INTENCIONALIDAD EN SISTEMAS ARTIFICIALES:
UNA APROXIMACIÓN DESDE JOHN SEARLE

FABIÁN BAUTISTA GONZÁLEZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE CIENCIAS HUMANAS

ESCUELA DE FILOSOFÍA

BUCARAMANGA

2017

INTENCIONALIDAD EN SISTEMAS ARTIFICIALES: UNA APROXIMACIÓN DESDE JOHN SEARLE

FABIÁN BAUTISTA GONZÁLEZ

TRABAJO DE GRADO PARA OPTAR AL TÍTULO DE FILÓSOFO

DIRECTOR:
JORGE FRANCISCO MALDONADO SERRANO
DOCTOR EN FILOSOFÍA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS HUMANAS
ESCUELA DE FILOSOFÍA
BUCARAMANGA
2017

De la serie de hechos inexplicables que son el universo o el tiempo, la dedicatoria de un [texto] no es, por cierto, el menos arcano. Se la define como un don, un regalo. Salvo en el caso de la indiferente moneda que la caridad cristiana deja caer en la palma del pobre, todo regalo verdadero es recíproco. El que da no se priva de lo que da. Dar y recibir son lo mismo. [...]

JORGE LUIS BORGES

A Paula

CONTENIDO

RESUMEN

TÍTULO: INTENCIONALIDAD EN SISTEMAS ARTIFICIALES: UNA APROXIMACIÓN DESDE JOHN SEARLE[*]

AUTOR: FABIÁN BAUTISTA GONZÁLEZ[**]

PALABRAS CLAVE: Intencionalidad, Sistemas artificiales, Inteligencia Artificial, GOFAI, clases de intencionalidad

La monografía se ocupa de la noción –estrictamente filosófica– de la intencionalidad. En concreto, el texto ensaya la formulación del problema que consiste en preguntarse acerca de la forma correcta de atribuir o adscribir estados intencionales a diferentes tipos de sistemas. Con base en la teoría de la intencionalidad del filósofo norteamericano John Searle, el objetivo es determinar la clase intencional que le es posible atribuir a un sistema artificial del tipo de la Inteligencia Artificial Fuerte y de tipo GOFAI (*Good Old-Fashioned Artificial Intelligence*). Se trata de comprender el sentido en el que, para Searle, la intencionalidad de estos tipos de sistemas no es intrínseca a ellos mismos sino que, más bien, es intencionalidad derivada. Lo anterior implica estudiar tanto la forma en la que Searle entiende la intencionalidad como las clases intencionales que pueden ser fijadas a partir de allí: intencionalidad intrínseca, intencionalidad derivada e intencionalidad como-si. Entre otras cosas, la clasificación de la intencionalidad de Searle compromete la distinción ontológica entre aquello que es relativo a nosotros, los seres humanos, y aquello que es independiente. En este contexto general, resulta bastante justo pensar que la monografía está enfocada –no por ello limitada– en presentar la posición de Searle respecto de uno de los programas investigativos de la Inteligencia Artificial, a saber, la Inteligencia Artificial Fuerte o GOFAI.

[*] Monografía

[**] Facultad de Ciencias humanas, Escuela de Filosofía, Director Jorge Francisco Maldonado Serrano

SUMMARY

TITLE: INTENTIONALITY IN ARTIFICIAL SYSTEMS: AN APPROACH FROM JOHN SEARLE*

AUTHOR: FABIÁN BAUTISTA GONZÁLEZ**

KEY WORDS: Intentionality, Artificial systems, Artificial Intelligence, GOFAI, types of intentionality

The monograph deals with the notion –strictly philosophical– of intentionality. Specifically, the text tests the formulation of the problem of asking the right way to assign or ascribe intentional states to different types of systems. Based on the intentionality theory of the American philosopher John Searle, the objective is to determine the intentional class that can be attributed to an artificial system of the type of Strong Artificial Intelligence and GOFAI (Good Old-Fashioned Artificial Intelligence). It is about to understand the sense in which, for Searle, the intentionality of these types of systems is not intrinsic to themselves but, rather, it is derived intentionality. This implies studying both the way in which Searle understands intentionality and the intentional classes that can be fixed from there: intrinsic intentionality, derived intentionality and as-if intentionality. Among other things, Searle's classification of intentionality compromises the ontological distinction between that which is relative to us, human beings, and that which is independent. In such a context, it is quite fair to think that the monograph is focused –but not limited– on presenting Searle's position regarding one of the research programs of Artificial Intelligence, namely Strong Artificial Intelligence or GOFAI.

* Monograph

** Faculty of Human sciences , School of Philosophy, Director Jorge Francisco Maldonado Serrano

INTRODUCCIÓN

La *intencionalidad*, me parece, es una noción rigurosamente filosófica. Con esto sólo quiero decir que la palabra 'intencionalidad', original de los escolásticos medievales y rescatada por el austriaco Franz Brentano, tiene resonancia tanto en la historia de la filosofía como en la filosofía actual. Por un lado, movimientos filosóficos de renombre han surgido en torno a la noción de intencionalidad (piénsese en la fenomenología) y han desarrollado lo que se puede llamar, sin ningún tipo de reparo, 'teorías de la intencionalidad'. Por otro lado, el giro mentalista que dio la tendencia analítica de la filosofía en los años setentas y ochentas del siglo pasado llevó a muchos filósofos anglosajones a ocuparse del fenómeno intencional. Cuando el lugar que en su momento ocupó el lenguaje, a causa del famoso giro lingüístico, fue ocupado por la mente, por lo mental –en el más amplio sentido de la expresión–, la intencionalidad despertó un interés notable, pues se consideró como una propiedad básica de los así llamados estados o fenómenos mentales: la propiedad que consiste en dirigirse a, o ser sobre o de, algo en el mundo[1].

Que la intencionalidad tenga un destacado interés filosófico quiere decir, cuando menos, que existe un conjunto de problemas teóricos o conceptuales que giran en torno suyo. De una parte, se pueden identificar problemas de tipo metafísico, como el problema que reside en aclarar la naturaleza de la intencionalidad o el problema del estatus ontológico del objeto intencional. De otra parte, es posible identificar problemas de tipo epistemológico, como el problema que consiste en precisar cómo se puede explicar de modo objetivo la intencionalidad[2]. Y finalmente, se pueden identificar también problemas que tentativamente llamaré 'de tipo normativo', como el problema de determinar bajo qué criterios es posible la atribución o la adscripción de estados intencionales. En líneas generales, es válido decir que

[1] Cfr. BRENTANO, Franz. Psychology from an Empirical Standpoint. London: Routledge. 2009, p.68.

[2] Este problema remite al desafío contemporáneo de brindar una explicación de la intencionalidad, de las propiedades de los estados mentales en general, que sea coherente con nuestra vision natural del universo, es decir, que se vincule con las demás explicaciones del mundo natural dadas por las ciencias naturales. De modo general se le ha denominado a este desafío 'la naturalización de la mente'.

esta monografía se ocupa del último problema, es decir, del problema acerca de la correcta adscripción o atribución de intencionalidad.

En concreto, 'el problema de la atribución de intencionalidad', como aquí lo llamo, hace referencia al hecho de que cotidianamente nos vemos usando oraciones como 'el perro *espera* mi llegada', 'mi teléfono *entiende* cuando le hablo' o 'la planta *está triste*'. Al emitir una oración de este tipo estamos atribuyendo –tal vez sin quererlo– estados mentales intencionales a sistemas que podrían no tenerlos. En ese sentido, el problema de la atribución de intencionalidad es el problema que consiste en determinar en qué casos tiene sentido atribuir o adscribir intencionalidad, esto es, estados intencionales (esperar, estar triste, entender, etc.), y bajo qué criterios es posible tal atribución. Nótese que en el primer caso estaríamos dispuestos a sostener que el perro que espera mi llegada *realmente espera* mi llegada. Pero, a diferencia del primero, el segundo caso es confuso (¿hasta qué punto aceptaríamos que un dispositivo digital entiende?) y el tercero es incluso divertido (¿realmente queremos decir que una planta puede sentir tristeza?).

Mi objetivo principal –al que esta monografía se restringe– es determinar si tiene sentido atribuir la propiedad de la intencionalidad a sistemas artificiales del tipo de la Inteligencia Arficial Fuerte[3] o de tipo GOFAI[4] (*Good Old-Fashioned Artificial Intelligence*), con base especialmente en la teoría de la intencionalidad del filósofo norteamericano John Searle. Asimismo, en caso de que tenga sentido, me interesa especificar qué clase de intencionalidad es la que podemos adscribir a este tipo de sistemas. Mi objetivo supone, pues, estudiar tanto la manera en la que Searle entiende la intencionalidad como las clases de intencionalidad que pueden ser fijadas a partir de allí. En ese contexto, es bastante justo pensar que mi monografía está enfocada –no por ello limitada– en presentar la posición de Searle respecto del proyecto de la Inteligencia Artificial. Esto se apreciará en el plan de trabajo, que pasaré a indicar enseguida.

---

[3] Cfr. SEARLE, John. Minds, Brains, and Programs. En: Behavioral and Brains Science, núm. 3 (1980). pp. 417-457. p. 417.
[4] Cfr. HAUGELAND, John. The Artificial Intelligence: The Very Idea. Cambridge: The MIT Press. 1989.

En el primer capítulo, el cual tiene un carácter preliminar, presentaré una breve revisión histórica del modo en que la tendencia analítica de la filosofía heredó algunos aspectos de los problemas relacionados con la intencionalidad. Esto me permitirá, entre otras cosas, justificar la formulación que hago del problema ('el problema de la atribución de intencionalidad') y entender parcialmente las razones por las que Searle propone una teoría que responde a tal problema. En la segunda parte del capítulo, a su vez, aclararé el término 'sistema artificial' vinculándolo con la Inteligencia Artificial Fuerte, que es una noción acuñada por Searle, y con los sistemas de tipo GOFAI, expuestos en su momento por John Haugeland. En ese sentido, con relación al segundo capítulo, el primero puede –y tal vez debe– ser considerado un preámbulo.

En el segundo capítulo, por su parte, expondré los elementos fundamentales de la teoría de la intencionalidad de John Searle. En un primer momento, con base en el modelo de los actos de habla, desarrollaré la idea según la cual la intencionalidad debe ser entendida como representación y brindaré una definición de las clases intencionales propuestas por Searle. En un segundo momento, ilustraré la posición de Searle respecto de los sistemas artificiales a partir de su famoso experimento mental de la Habitación China. El objetivo del capítulo es determinar, a partir de la clasificación de la intencionalidad de Searle, la clase intencional que le es posible atribuir a un sistema artificial de tipo de la Inteligencia Artificial Fuerte o de tipo GOFAI.

Finalmente, en las conclusiones, además de hacer un breve recuento de los resultados obtenidos, trazaré posibles caminos de investigación en filosofía y en disciplinas adyacentes.

