta forma, nuestro primer movimiento será buscar en el pasado haber explicitado que la respuesta a este caso era 125. Sin embargo, al ser nuestra experiencia finita y pretender que jamás habíamos aplicado esta función a estos números en particular, esta opción queda descartada. De forma parecida podríamos sentirnos tentados a responder que deberíamos hacer lo que hasta entonces hemos hecho. No obstante, ¿no es esto seguir la regla que hemos aplicado? En otras palabras, si hubiésemos aplicado la regla de cuadición, estaríamos obligados a responder que la respuesta es 5 con la misma validez.[51] No podemos, por tanto, hacer lo que antes hemos hecho, ni tampoco recurrir a una operación más abstracta que la defina, pues de ser este el caso, podríamos elevar el nivel de abstracción del problema a través de la definición de cualquier otro término utilizado en un nivel más abstracto.

La única respuesta aceptable para el reto escéptico tendría que mostrar un hecho superlativo por el cual estoy justificado a decir que la respuesta *tiene* que ser 125 y no 5. De esa manera, indicaría cuál es mi estado mental[52] al querer decir *más* y no *cuás* ni ninguna otra posible interpretación. Una vez llegados a este punto podemos advertir que nuestra respuesta implica un injustificado salto de fe al vacío al momento de aplicar las reglas, pues si algo ha quedado claro es que no podemos justificar ningún hecho superlativo que justifique mi pretensión de utilizar la adición, en lugar de la *cuadición*, por ejemplo.[53] En lo que al hablante que hace la operación aritmética respecta, no puede negar lo que quiso decir en un primer momento, pero si estas instrucciones fuesen un proceso cognitivo que tiene que ser formalizado para un programa informático, ¿de qué manera podría describir satisfactoriamente su aplicación? Es imposible programar un agente artificial sin la existencia de un

[51] Y otra opción podría ser recurrir a alguna otra regla o concepto para saber cuál de las dos funciones adición o cuadición estamos usando en el pasado. Sin embargo, el mismo argumento escéptico podría aplicarse a las otras reglas o conceptos con las adecuaciones pertinentes. Insistir en elevar el nivel en que se aplica la regla nos lleva a una regresión al infinito.

[52] Utilizamos esta noción como es descrita antes en el presente.

[53] En este punto es pertinente insistir en que el ejemplo ofrecido por Wittgenstein está relacionado con operaciones aritméticas justo porque se cree que el conocimiento matemático es el conocimiento mejor fundamentado y más claro, y sí incluso este tipo de conocimiento puede ser atrapado por este argumento escéptico, entonces el uso de otros conceptos no puede estar en mejor posición.

hecho en el mundo, pues tendría que inscribirse al conjunto de instrucciones que configura al agente y no son de naturaleza individual, como apuntaremos en la resolución del argumento.

Kripke describe dos soluciones posibles al argumento escéptico: una solución directa y una escéptica. La primera taja de manera cortante la discusión, pues da cuenta de las razones por las cuales el escepticismo es injustificado. Responderemos de manera directa a un escéptico que duda de la periodicidad del amanecer hasta que mostremos una prueba infalible e incontrovertible de garantizar que así sea (al menos hasta que las condiciones cambien, por ejemplo, que el sol explote y no parezca haber ningún otro amanecer). Respecto a la solución escéptica escribe lo siguiente: "comienza, por el contrario, concediendo que las aserciones negativas del escéptico son irrebatibles. No obstante, nuestra práctica o creencia [...] no tiene porqué requerir la justificación que el escéptico ha mostrado insostenible." (Kripke 1982, pp. 79-80)

Lo anterior será mucho más claro en algunos momentos, explicaré las razones por las cuales dicha justificación no es necesaria. Nos será de utilidad explicar una de las diferencias más importantes entre Wittgenstein en el *Tractatus Logicus-Philosophicus*[54] y las *Investigaciones Filosóficas.* En la primera obra se explica el significado como la relación de una oración con los hechos en el mundo de los cuales depende su valor de verdad, al mismo tiempo que las oraciones atómicas son verdaderas por correspondencia. Un ejemplo de esto último podría ser una oración compuesta de oraciones atómicas cuyo valor de verdad depende de la composición del valor de verdad de las oraciones atómicas que la componen. En lo que a la segunda respecta, no plantea de ninguna forma una teoría del lenguaje alterna que pueda rivalizar con la de la primera. Si antes el *quid* de la cuestión descansaba sobre las condiciones de verdad, ahora descansa sobre las condiciones de justificación, es decir, sobre las condiciones en que podemos aseverar una expresión. Kripke (2006) acertadamente concreta la idea afirmando que Wittgenstein describe el lenguaje en virtud de las condiciones en que podemos