CAPÍTULO I: INTENCIONALIDAD Y SISTEMAS ARTIFICIALES

El propósito de este primer capítulo es suministrar una comprensión del modo en el que la filosofía de tendencia analítica heredó algunos aspectos de los problemas que giran en torno a la intencionalidad. Además, en la segunda parte del capítulo preciso el término 'sistema artificial' y lo relaciono con la Inteligencia Artificial Fuerte y con GOFAI. Esto servirá para motivar el estudio de la pregunta que me interesa y que desarrollaré en el segundo capítulo del texto. En tal sentido, en esta ocasión no es mi objetivo tomar partido respecto de las discusiones que aquí presento –valga decir– sólo a manera de esbozo.

I.I ALGUNOS ASPECTOS DE LA INTENCIONALIDAD

Intencionalidad, del latín *intentio*, es la palabra que los filósofos de la Edad Media usaban para designar la 'dirección-hacia' algo. Como advirtió Brentano sobre el caso de Santo Tomás[5], su uso de la palabra se vio restringido a fines teológicos, y es probable que lo mismo pueda decirse también de San Agustín y de San Anselmo. Lo cierto es que en los tres casos la palabra 'intencionalidad' guardaba una estrecha relación con la teología y parece difícil interpretar un uso diferente a aquél. Sin embargo, para Brentano, que estaba interesado en brindar una descripción de lo mental, la palabra podía ser usada para señalar una característica general de los fenómenos mentales. En sus palabras:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is afirmed or denied, in love loved, in hate hated, in desire desired and so on. [...] This intentional in-existence is characteristic

[5] BRENTANO. Óp. cit., p. 67.

> exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves[6].

En ese famoso pasaje de *Psicología desde un punto de vista empírico*, de 1874, Brentano estaba en búsqueda de una definición unitaria para la totalidad de lo mental. Su estrategia era determinar las distinciones que existen entre los fenómenos mentales y los fenómenos físicos. Después de haber considerado a los fenómenos mentales como representaciones o como descansando en representaciones[7], y de haber señalado que los fenómenos mentales, a diferencia de los fenómenos físicos, carecen de extensión[8], Brentano estableció un sentido de 'intencional' en términos de 'dirección hacia un objeto', de 'referencia a un contenido' y de 'objetividad inmanente'. Con ello supo indicar que es un rasgo de los fenómenos mentales el hecho de que todos ellos tienen un contenido al cual refieren, un objeto al cual se dirigen. Una creencia, un deseo, una esperanza, una experiencia perceptual, todos estos ejemplos de fenómenos mentales siempre se dirigen a algo, siempre refieren a un contenido: creencia de que está lloviendo, deseo de ganar una competencia, esperanza de que las cosas mejoren, etc. Por lo que respecta a Brentano, no parece haber un fenómeno mental que no refiera a algo: una creencia de nada, un deseo de nada no serían ni creencia ni deseo, no serían fenómenos mentales.

---

[6] "Todo fenómeno mental está caracterizado por lo que los escolásticos de la Edad Media han llamado la inexistencia intencional (o mental) de un objeto, y que nosotros llamaríamos, aunque no sin ninguna ambigüedad, la referencia a un contenido, la dirección hacia un objeto (el cual no hay que entender aquí en el sentido de una cosa), o la objetividad inmanente. Todo fenómeno mental contiene en sí mismo algo como su objeto, si bien no todos lo hacen del mismo modo. En la representación hay algo representado; en el juicio hay algo afirmado o negado; en el amor, amado; en el odio, odiado; en el deseo, deseado, etc. [...] Esta in-existencia intencional es una característica exclusiva del fenómeno mental. Ningún fenómeno físico exhibe nada igual. Por tanto, podemos definir al fenómeno mental diciendo que es la clase de fenómeno que contiene un objeto intencional dentro de sí". BRENTANO. Óp. cit., p. 68. En cada caso se mencionará si la traducción es propia. En este caso, lo es.

[7] Ibíd., p. 61.

[8] Ibíd., p. 65.

1.1.1 'Inexistencia' del objeto intencional: A fin de precisar qué significa 'referencia a un contenido', 'dirección hacia un objeto', etc., Brentano afirmó al comienzo de la cita que estas expresiones deben entenderse en el sentido de los escolásticos como 'inexistencia intencional de un objeto'. Pero, lejos de ser una indicación clara para entender la intencionalidad, la formulación se torna irremediablemente confusa, sobre todo en lo que refiere a la palabra 'inexistencia'[9]. Por ello, surgieron debates alrededor de lo que Brentano quiso decir con la palabra: si acaso buscaba expresar el hecho de que los fenómenos mentales pueden dirigirse a objetos *inexistentes*, como cuando alguien desea un unicornio o cree que Papá Noel vendrá, o bien manifestar que los objetos a los que estos fenómenos se dirigen son internos, es decir, que *existen-en* la mente en el sentido de que se encuentran en ella.

Sea como fuere, lo importante ahora es notar que ambas opciones aluden a un aspecto de la intencionalidad que fue y es primordial para muchos, a saber, la cuestión acerca de la naturaleza de los objetos a los que se dirigen los fenómenos mentales, los así llamados 'objetos intencionales'. Tanto en el caso de los 'objetos inexistentes' como en el caso de los 'objetos existentes en la mente' se originan preguntas que remiten a la *naturaleza* de los objetos intencionales: ¿en qué sentido es inexistente el objeto intencional? ¿qué significa que el objeto intencional exista en la mente? ¿qué es, pues, exactamente un objeto intencional? Es reconocible el interés por este tipo de preguntas en la filosofía posterior a Brentano. En concreto, puede decirse que el tema planteado –si bien no desarrollado– por él llevó a Meinong[10], Searle[11] y Crane[12], entre otros, a estudiar la naturaleza de los objetos intencionales.

---

[9] La ambigüedad de la palabra 'inexistencia' resalta también en castellano. Al menos así lo parece pues, según la Real Academia de nuestra lengua, 'inexistencia' posee –lo mismo que en el latín de los escolásticos y en el alemán de Brentano– dos acepciones: 'falta de existencia' y 'existencia de algo en otra cosa'.

[10] Cfr. MEINONG, Alexius. The theory of objects. En: CHISHOLM, Roderick (Ed.). Realism and the Background of Phenomenology. Glencoe: The Free Press, 1960.

[11] Cfr. SEARLE, John. Intentionality. Cambridge: Cambridge University Press. 1983.

[12] Cfr. CRANE, Tim. The Objects of Thought. Oxford: Oxford University Press. 2013

Y como en el caso de Mally[13] y de Parsons[14], este debate promovió también la reflexión sobre los objetos inexistentes o ficticios.

1.1.2 Intencionalidad e intensionalidad: En la cita de Brentano se lee, pues, que un rasgo característico de los fenómenos mentales es que exhiben intencionalidad, la capacidad que consiste en ser sobre, referirse o dirigirse a algo. Esto se hace particularmente visible cuando decimos cosas como: 'creo que está lloviendo', 'deseo que ganes la competencia', 'espero que las cosas mejoren', es decir, cuando usamos enunciados que expresan lo que, en 1918, Bertrand Russell llamó 'actitudes proposicionales'[15]. Estos enunciados, que son reportes de fenómenos mentales, tienen *propiedades lógicas* propias, las cuales los definen. Parece que, en un primer momento, la filosofía analítica del pasado siglo concentró su estudio de la intencionalidad en el análisis de estas propiedades, en el estudio lógico de las actitudes proposicionales descritas por Russell[16]. Ilustraré esto con un caso concreto.

Casi en los sesentas, quizá impulsado por el famoso giro lingüístico, Roderick Chisholm[17] vio una oportunidad de entender los fenómenos mentales intencionales, su intencionalidad, estudiando los enunciados con los que se reportan tales fenómenos. Su objetivo fue analizar

---

[13] Cfr. MALLY, Ernst. Gegenstandtheoretische Grundlagen der Logik und Logistik. Leipzig:Barth. 1912.

[14] Cfr. PARSONS, Terence. Nonexistent Objects. New Haven: Yale University Press. 1980.

[15] El término 'actitudes proposicionales' remite a los enunciados con los que se reportan los fenómenos mentales. Este tipo de fenómenos –como la creencia– son actitudes proposicionales, pues son actitudes que un agente tiene respecto de una proposición. Cuando digo 'creo que está lloviendo' la actitud que tengo hacia la proposición 'está lloviendo' es la creencia. Lo mismo parece aplicar para otros casos de reportes de fenómenos mentales simples como: 'deseo que ganes la competencia', 'espero que las cosas mejoren', 'conozco la obra de Russell', 'comprendo el idioma alemán', etc. Cfr. RUSSELL, Bertrand. The philosophy of logical atomism. En: The Monist. (1918) pp. 177-281. p. 227.

[16] Es comprensible que el enfoque para estudiar la intencionalidad en ese momento fuera el análisis del lenguaje con el que hablamos de los fenómenos mentales pues, para la tendencia analítica de entonces, la filosofía era principalmente el estudio del lenguaje y el análisis lógico.

[17] Cfr. CHISHOLM, Roderick. Sentences about Believing. En: Minnesota Studies in the Philosophy of Science, núm. 1 (1957). pp. 510-520.

las propiedades lógicas de las actitudes proposicionales descritas por Russell bajo la sospecha de que existe una estrecha relación entre la *intencionalidad*, la característica de los fenómenos mentales, y la *intensionalidad*, la propiedad lógica de cierto tipo de enunciados. La similitud ortográfica entre ambas palabras, 'intencionalidad-con-c' e 'intensionalidad-con-s' (en inglés: 'intentionali-ty' e 'intensionality'), le hicieron pensar que los fenómenos mentales intencionales podían ser, también, intensionales, en el sentido de no-extensionales. Y de ese modo podría interpretarse la intencionalidad de los fenómenos mentales como intensionalidad, como no-extensionalidad.

Al estudiar los enunciados que usamos al hablar de los fenómenos mentales, tal como hizo Chisholm, se observa que se comportan de una manera lógicamente atípica. En concreto, no superan el test de generalización existencial ni el test de sustituibilidad, que son las dos pruebas básicas que permiten atribuir la propiedad de extensionalidad a los enunciados de un lenguaje[18]. Para ver esto, réparese primero en los siguientes enunciados que se comportan típicamente:

(1) Russell es filósofo
(2) El autor de *Principia Mathematica* es filósofo

Los enunciados (1) y (2), por un lado, permiten generalización existencial: con base en ellos podemos hacer enunciados generales como 'Hay un x tal que x es filósofo' (representado como '∃x(x filósofo)' en lógica de primer orden). Por otro lado, (1) y (2) permiten sustituir los términos que los componen sin que se afecte su carácter de verdad: esto se prueba porque no puede ser el caso que (1) sea verdadero y a la vez (2) sea falso (no puede ser el caso que Russell sea filósofo y a la vez el autor de los *Principia*, que es Russell, no sea filósofo). Dado que los enunciados (1) y (2) pasan estas dos pruebas podemos afirmar que son extensionales.

[18] Cfr. SEARLE, Óp. cit., p. 188-200.

Ahora bien, repárese en los siguientes enunciados que son reportes de fenómenos mentales:

(3) Paula cree que Russell es filósofo

(4) Paula cree que el autor de *Principia Mathematica* es filósofo

Los enunciados (3) y (4), por un lado, no permiten generalización existencial: del hecho de que Paula crea que Russell o el autor de *Principia* es filósofo no puede enunciarse que 'Hay un x tal que x Paula cree que es filósofo'. Por otro lado, la verdad del enunciado (3) no implica la verdad del enunciado (4), pues podría ser el caso que Paula creyera que el autor de los *Principia* no es Russell, sino –por ejemplo– Einstein. Y así, aunque (3) fuese verdad, no se garantiza la verdad de (4) al sustituir los términos. Dado que (3) y (4) no pasan las pruebas de sustitubilidad ni de generalización existencial no podemos afirmar que son extensionales. Y si esto en general aplica para todos los enunciados que reportan fenómenos mentales, como de hecho parece ser el caso, entonces estos enunciados son intensionales, exhiben la propiedad de la intensionalidad.

Que no podamos atribuir la propiedad lógica de la extensionalidad a los enunciados con los que se reportan los fenómenos mentales llevó a una discusión muy amplia en filosofía y en especial en la filosofía de tendencia analítica. En concreto, puede decirse que las reflexiones de Cornman[19], Kneale[20] y Urmson[21] comparten el espíritu del trabajo de Chisholm. Además, su reflexión sobre la intencionalidad de Brentano llevó a Searle[22] y Zalta[23], entre otros, a discutir la posible relación entre la intencionalidad-con-c y la intensionalidad-con-s. En resumen, si bien muy objetado, el trabajo de Chisholm fue determinante para que la filosofía analítica, en cuyo horizonte se veía ya el giro mentalista, se interesara por el tema de la intencionalidad.

---

[19] Cfr. CORNMAN, James. Intentionality and Intensionality. En: Philosophical Quarterly, XII (1962). pp. 44-52.
[20] Cfr. KNALE,William. Intentionality and Intensionality. En: Aristotelian Society, LXII (1968).
[21] Cfr. URMSON, James. Criteria of Intensionality. En: Aristotelian Society, LXII (1968).
[22] Cfr. SEARLE, Óp. cit., p. 180.
[23] Cfr. ZALTA, Edward. Intensional Logic and the Metaphysics of Intentionality. Cambridge: MIT. 1988.

1.1.3 La 'tesis de Brentano' bajo crítica: Desde finales del siglo pasado, bajo la influencia del giro mentalista, Tim Crane[24] ha supuesto que la posición de Brentano sobre la intencionalidad puede ser resumida en una sola tesis, la cual se sustenta principalmente en el pasaje citado. Según Crane[25], Brentano piensa que la intencionalidad, la propiedad que consiste en dirigirse o en referir a objetos, es la *marca exclusiva* de los fenómenos mentales. La 'tesis de Brentano', como él la ha denominado, afirma, por un lado, que la intencionalidad es la *marca* de lo mental y, por otro, que es *exclusiva* de esta clase de fenómenos. Enseguida indicaré algunos casos que ilustren que ambas partes de la tesis han sido objeto de debate en la tendencia analítica de la filosofía.

En primer lugar, cuando Brentano advirtió que la intencionalidad es la *marca* de lo mental quiso expresar que ella es una condición *necesaria* y *suficiente* para los fenómenos mentales[26]. Esto quiere decir, para Brentano, que todo fenómeno mental *necesita* ser intencional, esto es, que todo fenómeno mental debe dirigirse o referir a algo, y que este hecho, es decir, este dirigirse o referir a algo, es *suficiente* para que pueda afirmarse del fenómeno que es mental. Frente a esta posición, algunos autores han sostenido que la intencionalidad no puede ser una condición necesaria de lo mental, ni por tanto suficiente, ya que es posible concebir fenómenos mentales que no poseen la propiedad de la intencionalidad. McGinn[27], por ejemplo, piensa que el carácter cualitativo de la experiencia, el cual remite a los famosos *qualia*, no está dirigido a algo; Searle[28], por su parte, piensa que existen emociones, como ciertas formas de ansiedad, de angustia y de depresión, que no refieren a un objeto determinado. Asimismo, otros autores piensan que hay propiedades de los fenómenos

---

[24] Cfr. CRANE, Tim. Intentionality as the mark of the mental. En: O'Hear, A. (Ed.). Contemporary Issue in the Philosophy of Mind. Cambridge: Cambridge University Pres, 1998.
[25] Cfr. Íbid., 1.
[26] Recuérdese: Q es una condición necesaria para P si y sólo si, si P, entonces Q; por su parte, Q es una condición suficiente para P si y sólo si, si Q, entonces P.
[27] Cfr. MCGINN, Colin. The Character of Mind. New York: Oxford University Press. 1982.
[28] Cfr. SEARLE, Óp. cit., p. 1.

mentales que son más relevantes a la hora de describirlos y que por esa razón tendría más sentido afirmar que ellas son la marca de lo mental. Un ejemplo concreto es Strawson[29], quien afirma que los únicos fenómenos que inequívocamente pueden ser considerados mentales son los fenómenos de la experiencia consciente. Por su parte, siguiendo a Brentano, Crane[30] argumenta que la intencionalidad debe ser considerada la marca de los fenómenos mentales, pues los casos ofrecidos en contra (de Searle y de McGinn, por ejemplo) sólo prueban que no se ha entendido correctamente la intencionalidad ni tampoco sus relaciones con otras propiedades de la mente (como la experiencia consciente de Strawson). Según Crane, una comprensión adecuada de estas relaciones evidencia que todos los fenómenos mentales remiten a la intencionalidad, por lo que puede decirse con seguridad que ésta es la marca de lo mental.

En segundo lugar, cuando Brentano advirtió que la intencionalidad es una propiedad *exclusiva* de los fenómenos mentales quiso expresar que no hay fenómenos físicos que la exhiben, que la intencionalidad no puede, por tanto, predicarse de ningun fenómeno u objeto físico. Esta segunda parte de la tesis de Brentano ha sido quizá más polémica que la primera en la tendencia analítica de la filosofía. Para muchos autores, si la intencionalidad es definida como la propiedad que consiste en referir o dirigirse a algo, es evidente que hay fenómenos físicos que son intencionales. Por ejemplo, si bien es cierto que mi creencia de que Russell es filósofo es intencional (pues *es sobre* Russell), es evidente también que al escribir en un papel 'Russell es filósofo' las marcas de tinta, que tienen propiedades físicas, *refieren a* 'Russell' y lo hacen en un *sentido* específico como 'siendo filósofo'. Del mismo modo, los sonidos que salen de mi boca cuando digo 'Russell es filósofo' son, en lo esencial, objetos físicos y no obstante también *refieren a* Russell siendo filósofo. Esto mismo puede decirse de los mapas y de las señales en general. Autores con puntos de vista tan dispares acerca de la mente como Searle[31], Dretske[32], Haugeland[33] y Dennett[34] coinciden en que la intencionalidad

[29] Cfr. STRAWSON, Galen. Mental reality. Cambridge: MIT Press. 1994.
[30] Cfr. CRANE, Óp. cit.
[31] Cfr. SEARLE, Óp. cit., p. 26.
[32] Cfr. DRETSKE, Fred. Knowledge and the Flow of Information. Cambridge: MIT Press. 1981.
[33] Cfr. HAUGELAND, John. Having Thought. Cambridge: Harvard University Press. 1998.

de los fenómenos físicos es innegable y, por lo tanto, no es una propiedad exclusiva de los fenómenos mentales.

1.1.4 El problema de la atribución de intencionalidad: El hecho de que se haya negado la tesis de Brentano, en especial la parte en la que afirma que la intencionalidad es una propiedad exclusiva de lo mental, produjo un interés identificable en algunos autores de la tendencia analítica. Si la intencionalidad no es exclusiva de lo mental, es decir, si en efecto puede predicarse de objetos físicos tales como las manchas de tinta en un papel o los mapas, entonces surge el problema de establecer si la posible intencionalidad de tales objetos físicos es la misma intencionalidad que poseen los estados mentales. En el caso en el que se considere que se trata de una intencionalidad diferente, debe formularse una clasificación de la intencionalidad (pues habría que establecer varias clases o tipos intencionales) y además deben aclararse los criterios bajo los cuales es posible atribuir cada una de estas clases. En resumen, en este contexto surge lo que en la introducción llamé 'el problema de la atribución de intencionalidad'. Al considerar que existen distintas clases de intencionalidad, en efecto, el problema de la atribución de intencionalidad se hace patente ya que sería muy fácil confundir los sentidos en los que se dice que algo posee intencionalidad: aquí se hace válido preguntar acerca de la clase de intencionalidad que le es posible atribuir a sistemas naturales no-humanos, como los anima-les y las plantas, y acerca de la clase de intencionalidad que le es posible adscribir a otros sistemas, como los artificiales, entre ellos los sistemas de la Inteligencia Artificial. En buena parte, esta es la razón por la cual John Searle formula, en su teoría de la intencionalidad, una clasificación de la intencionalidad: una que distingue entre la intencionalidad de los fenómenos mentales y la intencionalidad de los fenómenos físicos.

---

[34] Cfr. DENNETT, Daniel. The Intentional Stance. Cambridge: MIT Press. 1987.

1.2 SISTEMAS ARTIFICIALES SIMBÓLICOS

La expresión 'sistema artificial' usada en el título de este trabajo es lo suficientemente ambigua como para merecer unas cuantas líneas de aclaración. Como sugerí en la introducción, al usar la palabra 'artificial' quiero hacer referencia a la disciplina de las Ciencias cognitivas que recibe el nombre de 'Inteligencia Artificial'. Sin embargo, como todas las disciplinas, la Inteligencia Artificial ha tenido distintos programas de investigación que han marcado profundamente su naturaleza. En ese sentido, aunque sea brevemente, en orden a aclarar la expresión 'sistema artificial', debo especificar con exactitud los modelos teóricos a los que me refiero.