[54] Para esta exposición seguiré el trazo de las líneas generales de Wittgenstein que describe Kripke, pues no parece prudente poner la obra original sobre la mesa, su tratamiento nos desviará lo suficiente del punto de nuestra discusión.

aseverar apropiadamente una construcción de palabras. Es claro en este punto que la segunda noción apunta a una descripción de las circunstancias en las que se usan las palabras, será ahí dónde serán claras sus condiciones de justificación.

No se pueden dar las condiciones de verdad bajo las cuales sea verdadero que sigamos una regla. No hay hechos en el mundo que hagan verdaderas nuestras pretensiones de decir más en lugar de una regla cuasiforme como *cuás*, por ejemplo. Pero esto no es debido a la arbitrariedad con la cual podríamos responder a problemas aritméticos de adición, sino a que hasta ahora hemos analizado el caso de una persona en solitario, sin relación con otras. Incluso, cuando podemos quedar realmente extrañados ante el ALP, nadie suele poner en duda la manera en que usa la adición en su día a día, pues la regla que seguimos está inmersa en una comunidad al igual que nosotros.

Quisiera dejar claro que no hay una solución directa al problema del seguimiento de reglas en términos del ALP a través de la siguiente cita: “al final alcanzamos un nivel donde actuamos sin ninguna razón por cuya virtud podamos justificar nuestra acción. Actuamos sin dudar, pero a ciegas.” (Kripke, 2006, p. 99) ¿Por qué seguimos una regla de la forma en que lo hacemos? Porque estamos inmersos en una red correctiva constituida de las otras personas entre las que nos desenvolvemos, seguimos las reglas sin tener pruebas, pero tampoco dudas. Si bien tenemos que descartar una respuesta intencionalista, pues no es posible determinar las intenciones de un hablante sin presuponer su intención, sostenemos que el seguimiento de una regla implica información que no podemos deducir de la misma regla. Esto es un poco más claro cuando consideramos que no podría dar respuesta a un error aritmético, pues tendría que sostenerse que era esta la intención inicial.

Si analizamos el estado mental de una persona de manera independiente, encontraremos que no existe forma en que podamos encontrar alguna discrepancia entre sus intenciones pasadas y su uso presente. Es claro que, si hablásemos de un conjunto distinto de seres vivos, con otras redes, que funcionan de maneras distintas y configuran de extrañas formas la manera en que se comunican,

probablemente las reglas cambiarían en función de su conveniencia u operación. Así, en un mundo extraño en que unir dos cosas idénticas se volviesen una, las reglas de la adición serían más parecidas a las reglas de la *cuadición* y no podríamos encontrar en ninguna mente particular una justificación por la cual cuadicionar de esa forma y no de otra.

3.2.1 El ajedrez como posible instancia del ALP

En el campo de investigación de la I.A. es muy común asociar explicaciones y adjuntar ejemplos de juegos, pues son entornos didácticos y fácilmente programables. A continuación, expondré la relación del argumento del lenguaje privado y la programación de un agente artificialmente inteligente a través de un ejemplo de la aplicación de reglas, en ocasión de un juego que no suele alejarse mucho de los estudios relacionados a la I.A.: el ajedrez. Éste es un juego que opera como un entorno completamente controlado, del cual podemos conocer de manera muy precisa todos sus elementos[55] y opera de manera discreta. Dadas estas características, deberíamos ser capaces de jugar de manera óptima conociendo sus reglas y el objetivo (amenazar al rey sin que éste pueda moverse, es decir, amenazar con su captura de manera inminente). No obstante, no es muy complicado advertir que esta es una estrategia muy poco conveniente y que construir un agente inteligente capaz de resolverlo, está más lejos de lo que podríamos creer al conocer los intentos más aparentemente avanzados.[56]

[55] Se compone de 32 piezas: de las cuales 16 son blancas y 16 negras y se subdividen en 2 conjuntos de 8 piezas, donde ocho son equivalentes y las 8 restantes se dividen en 4 pares ordenados de manera jerárquica, de los cuales tres son de idéntico valor al de su par, pero en la última pareja una de las piezas es capaz de desplazarse de la manera más versátil, mientras en la última, de menor movimiento posible, descansa la victoria del equipo contrincante. Éstas muy bien determinadas piezas, se mueven desde una configuración inicial y simétrica, en 64 casillas posibles, donde 32 son negras y 32 son blancas, e inician las blancas un intercambio de turnos alternados.