1.2.1 Inteligencia Artificial Simbólica: El origen de la idea de la Inteligencia Artificial puede remontarse, por lo menos, a Alan Turing. En *Computing Machinery and Intelligence*[35], de 1950, Turing abordó la pregunta de si una máquina puede considerarse inteligente y propuso una prueba de inteligencia, el famoso 'Test de Turing' o 'Juego de Imitación'. El objetivo de la prueba era determinar si una máquina podía comportarse como se comporta un agente inteligente, como una persona, lo que para Turing era una señal de la inteligencia de la máquina. El test puede ser resumido de la siguiente manera:

Imagínese un entrevistador que está en un cuarto y dos participantes que están en otro cuarto: un hombre y una mujer. Si bien el entrevistador sabe que hay dos participantes en el segundo cuarto, ignora por completo cuál es el hombre y cuál es la mujer. Su tarea es identificarlos. Para ello, el entrevistador les envía preguntas por medio de un monitor con el objetivo de que las respuestas que le devuelvan, también por el monitor, le brinden alguna pista de quién es la mujer y quién es el hombre. Ahora bien, el hombre tiene la tarea de confundir al entrevistador y miente en sus respuestas para que éste piense que es la mujer, siendo en realidad el hombre. Por su parte, la mujer tiene la tarea de responder sinceramente e

[35] TURING, Alan. Computing Machinery and Intelligence. En: Mind, 49 (1950). 433-460.

intentar convencer al entrevistador de que ella, en efecto, es la mujer. Imagínese que, como es presumible, el entrevistador acierta algunas veces en la identificación y falla en algunas otras. Ahora bien, por último imagínese que reemplazamos al hombre por una máquina. Y aquí viene el punto: si la máquina es capaz de realizar perfectamente la tarea que tiene el hombre, es decir, de responder a las preguntas del entrevistador manteniendo una conversación fluida y convincente, confundiéndolo y logrando que falle en ocasiones, entonces puede afirmarse que la máquina tiene un comportamiento inteligente[36]. Al fin y al cabo ¿en qué se distingue su comportamiento del comportamiento del hombre?

Lo esencial de la prueba de Turing, a mi parecer, es la influencia que ejerció al modificar la pregunta central: ¿qué es pensar, qué es la inteligencia, qué es la mente? se transformó en ¿cómo podemos hacer que un sistema artificial, como una computadora, se comporte como se comporta un ser humano? Mientras que la primera pregunta puede ser considerada desorientadora, y para algunos incluso sería especulativa, la segunda brinda la posibilidad de un programa de investigación: el diseño de máquinas que tengan la capacidad de imitar el comportamiento humano inteligente. Por esta razón, la presentación del Test de Turing en 1950, por parte de Turing, con frecuencia suele ser considerado un hecho decisivo en la consolidación de la Inteligencia Artificial como disciplina y como proyecto investigativo.

Otro de los grandes aportes de Turing fue dado años antes cuando acuñó la noción 'Máquina de Turing', la cual expresa la idea de una máquina abstracta que lleva a cabo operaciones computacionales usando sólo dos tipos de símbolos (ceros y unos, por ejemplo). A partir del concepto de Máquina de Turing se estableció lo que se ha llamado la 'Tesis Church-Turing', según la cual cualquier algoritmo que sea computable puede ser computable en una Máquina de Turing. Searle afirma[37] que paralelamente a esta idea Turing sostuvo el concepto de Máquina de Turing Universal, cuyo valor es que puede simular el comportamiento de cualquier Máquina de Turing regular. Esto es particularmente relevante pues al vincular el programa investigativo del diseño de máquinas inteligentes (Inteligencia Artificial) con la

[36] Esta descripción tendrá relevancia también en el segundo capítulo del texto.
[37] SEARLE, John. Mind: A brief introduction. New York: Oxford University Press. pp. 66-72.

prueba que nos permite determinar cuándo una máquina tiene comportamiento inteligente (Test de Turing) y con la idea de una máquina abstracta que puede llevar a cabo cualquier cómputo (Máquina de Turing Universal) se obtiene un modelo teórico del modo en el que la mente podría funcionar, a saber, el modelo computacional de la mente[38].

1.2.2 Inteligencia Artificial Fuerte: El modelo computacional de la mente es el paradigma de las ciencias cognitivas que suele ser expresado con el eslogan: 'la mente es al cerebro lo que el software es al hardware de un computador'. En específico, lo que supone el modelo computacional de la mente es que los procesos cognitivos particulares, como la percepción, la memoria y el aprendizaje, son procesos algorítmicos que consisten en la manipulación o transformación de símbolos representados. En ese sentido, una de sus implicaciones más notables es la *tesis de la realizabilidad múltiple*, según la cual, dado que los tipos de estados mentales no pueden identificarse con los tipos de estados cerebrales, aquéllos pueden tener múltiples realizaciones físicas[39]. En ese contexto, un sistema artificial de circuitos o *chips*, por ejemplo, podría llevar a cabo procesos cognitivos tal y como lo hace un cerebro biológico.

Las críticas al programa investigativo de diseñar máquinas que puedan demostrar inteligencia han sido numerosas. Entre los más destacados críticos en filosofía está Hubert Dreyfus, quien se enfocó en develar los presupuestos del programa cognitivista de la Inteligencia Artificial y John Searle, quien escribió un artículo[40] en el cual atacó la tesis del computacionalismo sostenida, según él, por la Inteligencia Artificial (IA). En el artículo Searle divide la IA en dos

---

[38] Evidentemente hubo más elementos que influyeron. En general, implicó todo lo que se conoce como revolución cognitiva: la inviabilidad del conductismo, los desarrollos de Newell y Simon en computación, de Chomsky en lingüística, etc.

[39] En la medida en que lo que define a un estado mental no es su correlato físico sino su rol funcional, diferentes instancias físicas –como los chips de un computador– pueden tener estados mentales. Esto se debe principalmente a Putnam. Cfr. PUTNAM, Hilary. The nature of mental states. Cambridge: Cambridge University Press. 1975.

[40] SEARLE. Minds, Brains... Óp. cit.

categorías: IA fuerte e IA débil. La distinción, que ha mantenido desde entonces y que cobró cierta fama, se basa en los objetivos que tiene cada una de las versiones. Según Searle, el objetivo de la IA en su versión débil es utilizar el computador como una herramienta para formular y comprobar hipótesis acerca de la mente; en cambio, dice Searle, el objetivo de la IA en su versión fuerte es diseñar computadores que tengan la capacidad de llevar a cabo procesos mentales. Para la IA fuerte, el computador no es una simple herramienta que puede ser utilizada al estudiar la mente sino que, en realidad, *es* una mente. Searle asegura, además, que la IA fuerte no sólo afirma que el computador programado es capaz de realizar procesos mentales o cognitivos, sino que también los explica, en el sentido de que clarifica en qué consiste llevar a cabo tales procesos mentales[41].

Ahora bien, la IA en su versión fuerte entiende los estados o procesos mentales como estados o procesos computacionales, pues supone que llevar a cabo tales procesos consiste, básicamente, en instanciar un programa de computador[42]. En concreto, para la IA fuerte los procesos mentales consisten en recibir entradas sensoriales (*inputs*), instanciar programas que se definen de manera formal o sintáctica, es decir, manipulando símbolos representados, y en proporcionar salidas (*outputs*). Un ejemplo paradigmático que Searle menciona en su artículo es el programa de IA de Schank y Abelson, un programa de comprensión-de-relatos en el que parece que un computador es capaz de responder preguntas (*outputs*) acerca de un relato (*inputs*), tal y como lo haría una persona en condiciones normales. El programa de comprensión-de-relatos podría superar la prueba de Turing en la medida en que es capaz de mantener una conversación fluida y convincente que podría engañar a cualquier entrevistador. Parte de la crítica de Searle a la IA fuerte, como veremos en el segundo capítulo, se enfoca en este aspecto.

1.2.3 *Good Old-Fashioned Artificial Intelligence*: John Haugeland introdujo el acrónimo GOFAI (*Good Old-Fashioned Artificial Intelligence*) en el tercer capítulo de *Artificial*

[41] Cfr. Íbid., 1.
[42] Cfr. Íbid., 1.

Intelligence: The Very Idea[43]. Las siglas remiten a la fase temprana de la Inteligencia Artificial, la cual podría identificarse con la IA fuerte de Searle (más que con la IA débil) pues comparte su supuesto básico: llevar a cabo procesos mentales es lo mismo que llevar a cabo procesos computacionales. Según Haugeland, "GOFAI, as a branch of cognitive science, rests on a particular theory of intelligence and thought-essentially Hobbes's idea that ratiocination is computation"[44]. Recuérdese que el filósofo inglés Thomas Hobbes sostuvo una posición sobre la capacidad de pensar que poseen los seres humanos en la que incluyó la idea de computar. Tal posición, que suele ser expresada bajo la frase 'razonar es computar', puede ser extraída de su *Tratado sobre el cuerpo*:

> Por razonamiento entiendo la computación. Y computación es hallar la suma de varias cosas añadidas o conocer lo que queda cuando de una cosa se quita otra. Por lo tanto razonar es lo mismo que sumar y restar, y si alguien añade a esto multiplicar y dividir, no estoy de acuerdo ya que la multiplicación es la suma de cosas iguales, y la division la resta de cosas iguales cuantas veces se pueda hacer. Por lo tanto todo razonamiento se reduce a estas dos operaciones de la mente: la suma y la resta[45].

Para Hobbes, pues, el razonamiento puede ser reducido a la capacidad de realizar operaciones de cómputo. Y a Haugeland le interesa mencionarlo porque, según él, las teorías GOFAI se caracterizan en que sostienen que nuestra capacidad de razonar, de pensar, etc. equivale a la manipulación de símbolos que podría llevar a cabo un artefacto, un sistema artificial. Como

---

[43] "GOFAI es una rama de la ciencia cognitiva que descansa sobre una particular teoría de la inteligencia y del pensamiento: esencialmente la idea de Hobbes según la cual razonar es computar". Traducción propia. HAUGELAND. The Artificial... Óp. cit. p. 112.

[44] Íbid., p. 112.

[45] Páginas después Hobbes redujo las proposiciones y los silogismos, es decir, formas de razonar del hombre, a operaciones de adición. Su postura cobra sentido si se tiene en cuenta además el entusiasmo intelectual de su época por el desarrollo de la matemática y el desafío –que Hobbes tomó a nombre propio– de brindar explicaciones materialistas del hombre. HOBBES, Thomas. Tratado sobre el cuerpo. Madrid: Universidad Nacional de Educación a Distancia. 2009. p. 38

dije, me parece que es posible identificar a las teorías GOFAI con la Inteligencia Artificial Fuerte y en general con toda Inteligencia Artificial Simbólica.

En resumen, al usar la expresión 'sistema artificial' estoy pensando en la fase temprana de la Inteligencia Artificial que suele ser vinculada con el cognitivismo de las Ciencias cognitivas y en general con el modelo computacional de la mente. Como mencioné, la IA fuerte y GOFAI sostuvieron en su momento que un artefacto diseñado para imitar el comportamiento humano, si es programado de modo adecuado, puede realizar procesos cognitivos bajo representaciones. Desde el punto de vista de la intencionalidad, esto significa que tales artefactos tendrían estados intencionales. Tendrían creencias y deseos, podrían comprender (como el programa de comprensión-de-relatos) y por lo tanto habría que considerarlos como agentes intencionales. Es en este contexto en el que surgen preguntas como ¿qué podemos decir acerca de la intencionalidad de los sistemas artificiales, los cuales, según cierta rama de la Inteligencia Artificial, pueden poseer estados intencionales? Más aun ¿qué podemos decir acerca de nuestra propia intencionalidad? ¿En qué sentido decimos de ella que es *propia*? Si se considera que las anteriores preguntas tienen algún sentido y permiten alguna respuesta determinada, doy por logrado el objetivo del capítulo. En ese caso, a partir de ahora quiero enfocarme en presentar la posición de Searle acerca de la intencionalidad de los sistemas artificiales.