[56] Antes de proceder al ejemplo, será necesario advertir que parecemos evadir el problema del marco al desarrollarnos en un entorno completamente controlado, sin embargo no hay que perder de vista que fácilmente podríamos regresar a él si pretendiésemos por ejemplo, que fuera capaz de desarrollar otras capacidades posiblemente útiles (cómo emplear las estrategias en operaciones más complejas que los seres racionales fácilmente podrían analogar al juego del ajedrez, o utilizar estrategias que impliquen la intimidación, al que también recurren algunos de ellos constantemente).

Muy probablemente, el procedimiento que seguiría un ser humano con conocimiento de las reglas y la operación del juego consistiría en trazar una línea entre el estado inicial del juego y aquel en que gana la partida. Una vez trazado dicho estado, movería sus piezas hasta conseguir avances parciales, evitaría que su dama se encuentre en riesgo y quizá evitaría doblar sus peones (ponerse uno frente al otro) porque intuitivamente no le parecería adecuado. El problema es que su agente contrincante no podría establecer dicha línea. No hay manera, por ejemplo, de que pueda deducir de las reglas la relación entre los estados inicial y deseado.

Incluso si encontrase una forma de desarrollar un curso de acción a través de las respuestas a las partidas propuestas por su contrincante, no podría encontrar relaciones que rápidamente podría encontrar y perfeccionar el agente natural humano. Si en este último, por ejemplo, pretende llegar al rey enemigo, difícilmente pondría en riesgo su pieza más poderosa, pues podría asumir que quién haya diseñado el juego probablemente lo hizo así por alguna razón. Muchas de las jugadas del agente natural podrían venir de relaciones que son producto de las reglas y de procesos creativos, errores de cálculo que probablemente una máquina no cometería o la propia audacia que podría ostentar el ser humano.

Las pautas generales con las cuales aplicamos las reglas que tenemos a nuestra disposición de manera general, en el lenguaje o en nuestra operación en el día a día, funcionan de la misma manera: elaboramos estrategias que relacionan conceptos que se encuentran dentro y fuera de las reglas del juego, intuimos y desarrollamos reglas propias que no podrían llevar a cabo agentes definidos en términos funcionales con el fin de obtener los mejores resultados posibles. Incluso si nuestro agente artificial tuviese una hipotética capacidad ilimitada, no podría sino calcular una cantidad indeterminada (posiblemente tendiente al infinito) de partidas completas posibles a raíz del estado inicial del tablero.

Claramente no es ésta la naturaleza de los procesos detrás de nuestras posibles jugadas. No podríamos desarrollar una cantidad tan complicada de información. Entonces, ¿de qué manera es que generamos estas relaciones creativas descritas? Bien, dicho proceso atiende a estos factores externos, pues

podríamos enunciar las reglas en las que estos factores se resuelven, pero no podríamos explicitar los procesos detrás de ellas. Lo anterior es debido a la ausencia de hechos (físicos o internos) que impliquen un correcto seguimiento de reglas. Es esta la operación del ALP, pues si hubiese alguna forma de formalizar este proceso, si fuese un hecho en el mundo, serían posibles los lenguajes privados, es decir, la aplicación privada de reglas. No obstante, hemos descartado ya esta posibilidad.

3.3 Complejidad: P y NP

Habiendo llegado a este punto, el lector o lectora podría haber notado que, tanto el problema del marco como el ALP, presuponen un sutil límite en el nivel de complejidad de los problemas que puede resolver un programa. Este problema es el problema de las clases de complejidad entre *P* = *NP*. La complejidad[57] es uno de los problemas que atraviesa el campo de la I.A. de manera completamente transversal. En este último apartado, con miras a la conclusión del presente, me gustaría explicitar las dimensiones del mismo, pues si bien no deseo dar una descripción absoluta de todos los problemas que configuran la escena de investigación, este se remarca de manera importante como necesario.

Recurriremos a las matemáticas, pues nos proveen de herramientas conceptuales más adecuadas que la filosofía, pero no expondré el problema con todo su detalle formal, pues no es mi campo y no es necesario para introducir los elementos relevantes en esta discusión. Dicho lo anterior, podríamos entender la complejidad de una función en términos de los pasos que serían necesarios para garantizar obtener una respuesta, el valor de salida de la función dados los valores de entrada para la misma. Si consideramos que una función puede escribirse como una iteración de funciones, cada una de estas interacciones agregaría una cantidad determinada de pasos, que tendrían que resolverse para dar solución a la función.

En el caso de las funciones clasificadas como polinomiales (P), el crecimiento, aunque posiblemente grande, es lo suficientemente mesurado como

[57] En adelante, cuando se hable de complejidad se hablará de complejidad computacional.

para que la cantidad de pasos que se tengan que realizar para garantizar obtener una respuesta a la aplicación de la función (obtener el valor de salida), no aumente de manera que nos sea prácticamente imposible realizar el proceso, con una cantidad relativamente pequeña de iteraciones de la función (esto dependerá de cuál sea la información de entrada a la función o procedimiento en términos de cantidad de información). Decir que el proceso sea prácticamente imposible de realizar es, en términos absolutos, simplemente que la cantidad de recursos que vamos a gastar en resolverlo crece demasiado rápido (tal que no pueda realizarse en una cantidad de tiempo polinomial)[58]. El costo es, por tanto, demasiado alto e incluso puede ser tan alto que requiera más recursos de los que existen en el universo mismo.[59]

Aquellas funciones que de hecho crecen de manera desmesurada con unas cuantas iteraciones se conocerán como funciones NP[60] y normalmente la cantidad de casos a considerar, aumenta a una velocidad exponencial.[61] La tesis P = NP sostiene que, a pesar de que una función bajo alguna presentación para ser considerada NP y, por tanto, que sus casos a considerar aumenten rápidamente, existe otra presentación de la misma función que es polinomial. Es decir, cuyos casos no crecen de manera tan acelerada, al final del día lo que se afirma más bien es que: existe una función que ante las mismas entradas arroje las mismas salidas, pero que requiera una menor capacidad de cómputo, una consideración de menos casos, ya sea en espacio, en tiempo, en memoria o en turnos. Es necesario advertir que en lo que a procesos efectivos refiere, es necesario que dichos procesos tienen que ser tractables, es decir, que puedan ser tratados con los recursos informáticos disponibles.

La relevancia de la tesis P = NP en este contexto es que, de ser verdadera, aminorara la complejidad que implica modelar algunos procesos relacionados con la

[58] En palabras del dr. Samuel Lomelí: el problema no consiste en lo largo que sea, sino en que lo rápido que crece.
[59] Una función polinomial otorga a cada valor de entrada un valor calculado con un polinomio. Es decir, una suma o resta de una cantidad finita de términos (monomios).
[60] No pretendo abarcar todas las categorías de complejidad inmiscuidas en el problema, pues describe una jerarquía de complejidad mucho más amplia de lo que una división entre P y NP sugiere.
[61] Crecer demasiado rápido en este contexto significa no poder ser descrito en una cantidad de tiempo polinomial.

I.A. En algún sentido, podríamos lograr ciertos resultados mediante la interacción de funciones, es decir, en el programa o paradigma de programación descrito al final de la sección 2.2.2. del presente. El problema descansa sobre la complejidad, en términos de espacio y tiempo, configura no sólo las posibles soluciones que podamos concebir, sino los términos en que podamos plantear los problemas. De esta manera, el problema P = NP establece una frontera precisa, en ocasión de la cual deberemos trazar el marco dentro del cual se desenvuelve la investigación de la I.A.

Resolver el problema en términos de los recursos suficientes para realizar las computaciones que parecen ser necesarias para computar un comportamiento generalmente inteligente, aunque interesante, no resuelve el problema de fondo. Esto es porque al final, de ser cierta la tesis P = NP (es decir, que los problemas que pertenecen al conjunto P sean idénticos a los problemas en el conjunto NP, clases de complejidad), simplemente sabremos que no necesitamos tantos recursos para computar funciones muy complejas. Esto no resuelve el problema de cómo es que podemos recuperar un mundo, que puede ser muy complejo, en términos de una descripción del mismo que únicamente recurre a la iteración de funciones matemáticas.