CAPÍTULO 2: INTENCIONALIDAD EN SISTEMAS ARTIFICIALES

En el capítulo anterior se consideró la posibilidad de atribuir la propiedad de la intencionalidad a fenómenos físicos y se terminó sugiriendo que los sistemas artificiales, como los de la IA fuerte o GOFAI, ostentan alguna clase intencional que requiere ser examinada. El presente capítulo, por su parte, busca determinar el sentido en el que es posible decir que un sistema artificial tiene intencionalidad. La fuente cardinal para el examen, como mencioné en la introducción, será la teoría de la intencionalidad del norteamericano John Searle. Mi decisión de escoger su teoría de la intencionalidad se debe ante todo a que brinda una clasificación lo suficientemente clara, aunque no por ello correcta, del fenómeno intencional. En un primer momento presentaré su teoría de la intencionalidad en orden a exponer sus clases intencionales. En un segundo momento, con el propósito de determinar la clase de intencionalidad que le corresponde a los sistemas artificiales de IA fuerte o GOFAI, presentaré su experimento mental de la Habitación China.

2.1 LA TEORÍA DE LA INTENCIONALIDAD DE JOHN SEARLE

A causa del giro mentalista, la mayoría de filósofos de tendencia analítica que se formaron en torno a los problemas del lenguaje situaron su mirada en las preguntas sobre la mente y sus propiedades. El norteamericano John Searle es un excelente ejemplo de lo anterior pues, tras haber escrito dos libros sobre problemas de filosofía del lenguaje (*Speech Acts*[46] y *Expression and Meaning*[47]), escribe un tercer libro que expone una teoría de la intencionalidad (*Intentionality*[48]) y que lleva como subtítulo 'An essay in the Philosophy of Mind' ('Un ensayo en la filosofía de la mente')[49]. Con base en este último libro, *Intentionality*, a

[46] SEARLE, John. Actos de habla. España: Ediciones Cátedra. 1994.
[47] SEARLE, John. Expression and Meaning. Cambridge: Cambridge University Press. 1979.
[48] SEARLE, John. Intencionalidad. Madrid: Editorial Tecnos. 1992.
[49] Hay además una razón filosófica de peso que explica el paso que da Searle de la filosofía del lenguaje a la filosofía de la mente. Para Searle, la capacidad que posee el lenguaje para representar es una extension de las capacidades biológicas de la mente (para representar). Por ello Searle cree que la filosofía del lenguaje debe ser entendida como una rama de la filosofía de la mente.

continuación presentaré los aspectos elementales de la teoría de la intencionalidad de Searle, aspectos que me permitirán, más adelante, exponer su clasificación de la intencionalidad.

En *Intentionality*, Searle retoma la formulación según la cual la intencionalidad es la propiedad de los estados mentales que consiste en dirigirse a, ser sobre o de, un objeto o estado de cosas del mundo[50]. Esto significa que, en principio, Searle acepta el uso tradicional de la palabra 'intencionalidad', establecido, como se vio, por Brentano. Sin embargo, aclara que bajo su punto de vista no todos los estados mentales son intencionales, pues hay emociones que no están dirigidas a algo, como ciertos casos de nerviosismo y de ansiedad. En ese sentido, aunque por un lado acepta el uso tradicional de la palabra, por otro lado Searle niega la tesis de Brentano que establece que la intencionalidad es una condición necesaria y suficiente de los estados mentales[51]. En realidad, la definición de Brentano le sirve a Searle como formulación preliminar a partir de la cual va a desarrollar su teoría de la intencionalidad. Para Searle, la definición postula una *relación* entre los estados intencionales y los objetos a los que se dirigen. Su propósito, en un primer momento, es aclarar en qué consiste exactamente tal relación. Para ello, en el primer capítulo de *Intentionality*, Searle usa como recurso heurístico el trabajo que había realizado años atrás en *Speech Acts*, el libro en donde expone su teoría del lenguaje con base en una interpretación de los 'actos de habla' de Austin.

2.1.1 Intencionalidad: el modelo de los actos de habla: Searle asume como hipótesis que los estados intencionales 'representan' objetos y estados de cosas en el mismo sentido en el que los actos de habla *representan* objetos y estados de cosas[52]. Así, establece un plan de trabajo: si se quiere entender la relación entre los estados intencionales y los objetos a los que se dirigen, si se quiere entender la intencionalidad, debe prestarse atención al sentido en el que los actos de habla representan sus objetos. Para llevar a cabo el plan de trabajo, la estrategia de

[50] Íbid., p. 17.
[51] Por esa razón, para referirme a los estados mentales que sí son intencionales abreviaré de ahora en adelante usando la expresión 'estados intencionales'.
[52] Íbid., p. 17.

exposición de Searle consiste en mostrar las conexiones que existen entre los estados intencionales y los actos de habla. Según Searle, hay –por lo menos– cuatro puntos de similitud importantes que deben ser tomados en cuenta.

En primer lugar, la distinción que aparece en *Speech Acts*[53] entre el contenido proposicional y la fuerza ilocucionaria también se aplica a los estados intencionales. El concepto 'fuerza ilocucionaria', tomado directamente de Austin[54], y que para Searle es equivalente a 'acto ilocucionario', alude al hecho de que al emitir una oración un hablante de una lengua puede realizar distintos actos respecto del contenido de esa oración. Básicamente, la idea que subyace al concepto de fuerza ilocucionaria y de acto ilocucionario es que los hablantes de una lengua natural pueden *hacer cosas con palabras*. Repárese en las siguientes oraciones:

(5) Paula aprende a leer
(6) ¿Paula aprende a leer?
(7) ¡Paula, aprende a leer!

Según Searle, al emitir (5), (6) y (7) el hablante tiene una misma referencia: 'Paula'. Asimismo, tiene un mismo contenido proposicional: 'aprende a leer'. Pero, aún compartiendo la referencia y el contenido proposicional, al emitir (5), (6) y (7) el hablante está realizando cosas totalmente distintas. En (5), por ejemplo, el hablante está enunciando algo; en (6), el hablante está haciendo una pregunta, y en (7) está dando una orden. Al proferir estas oraciones, el hablante las está usando de una manera específica y, al usarlas, *está haciendo cosas en el mundo*. A estos actos que el hablante realiza cuando profiere las oraciones, tanto Austin como Searle los llaman 'actos ilocucionarios'. Ahora bien, Searle afirma que la distinción entre el contenido proposicional ('aprende a leer') y el acto ilocucionario ('enunciar', 'ordenar', 'preguntar') también se presenta en los estados intencionales. Repárese en las siguientes oraciones de actitudes proposicionales:

[53] SEARLE. Actos... p. 32-33.
[54] Cfr. AUSTIN, John. How to do things with words. London: Oxford University Press. 1962.

(8) Paula cree que hoy es viernes

(9) Paula desea que hoy sea viernes

(10) Paula teme que hoy sea viernes

En (8), (9) y (10) la referencia es la misma, a saber: 'Paula'. El contenido, que en este caso Searle llama a veces 'representativo' y a veces 'proposicional', también es el mismo: 'que hoy es (o sea) viernes'. Sin embargo, (8), (9) y (10) se diferencian en cuanto tienen un aspecto diferente, lo que Searle llama el 'modo psicológico'. En (8) el modo psicológico que se tiene respecto del contenido representativo 'hoy es viernes' es la creencia; en (9), el modo psicológico es el deseo, y en (10) es el temor. Pero el modo psicológico no se confunde con el contenido representativo, es decir, con el contenido que se representa al tener los estados intencionales. En resumen, según Searle, todo estado intencional se caracteriza por tener un modo psicológico que, lo mismo que el acto de habla respecto del contenido proposicional, se diferencia de su contenido.

En segundo lugar, tanto los actos de habla como los estados intencionales se caracterizan porque poseen lo que Searle llama 'direcciones de adecuación'. Según Searle, la mayoría de actos de habla tienen alguna de las dos direcciones posibles: palabra-a-mundo o mundo-a-palabra[55]. Por ejemplo, el acto de habla (5), en el que se enuncia que Paula aprende a leer, tiene la dirección de adecuación palabra-a-mundo pues el enunciado *debe adecuarse al mundo* para que podamos decir de él que es verdadero. Por su parte, el acto de habla que se da en (7), en el que se le ordena a Paula que aprenda a leer, tiene una dirección de adecuación mundo-a-palabra, pues al usar (7) el hablante espera de algún modo que *el mundo se adecúe a su orden*, es decir que se genere un cambio en el mundo: en ese caso específico, que Paula aprenda a leer[56]. Según Searle, la mayoría de actos de habla tienen alguna de las dos direcciones de adecuación.

[55] SEARLE. Intencionalidad... pp. 22-23.

[56] Searle da un ejemplo fuera del contexto, muy oportuno, sobre las direcciones de adecuación: "Si Cenicienta entra en una zapatería a comprar un par de zapatos nuevos, ella considera la talla de su pie como dada y busca zapatos que se ajusten a ella (dirección de ajuste zapato-a-pie) pero cuando el príncipe busca a la propietaria del zapato él toma el zapato como dado y busca un pie que se ajuste al

De la misma manera, el estado intencional (8), en el que Paula cree que es viernes, tiene la dirección de adecuación mente-a-mundo, pues es la creencia la que *debe adecuarse al mundo* para que podamos decir de ella, de la creencia, que es verdadera. En el caso de (9), en donde Paula desea que hoy sea viernes, el estado intencional tiene una dirección de adecuación mundo-a-mente, pues, por decirlo así, *es el mundo el que debe adecuarse al deseo* para que podamos decir del estado intencional, del deseo, que se satisfizo. Asimismo, el estado intencional (10), en el que Paula teme que hoy sea viernes, tiene dirección de adecuación mundo-a-mente, pues es el mundo –y no el estado intencional– el que *debe adecuarse* para que pueda decirse del temor que se cumplió. En otras palabras, en el caso de la creencia, es ella la que debe ser de determinada manera para que sea verdadera; en el caso del deseo y del temor, al contrario, es el mundo el que debe ser de determinada manera para que el deseo se satisfaga y para que el temor se cumpla. Según Searle, la mayoría de estados intencionales tienen, dependiendo de su modo psicológico, una dirección de adecuación: mente-a-mundo o mundo-a-mente.

Por otro lado, una tercera conexión consiste en que tanto los actos de habla como los estados intencionales se caracterizan porque poseen lo que Searle llama 'condiciones de satisfacción'. En palabras de Searle: "Las condiciones de satisfacción son aquellas condiciones que [...] deben darse si el estado se satisface"[57]. Para explicar esto recuérdese que en la anterior conexión Searle establece que los actos de habla y los estados intencionales pueden tener 'dirección de adecuación'. Para Searle, si el acto de habla tiene una dirección de adecuación, ya sea palabra-a-mundo o mundo-a-palabra, en cada caso podrá determinarse el éxito o fracaso del acto de habla. Por ejemplo, si el hablante en (7) le ordena a Paula que aprenda a leer y Paula en efecto aprende a leer (o por lo menos lo intenta), puede decirse de ese acto de habla, de la orden, que fue obedecida. Decir que la orden fue obedecida significa que cumplió sus condiciones de satisfacción, que se satisfizo. En el mismo sentido, si el estado intencional tiene dirección de adecuación, ya sea mente-a-mundo o mundo-a-mente, podrá determinarse el éxito o el fracaso del estado intencional. Si, como en (8), Paula cree que hoy es viernes y en

---

zapato (dirección de ajuste pie-a-zapato". Aunque el traductor opta por traducir aquí 'direction of fit' como 'dirección de ajuste', me parece más exacta la expresión 'dirección de adecuación'. Íbid., p. 23.

[57] Íbid., p. 28.

efecto hoy es viernes, puede decirse de su creencia que es verdadera. Y decir que una creencia es verdadera significa que cumple sus condiciones de satisfacción. En resumen, si el acto de habla o el estado intencional posee dirección de adecuación podrá determinarse en cada caso si cumple sus condiciones de satisfacción, es decir, podrá determinarse si el acto de habla o el estado intencional es satisfecho.

Por último, la cuarta conexión[58] está en el hecho de que los actos de habla son expresiones de estados intencionales, una idea que ya aparecía con timidez en *Speech Acts*[59]. Este caso es más que una simple similitud, pues, para ponerlo en palabras de Searle, "[...] el estado intencional expresado no es sólo un acompañamiento de la realización del acto de habla"[60]. Con esto Searle quiere decir que se trata de una conexión *interna*, una que puede ser ilustrada de la siguiente manera. Cuando un hablante de una lengua usa el acto de habla (7), en el que se le ordena a Paula que aprenda a leer, además de estar realizando el acto ilocucionario 'ordenar', está expresando un estado intencional particular, a saber, el deseo de que Paula aprenda a leer. En el mismo sentido, si el hablante usa el acto de habla (5), en el que se enuncia que Paula está aprendiendo a leer, además de realizar el acto de enunciar, el hablante está expresando su creencia de que Paula está aprendiendo a leer. De modo general, si el hablante enuncia *p*, expresa la creencia de que *p*; si da la orden de hacer *q*, expresa el deseo o el anhelo de que se haga *q*; si promete *r*, expresa la intención de hacer *r*, y así sucesivamente. En resumen, es una característica de los estados intencionales el hecho de que pueden ser expresados por actos de habla.

Ahora bien, decir que los actos de habla son expresiones de estados intencionales significa, para Searle, que si un acto de habla posee dirección de adecuación, se satisfará (cumplirá sus condiciones de satisfacción) si y sólo si el estado intencional se satisface (si cumple sus condiciones de satisfacción). En ese caso, las condiciones de satisfacción del acto de habla se muestran siendo idénticas a las condiciones de satisfacción del estado intencional. Por

[58] En la exposición que lleva a cabo Searle esta conexión es la tercera y no la cuarta. Hice el cambio porque, a mi juicio, así resulta pedagógicamente más fácil la presentación.
[59] Cfr. SEARLE, Actos... p. 78.
[60] SEARLE. Intencionalidad... p. 24.

ejemplo, el acto de habla (5), en el que se enuncia que Paula aprende a leer, expresa el estado intencional de la creencia de que Paula aprende a leer. Las condiciones de satisfacción del acto de habla (5) consisten en que en efecto Paula aprenda a leer y las condiciones de satisfacción de la creencia de que Paula aprende a leer son, también, que Paula en efecto aprenda a leer. En ambos casos, pues, se trata de las mismas condiciones de satisfacción.

2.1.2 La estructura de la intencionalidad: Después de exponer los puntos de conexión entre actos de habla y estados intencionales, se obtiene una imagen general de la estructura de la intencionalidad. Recuérdese que el objetivo era aclarar en qué sentido se dice que los estados intencionales 'representan' sus objetos y estados de cosas. En palabras de Searle:

> El sentido de «representación» en cuestión pretende ser enteramente agotado por la analogía con los actos de habla: el sentido de «representar» en el que una creencia representa sus condiciones de satisfacción es el mismo sentido en el que un enunciado representa sus condiciones de satisfacción. Decir que una creencia es una representación es simplemente decir que tiene un contenido proposicional y un modo psicológico, que su contenido proposicional determina un conjunto de condiciones de satisfacción bajo ciertos aspectos, que su modo psicológico determina una dirección de ajuste [de adecuación] de su contenido proposicional, en el sentido en que todas estas nociones –contenido proposicional, dirección de ajuste [de adecuación], etc.– son explicadas por la teoría de los actos de habla[61].

La estructura de los estados intencionales es similar, pues, a la estructura de los actos de habla: en ambos casos hay una diferencia entre el contenido proposicional y el modo en que se da ese contenido; en ambos casos el modo en que se da ese contenido determina la dirección de adecuación y el contenido proposicional determina las condiciones de satisfacción, y en ambos casos son las condiciones de satisfacción aquello que se representa bajo un modo

[61] Íbid., p. 27.

particular. Es este el sentido en el que Searle afirma que los estados intencionales *representan* sus objetos y estados de cosas.

2.1.3 La intencionalidad de los actos de habla: Ahora bien, cuando Searle afirma que los estados mentales son intencionales en el sentido de que *representan* tal y como los actos de habla *representan*, también está afirmando que los actos de habla son intencionales. En otras palabras, si el sentido de representar en ambos casos es el mismo, si los actos de habla están dirigidos, son sobre o de, objetos y estados de cosas como los estados intencionales lo están y lo son, entonces puede afirmarse que los actos de habla también ostentan intencionalidad. Sin embargo, pese a que los estados intencionales representan sus objetos y estados de cosas *en el mismo sentido* en el que los actos de habla representan los suyos, no lo hacen *por los mismos medios ni en la misma forma*[62]. Por un lado, para que un acto de habla represente sus objetos o estados de cosas es necesario que se realice en una entidad física. Es decir, para que el acto de habla exista tiene que ser expresado físicamente de manera escrita o hablada. Los estados intencionales, por su parte, *en tanto estados intencionales*, no lo requieren. Para que Paula tenga la creencia de que Russell es filósofo no tiene que escribirlo en un papel ni expresarlo con su voz: simplemente tiene la creencia. Por otro lado, según Searle, la intencionalidad de los estados mentales es 'intrínseca' a ellos mismos y la de los actos de habla no lo es. En sus palabras:

---

[62] Searle escribe exactamente: "Intentional states represent objects and states of affairs in the same sense that speech acts represent objects and states of affairs (though, to repeat, they do it by different means and in a different way)". SEARLE, Intentionality... p. 11. La traducción al castellano de esa frase, en la versión que he estado citando, es innecesariamente confusa: "Los estados Intencionales representan objetos y estados de cosas en el mismo sentido en que los actos de habla representan objetos y estados de cosas (aunque, insisto, lo hacen en sentidos distintos y de una manera diferente)". SEARLE. Intencionalidad... p. 26.

> La realización efectiva en la que el acto de habla se lleva a cabo, implicará la producción [...] de alguna entidad física, tales como los sonidos hechos con la boca o las marcas en el papel. Las creencias, temores, esperanzas y deseos, por otro lado, son intrínsecamente intencionales. Caracterizarlos como creencias, temores, esperanzas y deseos es atribuirles ya intencionalidad. Pero los actos de habla tienen un nivel de materialización física, *qua* actos de habla, que no es intrínsecamente intencional. No hay nada intrínsecamente intencional en los productos del acto de emisión, esto es: los ruidos que salen de mi boca o las marcas que hago sobre el papel. Una emisión puede tener intencionalidad, como una creencia tiene intencionalidad, pero, mientras que la intencionalidad de la creencia es *intrínseca*, la intencionalidad de la emisión es *derivada*[63].

Con ello se establece que hay dos formas, por lo pronto, de tener intencionalidad: intrínsecamente, como en el caso de los estados mentales, y derivadamente, como en el caso de los actos de habla[64]. Ahora bien, en *Intentionality* Searle no brinda como tal una definición explícita de 'intencionalidad intrínseca' ni de 'intencionalidad derivada'. En los capítulos siguientes del libro su interés se enfoca, más bien, en desarrollar aspectos de su teoría de la intencionalidad, como es el caso de la intencionalidad de la percepción y de la acción, y en profundizar en algunos elementos como el trasfondo (*background*). Sin embargo, me parece que es posible formarse una idea bastante clara de las formas o clases intencionales si se atiende, por ejemplo, a textos como *Mind, Language and Society*, de 1999.

2.1.4 Clases intencionales: intrínseca, derivada y "como-si": En *Mind, Language and Society*[65], que es el libro que Searle escribe para mostrar el modo en el que sus posturas filosóficas encajan en nuestra concepción natural del universo, se introduce una distinción ontológica que, según dice, es la que le permite sostener a su vez la distinción entre intencionalidad

[63] Íbid., p. 41.
[64] En realidad, la intencionalidad de los actos de habla aplica en general al lenguaje, pues todo lenguaje tiene la capacidad de estar dirigido a, ser sobre o de, objetos y estados de cosas del mundo, es decir, tiene la capacidad de representar.
[65] SEARLE, John. Mind, Language and Society. New York: Masterminds. 1999.

intrínseca e intencionalidad derivada. Explicaré la distinción ontológica para entender la distinción entre las dos clases intencionales.