Conclusiones

No cabe duda de que no existe un momento en la historia donde podamos ubicar el inicio del proyecto de la I.A., así como no podemos determinar qué fines tiene o qué conceptos reconoce en los términos que lo configuran. A lo largo del presente trabajo de investigación han sido expuestas algunas de las posturas más relevantes en el panorama del proyecto de la I.A. y en ocasión de éstas se han descrito términos que resultan fundamentales para la descripción, no sólo de los objetivos del proyecto, sino también de los términos en que planteamos las preguntas a las que nos enfrentamos. Eso se debe a que, como en otros campos del conocimiento, el proyecto determina sus objetivos y los compromisos que estamos dispuestos a asumir, conforme nuevos campos, descubrimientos e intereses, se integran al cuerpo de creencias y métodos que entran en juego.

El inicio del presente trabajo tiene el objetivo de ubicar los orígenes de este proyecto, dichos orígenes fueron rastreados desde el proyecto de la cibernética de primer orden. Además de esto, fueron descritas las principales características de este proyecto y la relevancia de los modelos mecánicos de computabilidad, en particular el modelo de máquinas de Turing universales. En el caso particular de este trabajo se buscaba evaluar el proyecto de la inteligencia artificial basada en lógica, específicamente si es que este proyecto podría modelar, de manera adecuada, un comportamiento racional en agentes programables. Con este fin fueron descritas las nociones de “inteligencia” y “programación”, lo cual sirvió como punto de inicio para evaluar los límites que nos compelen. Dichos límites establecían características imposibles para nuestros agentes programables, pero también configuraron las respuestas a dichos problemas.

Para poder evaluar de manera adecuada sí estos modelos podrían recuperar nuestra noción de agente inteligente en el segundo capítulo. En un primer momento definimos de manera lo más precisa posible lo que consideraríamos como agente y, en particular, un agente artificial. En un segundo momento se planteó el concepto de inteligencia general, en el cual se destacaban la necesidad de no solamente dominar ciertas herramientas o técnicas, sino la habilidad de adaptar dichas

herramientas a nuevos contextos, en muchos casos, sin relación con contextos previos. La idea a recuperar es que la inteligencia general no debería simplemente atribuirse a una agente que tienen una base de datos enorme, sino a uno que no requiera de un plan acabado para enfrentar de manera adecuada cada situación posible. La adaptabilidad se convirtió en uno de los rasgos principales de la inteligencia general. Finalmente, en dicho capítulo se presentó el tipo de programas que se utilizan o podrían utilizarse dentro de los límites del proyecto descrito para modelar la noción de inteligencia general. Los límites están establecidos por la noción de computabilidad en una máquina de Turing. La conclusión del segundo capítulo termina por enfatizar la naturaleza de los límites de la programación y la inteligencia general.

En el último capítulo se evaluó si los límites de la computabilidad Turing eran lo suficientemente amplios como para recuperar una noción de inteligencia general. Si bien no se ofrecieron argumentos contundentes en contra de dicha tesis, se plantearon algunos problemas a los cuales se debería enfrentar el proyecto. El primero de ellos era el argumento de seguir una regla (argumento de lenguaje privado) ofrecido por Wittgenstein y descrito por Kripke. En dicho argumento, de alguna manera se sostenía que deberíamos poder recuperar lo suficiente del contexto para poder determinar si es que nosotros estamos siguiendo una regla o por lo menos no de manera adecuada. Esto está estrechamente relacionado con nuestra capacidad de adaptarnos a diferentes contextos y es de la mayor relevancia considerando que los programas pueden ser vistos justo como una serie de reglas bien estructuradas. Lo más relevante de este argumento es su solución. La solución escéptica considerada en el ALP, es útil para poder determinar cuándo es que nosotros seguimos correctamente una regla, usamos bien un concepto o cualquier otra formulación similar, es necesario apelar a un contexto dado (no solamente por el mundo y sus propiedades, sino también por una comunidad, que nos ayuda a fortalecer nuestro sistema conceptual o incluso crearlo). El principal problema con esta solución es que parece muy poco plausible que la recuperación del contexto se pueda describir por completo y de manera adecuada por el paradigma de computación que tiene de fondo la I.A. lógica, pues se pretende ofrecer descripciones puramente funcionales.