Para Searle, hay rasgos del mundo que existen al margen de lo que los seres humanos pensemos o hagamos, existen independientes de nosotros, y hay otros rasgos que son relativos a nosotros, que dependen de lo que pensemos y hagamos. Piénsese, por ejemplo, en una silla. Por un lado, la silla se caracteriza porque tiene propiedades físicas (masa, átomos, estructura molecular, etc.); por otro lado, también se caracteriza por el hecho de que es silla, de que alguien hizo que fuese silla y de que alguien la usa como silla. La distinción radica en que lo primero no depende de nosotros los seres humanos (si no existiéramos ese pedazo de objeto seguiría teniendo masa, átomos, etc.) y lo segundo sí depende de nosotros (la silla es un artefacto para sentarse porque nosotros los seres humanos la usamos de esa manera, porque nosotros la pensamos como silla). Ahora bien, a los rasgos que son independientes de nosotros Searle los llama 'independientes del observador' (*observer-independent*) y a los que dependen de nosotros los llama 'dependientes del o relativos al observador' (*observer-dependent/observer-relative*)[66]. En este contexto, Searle afirma que la intencionalidad de los estados mentales, como las creencias y los deseos, es intrínseca a ellos porque es independiente de cualquier observador externo. Afirma, también, que la intencionalidad de los actos de habla, del lenguaje en general, es relativa al observador, pues depende del significado que nosotros los seres humanos le demos a las palabras. Repárese en los siguientes casos:

(11) Paula está hambrienta

(12) En inglés 'Paula is hungry' significa 'Paula está hambrienta'

[66] Al considerarlo desde el punto de vista del objeto de las ciencias, Searle afirma: "In general, the natural sciences deal with observer-independent phenomena, the social sciences with the observer dependent" Íbid., p. 6. ["En general, las ciencias naturales tratan con los fenómenos independientes del observador, las ciencias sociales con los que dependen del observador"]. Traducción propia. Lo anterior podría hacer pensar que la distinción no pasa de ser la distinción objetivo/subjetivo, pero esto no resulta del todo claro. Sin lugar a dudas, la cuestión requiere un estudio propio.

Por un lado, (11) es un estado intencional (que Paula esté hambrienta quiere decir, entre otras cosas, que tiene el deseo de comer) que no es relativo al observador, pues nadie tiene que pensar que Paula está hambrienta para que en efecto Paula lo esté. Ella simplemente lo está sin importar si alguien lo considera así. Por otro lado, (12) es un acto de habla que es dependiente del observador pues el hecho de que en inglés 'Paula is hungry' signifique 'Paula está hambrienta' implica que existan personas (los usuarios del idioma inglés) que piensen que esa serie de palabras, ordenadas de esa manera, tiene ese significado. En otras palabras, (12) depende totalmente de que las personas lo piensen así y en ese sentido se dice que es relativo al observador.

Ahora bien, Searle también introduce en *Minds, Language and Society* otra distinción respecto de las formas en las que algo puede tener intencionalidad. Considérese lo siguiente:

(13) Las plantas del jardín están hambrientas de nutrientes

Se dijo ya que (11) ostenta intencionalidad intrínseca y que (12) tiene intencionalidad derivada pero, en el caso de (13) ¿hay intencionalidad intrínseca o derivada? Para mostrarlo, Searle introduce el concepto de 'intencionalidad como-si'[67] (*as-if intentionality*), que no es como tal una clase intencional, pues no puede decirse, en estricto sentido, que tenga intencionalidad. La diferencia entre (11) y (13), por ejemplo, radica en que cuando se dice (11) la afirmación es usada *literalmente* para hacer referencia al hecho de que Paula *realmente* tiene hambre (es decir, para hacer referencia a que Paula tiene un estado mental, el deseo de comer) y que ese hecho es independiente de cualquier observador; en (13), por su parte, la afirmación del hablante es usada *metafóricamente* para indicar que las plantas se han marchitado por la falta de nutrientes, no queriendo decir con ello que la planta *realmente* tiene el deseo de comer. Ahora bien, la diferencia entre (12) y (13) radica en que cuando se dice (12) la afirmación es usada *literalmente* para hacer referencia al hecho de que el inglés 'Paula is hungry' significa realmente 'Paula está hambrienta', aún cuando este hecho dependa del observador, que son

[67] Íbid., p. 93.

los hablantes del inglés; pero en (13), repito, la afirmación es metafórica, no es literal. En ese sentido, (13) tiene intencionalidad *como-si* la tuviera, pero en realidad no la tiene.

Dado lo anterior, en orden a suministrar claridad respecto de la tres formas de atribuir intencionalidad para Searle (intencionalidad intrínseca, intencionalidad derivada e intencionalidad como-si), una opción de definición de estas clases intencionales podría ser la siguiente:

Intencionalidad intrínseca: También llamada intencionalidad original[68]. Esta es la primera clase de intencionalidad real. Es intrínseca, es decir, es inherente a los estados intencionales porque es independiente del observador. Esto significa que su intencionalidad no deriva de nada, sino que se da naturalmente. Es la clase de intencionalidad que ha sido desarrollada biológicamente por los cerebros y, en ese sentido, sólo la ostentan personas y animales[69].

Intencionalidad derivada: Esta es la segunda clase de intencionalidad real. Es real pues no hay nada metafórico en ella; pero, a diferencia de la intencionalidad intrínseca, ésta depende del observador. Se dice que es derivada porque deriva de la intencionalidad intrínseca y obtiene su intencionalidad por ella. Es la clase de intencionalidad que posee el lenguaje (las palabras, las marcas de tinta en el papel, los sonidos), las imágenes, las señales, los mapas, elementos físicos a los que (la intencionalidad intrínseca de) las personas han dotado de intencionalidad[70].

Intencionalidad como-si: Esta no es propiamente una clase intencional. Los casos de intencionalidad-como-si son casos en los que se le atribuye intencionalidad a algo que es como si la tuviera. En ese sentido, es usada como una metáfora, no es literal ni real. Es importante mencionarla ya que suele generar confusiones. En general, en estos casos es común decir que algo tiene estados intencionales cuando pareciera que ostenta intencionalidad pero en

[68] Cfr. SEARLE. Mind: a brief... pp. 7-29.
[69] Cfr. Íbid., p. 93-94.
[70] Cfr. Íbid., p. 93-94.

realidad, literalmente, no lo hace. Un ejemplo son los sistemas naturales no-animales, como las plantas. Es también la intencionalidad de los artefactos[71].

Habiendo establecido distinciones claras respecto de la formas o clases intencionales, ahora tiene sentido preguntar a cuál de aquéllas corresponde la intencionalidad que podría atribuirse a un sistema artificial del tipo de la Inteligencia Artificial Fuerte o de tipo GOFAI. ¿Cuál es la respuesta de Searle? En principio, pareciera que la intencionalidad que le es posible atribuir a un sistema artificial es la intencionalidad como-si. Si se repara en el hecho de que los sistemas artificiales son artefactos creados por el hombre y en que por supuesto no son cerebros biológicos, entonces la intencionalidad que habría que atribuirles es, en efecto, metafórica. Por ejemplo, cuando la computadora que está adelante muestra un aviso que dice 'conecte cargador para alimentación' no quiere decir *en realidad* que tiene el deseo de alimentarse, que su batería está, por decirlo así, sedienta de electricidad. Su aparente estado intencional no parece ser ni real ni tampoco literal; sólo parece metafórico. Sin embargo, lo cierto es que los sistemas artificiales de IA fuerte y de GOFAI, al igual que las computadoras digitales, no son simples artefactos, pues –si se recuerda la segunda sección del primer capítulo– tales sistemas pueden llegar a poseer la habilidad de sostener conversaciones tal como los seres humanos e incluso, como los programas de IA de Shack y Abelson, pueden superar satisfactoriamente pruebas de inteligencia como el Test de Turing. En diversos textos Searle parece pensar que el caso de los sistemas artificiales de este tipo es, en efecto, más complicado. En *Mind: a brief introduction*, Searle afirma:

> [...] if you think about it you will see that computation and syntax are observer relative. [...] Intrinsically it is a complex electronic circuit that we use to compute with. The electrical state transitions are intrinsic to the machine, but the computation is in the eye of the beholder. [...] The sense in which computation is in the machine is the sense in which information is in a book. It is there alright, but it is observer relative and not intrinsic[72].

[71] Cfr. Íbid., p. 93-94.

[72] "[...] Si piensas al respecto notarás que la computación y la sintaxis son relativas al observador. [...] Intrínsecamente es un circuito electrónico complejo que usamos para computar. Las transiciones de

De modo que no se trata propiamente de intencionalidad como-si, sino de la misma clase intencional que posee, por ejemplo, un libro. Recuérdese que la intencionalidad del libro sólo refiere a, es sobre o de, objetos y estados de cosas del mundo en cuanto que las personas lo piensan así, su intencionalidad no es intrínseca sino que depende totalmente del "ojo del espectador", como dice Searle. Y por lo tanto, su intencionalidad es derivada. Ahora bien, Searle afirma que es éste el sentido en el que se debe entender la intencionalidad de los sistemas artificiales que realizan computaciones. Su afirmación se debe a que estos sistemas (de la Inteligencia Artificial Fuerte y de tipo GOFAI) llevan a cabo operaciones que involucran *representaciones* simbólicas que, por definición, refieren a objetos y estados de cosas del mundo. Pero ¿en qué sentido exacto es que la intencionalidad de tales sistemas es derivada? En un primer momento, por cierto, pareció que la intencionalidad de estos sistemas artificiales era "como-si", simplemente metafórica; no obstante, Searle parece pensar que su intencionalidad es, más bien, derivada. Me parece que el sentido en el que la intencionalidad de estos sistemas es derivada quedará claro al exponer el experimento mental de la Habitación China de Searle, que es el objetivo de la siguiente sección.

2.2 LA HABITACIÓN CHINA Y LA INTENCIONALIDAD

Los experimentos mentales han sido un recurso, tanto de la ciencia como de la filosofía, desde los tiempos de la Antigua Grecia[73]. Aun cuando hoy día se discuta intensamente en filosofía de la ciencia acerca de su justificación epistemológica, sobre su validez científica, no puede negarse la influencia que ejercen, sobre todo en filosofía contemporánea. En concreto, el

---

estados eléctricos son intrínsecos a la máquina, pero la computación está en el ojo del espectador. [...] El sentido en el que la computación está en la máquina es el sentido en el que la información está en un libro. Está justo ahí pero es relativa al observador y no intrínseca". Traducción propia. SEARLE. Mind: a brief... p. 91.

[73] En ciencia, el Barco de Galileo, la famosa Paradoja de los Gemelos de Einstein, el Gato de Schrödinger y el Demonio de Maxwell, son un buen ejemplo de experimentos mentales. Asimismo, entre los casos más populares en filosofía se encuentran las famosas paradojas de Zenón sobre el infinito, la Hipótesis del Genio Maligno de Descartes durante su duda metódica, los ensayos imaginativos de Locke sobre el Problema de la identidad personal, el Problema de Molyneux en filosofía de la percepción, el Cuarto de Mary de Jackson y las Tierras Gemelas de Putnam en filosofía de la mente. Con este inventario, no obstante, no pretendo pasar por alto las obvias diferencias. Me interesa hacer evidente que se trata de un recurso muy común en filosofía. Diré algo al respecto en las conclusiones.

Argumento de la Habitación China de Searle (*The Chinese Room Argument*), el experimento mental presentado en el artículo de 1980 *Minds, Brains, and Programs* y reproducido en el libro *Minds, Brains and Science*[74], entre otros, ha recibido en los últimos treinta y seis años buena parte de la atención filosófica. En esta sección presentaré el experimento mental de Searle en orden a aclarar la clase intencional que puede ser atribuida a sistemas artificiales.