El segundo problema tratado en este capítulo fue el problema del marco. Dicho problema sostiene que la búsqueda o selección de las variables relevantes para ser procesadas por una máquina, dado un contexto, es un problema de una complejidad considerable: un problema de complejidad NP. Esto, si bien no imposibilita que pueda ser resuelto, sí indica que la manera en la cual una computadora, asociada con el programa Inteligencia artificial basado en lógica, procesa el contexto y requiere una cantidad enorme de recursos para poder determinar cuáles son las variables relevantes del mismo. Lo anterior sería el caso, claro, a menos que la tesis P = NP sea el caso. Algo que contrasta mucho con la cantidad de recursos que requiere un ser humano para procesar contextos similares.

No creo haber podido demostrar de manera contundente que el proyecto de I.A. basado en lógica debería ser descartado. No obstante, creo haber apuntado a que la intencionalidad y algunos otros factores parecen indicar que, aunque se pueden obtener buenos resultados, será complicado (sino es que imposible) obtener un modelo satisfactorio de comportamiento inteligente desde este paradigma de la I.A. Aunque reconozco que, para obtener un resultado más sólido, es necesario considerar algunos otros factores asociados con la interacción de la máquina con el medio, por ejemplo, lo que espero poder hacer en trabajos posteriores.

Obras consultadas

Dennett, C. D. (1984), *Cognitive Wheels: The Frame Problem in Artificial Intelligence*. En I. Hookway Christopher (ed.), *Minds, Machines and Evolution* (pp. 129–150). Cambridge, Inglaterra: Cambridge University Press.

Dupuy, J. P. (2009). *The Mechanization of the Mind: on the Origins of Cognitive Science*. Cambridge, Estados Unidos de América: MIT Press.

Jefferson, G. (1949). The Mind of Mechanical Man. *The British Medical Journal*, *1*, (4616), 1105-1110.

Kripke, S. (2006), *Wittgenstein, a propósito de reglas y lenguaje privado*. Madrid, España: Tecnos.

McCarthy, J. Minsky, M. L. Rochester y N. Shannon, C. E. (1955). *A Proposal for the Dartmouth Research Project on Artificial Intelligence*. Hanover, Estados Unidos de América: sin publicar.

McCarthy, J. & Hayes, P.J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. En Webber, Bonnie L. y Nilson J., Nils. (ed.) *Readings in Artificial Intelligence* (pp. 431-450) Los Altos, Estados Unidos de América: Morgan Kaufmann Publishers, Inc.

McCorduck, P. (2004). *Machines Who Think*. *A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. Natick, Estados Unidos de América: AK Peters Ltd.

Pias, C. (ed.). (2016). *Cybernetics. The Macy Conferences 1946-1953. The Complete Translations*. Chicago, Estados Unidos de América: Chicago University Press.

Rescorla, M., (2020) "The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), recuperado de: https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>.

Russell S. y Norvig P.r. (2004). *Inteligencia Artificial. Un Enfoque Moderno.* Madrid, España: Pearsons Education.

Simon, H., Newell, A. y Shawn, J. (1959). *Report on a General Problem-Solving Problem*. Pennsylvania, Estados Unidos de América: Carnegie Institute of Technology.

Shanahan, M. (2016). *The Frame Problem*, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Recuperado de: https://plato.stanford.edu/archives/spr2016/entries/frame-problem/.

Turing, A. (1950). Computing Machinery and Intelligence. En *MIND*, *LIX* (236) 433-460.

Wittgenstein, L. (2009), Tractatus Logicus-Philosophicus. En *Wittgenstein I* (Veiga, Jacobo Muñoz. y Suárez García, Alfonso. trad.). Madrid, España: Gredos.

Literatura complementaria.

Bloor, D. (1997). Meaning Finitism. En *Wittgenstein, Rules and Institutions* (pp. 9-26). Londres, Inglaterra: Routledge.

Cherniak, C. (1990). *Minimal Rationality*. Cambridge, Estados Unidos de América: MIT Press.

David, M. (2000). *The Universal Computer. The Road from Leibniz to Turing*, New York, Estados Unidos de América: W. W. Norton & Company.

Pollock, L. J. (1995). *Cognitive Carpentry. A Blueprint for How to Build a Person*. Cambridge, Estados Unidos de América: MIT Press.

McFarland, D. (2008). Guilty Robots, *Happy Dogs. The Question Alien Minds.* Oxford, Inglaterra: Oxford University Press.

Kusch, M. (2006). Kripke 's interpretation of Wittgenstein. En *A Sceptical Guide to Meaning and Rules* (pp. 237-264). Acumen Publishing. doi:10.1017/UPO9781844653782.009