El objetivo de Searle al presentar su experimento mental, tal como se lee en el artículo de 1980, es atacar lo que allí llama 'Inteligencia Artificial Fuerte' (IA fuerte). Recuérdese que –como vimos en la segunda parte del primer capítulo– Searle atribuye a la IA fuerte el modelo computacional de la mente, en concreto en lo que respecta a estas dos afirmaciones: 1) un sistema artificial puede llevar a cabo procesos mentales si es programado de modo adecuado y 2) tal hecho explica la capacidad humana de llevar a cabo procesos mentales. Es decir, según Searle, la IA fuerte y –de paso GOFAI– afirma que un sistema artificial tiene, por ejemplo, la capacidad de comprender y este hecho explica la capacidad humana de comprender. Pues bien, la estrategia de Searle consiste en situar a una persona en un cuarto y en crear las condiciones para que simule el funcionamiento de un sistema artificial. En concreto, se trata de mostrar que los procesos que la persona lleva a cabo al simular al sistema se distinguen sustancialmente de sus procesos mentales. De ese modo, para Searle, las afirmaciones de la IA fuerte se verían seriamente debilitadas.

2.2.1 Searle en la Habitación China: Imagínese que encierran a Searle, quien no sabe ni una sola palabra del idioma chino, en una habitación en donde le suministran un conjunto de reglas escritas en inglés, un idioma que Searle comprende. Las reglas le permiten manipular símbolos chinos, que han sido introducidos a la habitación, estableciendo relaciones de tipo formal, es decir, identificando los símbolos por sus formas. Searle, por su parte, ignora que los símbolos son caracteres chinos: para él no se trata más que de una serie de garabatos sin significado alguno. Ahora bien, las reglas que le han suministrado le permiten responder a

[74] SEARLE, John. Mentes, cerebros y ciencia. España: Ediciones Cátedra. 1985.

algunas preguntas en chino que son introducidas a la habitación, aún cuando Searle no sabe que son preguntas. Simplemente, las reglas dicen cosas como: si introducen un signo 'changyuan-changyuan' devuelva un signo 'chongyuon-chong-yuon', de modo que desde el exterior pareciera que Searle está respondiendo a las preguntas en chino que le están haciendo. Más aún, las respuestas de Searle resultan tan buenas que son indistinguibles de las de un hablante nativo del chino. Sin embargo, pese a que en efecto son bastante buenas, lo cierto es que Searle no entiende ni una palabra del idioma chino: sólo está siguiendo reglas de tipo formal.

2.2.2 Intencionalidad derivada en Sistemas artificiales: En principio, el experimento mental de Searle es un ataque frontal al Test de Turing. Recuérdese que la prueba de Turing establecía que un sistema artificial podía ser considerado inteligente si imitaba el comportamiento de una persona en una conversación y si además lograba engañar a una persona externa. Esto es exactamente lo que hace Searle al simular el computador, pues al responder las preguntas que le han hecho en idioma chino parece –a los ojos de las personas que están afuera de la habitación– que él está manteniendo una conversación fluida en ese idioma. Sin embargo, resulta claro que su comprensión de las preguntas, y de hecho la compresión de sus propias respuestas, es completamente nula. Pero ¿por qué exactamente? Según Searle:

> Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in

> the minds of those who program them and those who use them, those who send in the input and those who interpret the output[75].

Además de atacar al Test de Turing, el experimento mental también se enfrenta a la tesis computacionalista según la cual los procesos mentales son procesos computacionales simbólicos, son procesos de manipulación de símbolos. En efecto, en la Habitación China Searle está simulando los procesos que llevaría a cabo un sistema artificial GOFAI o de IA fuerte, es decir, procesos computacionales, y sin embargo no comprende los símbolos que manipula. Esto se debe a que el significado de los símbolos, aquello que *representan*, a lo que refieren, está en la mente de los programadores, de los usuarios, de las personas que insertan las entradas (*inputs*) e interpretan las salidas (*outputs*). Incluso en el caso en el que se piense que el sistema artificial realmente podría tener representaciones, lo cierto es que éstas no tienen un valor independiente, son relativas al observador, dependen completamente de él. Por lo tanto, según Searle, habría que concluir que la intencionalidad de los sistemas artificiales de la Inteligencia Artificial Fuerte y de tipo GOFAI no es otra que intencionalidad derivada.

[75] "Porque las manipulaciones formales de símbolos como tal no tienen ninguna intencionalidad; no significan nada; ni siquiera son manipulaciones de símbolos, pues los símbolos no simbolizan nada. En la jerga lingüística, sólo tienen una sintaxis, pero no tienen semántica. Dicha intencionalidad que los computadores parecen tener está únicamente en las mentes de quienes los programan y aquellos que los utilizan, los que envían los inputs e interpretan los outputs". SEARLE. Minds, Brains... p. 422. Traducción propia.

CAPÍTULO 3: CONCLUSIONES

A lo largo de la presente monografía se estudió la naturaleza de la intencionalidad, así como sus tipos o clases, bajo el supuesto de que es posible determinar un uso correcto de la atribución de estados intencionales. En específico, el objetivo era determinar la clase intencional que le es correcto atribuir a un sistema artificial de tipo de la Inteligencia Artificial Fuerte o de tipo GOFAI. La respuesta, que –como es usual en filosofía– genera más preguntas que certezas, es que la intencionalidad de estos sistemas artificiales es derivada, es decir, no es intrínseca a ellos mismos. Esta sección tiene el objetivo de hacer un breve recuento de lo dicho en la monografía, de enunciar algunas consecuencias que se pueden extraer de allí y de presentar los posibles caminos de investigación que, concluido el trabajo, se han abierto.

En el primer capítulo hice básicamente dos cosas: 1) a partir de una revisión histórica del concepto de intencionalidad rescatado por Brentano, expuse algunos aspectos del fenómeno intencional que ocupó a ciertos filósofos de la tendencia analítica de la filosofía. Primero, expliqué el sentido de la discusión sobre la existencia o, más bien, la inexistencia del objeto intencional; segundo, por medio del caso de Chisholm, ilustré la forma en la que la filosofía analítica abordó en un primer momento los problemas acerca de la intencionalidad, entendiéndola como una propiedad del lenguaje con el que se reportan los fenómenos mentales intencionales, esto es, como intensionalidad o no-extensionalidad; tercero, mostré los dos sentidos en los que en la tendencia analítica se atacó la así llamada 'tesis de Brentano', según la cual la intencionalidad es la marca exclusiva de lo mental. Lo anterior fue particularmente útil para justificar mi modo de abordar la intencionalidad, modo que consistió en preguntarse por las clases de intencionalidad que pueden ser atribuidas a cierto tipo de fenómenos y de sistemas. Por otro lado, 2) también aclaré la expresión 'sistema artificial' al vincularla con la rama de la Inteligencia Artificial que se enfoca en diseñar sistemas que sean capaces de realizar operaciones computacionales simbólicas. Para ello, brindé una breve descripción de la Inteligencia Artificial Fuerte, la expresión acuñada por Searle, y de las teorías GOFAI expuestas en su momento por Haugeland. Esto dio lugar, a su

vez, para la pregunta acerca de la intencionalidad de tales sistemas artificiales, que abordé en el segundo capítulo de la monografía.

En el segundo capítulo, por su parte, me enfoqué en la pregunta acerca de la intencionalidad de los sistemas artificiales de Inteligencia Artificial Fuerte y de GOFAI. En primer lugar, expliqué el sentido en el que los estados mentales intencionales representan sus objetos y estados de cosas, que no es otro que el sentido en el que los actos de habla lo hacen: los estados intencionales representan sus condiciones de satisfacción, que son determinadas por el contenido proposicional bajo una cierta dirección de adecuación, la cual, a su vez, es determinada por el modo psicológico. Al considerar las conexiones entre los actos de habla y los estados intencionales, fue necesario considerar también las posibles diferencias: si bien los estados intencionales representan sus objetos en el mismo sentido que los actos de habla, no lo hacen de la misma manera ni por el mismo medio. El medio es diferente pues los actos de habla requieren cierta materialización física que los estados intencionales no necesitan. Y la manera es diferente pues la intencionalidad de los actos de habla se deriva de la intencionalidad de los estados mentales. Esto me llevó a presentar una definición de las clases intencionales propuestas por Searle. Pero esto no fue suficiente para aclarar la clase intencional que le corresponde a los sistemas artificiales de la Inteligencia Artificial Fuerte y de GOFAI, por lo que fue preciso atender al experimento mental de la Habitación China. En concreto, el experimento mental permitió entender que, pese a que los sistemas artificiales representan objetos y estados de cosas, sólo lo hacen en la medida en que hay personas que los programan, que introducen entradas y que interpretan las salidas. Son las personas las que pueden representar sus estados mentales; las representaciones de un sistema artificial no son, en efecto, diferentes a las representaciones de un libro o de un mapa.

Concluido lo anterior ¿qué posibles caminos se han abierto? El más evidente, sin duda, es el que refiere a la intencionalidad. Por un lado, la clasificación de la intencionalidad de Searle, su manera tripartita de entenderla (intrínseca, derivada y como-si), depende de su distinción ontológica entre aquello que es relativo a nosotros y lo que es independiente. Pero ¿hasta qué punto y en qué sentido se está dispuesto a aceptar tal distinción? Por supuesto, y por fortuna,

la suya está lejos de ser la única clasificación de la intencionalidad posible: un ejemplo concreto, y no tan famoso, es la clasificación de la intencionalidad de Haugeland, también tripartita. Haugeland[76] distinguió entre 'intencionalidad auténtica', 'intencionalidad ordinaria' y lo que él llamó 'intencionalidad *ersatz*', que sería, a su juicio y en contraste con Searle, la intencionalidad de los sistemas GOFAI y de Inteligencia Artificial Fuerte. Estas clases intencionales de Haugeland difícilmente encuentran un correlato en las clases intencionales de Searle, de modo que un trabajo comparativo al respecto es un posible camino a seguir. Por otro lado, es claro que la inteligencia artificial simbólica no es el único programa investigativo de la Inteligencia Artificial. ¿Qué se puede decir, por ejemplo, acerca de los modelos más recientes, como el conexionista o el de redes neuronales? ¿Cuál es la clase de intencionalidad que les corresponde? ¿Las representaciones de sus sistemas artificiales también dependen de nuestras representaciones? Preguntas de este tipo involucran a disciplinas tan diversas como aquellas que forman parte de las Ciencias cognitivas. Y, además de ellas, por ejemplo, al preguntarse por la intencionalidad de los animales no-humanos se involucran también disciplinas como la biología cognitiva, la etología, la psicología comparada y la filosofía de la biología. Resulta claro que un posible trabajo interdisciplinar queda abierto.

En filosofía hay, no obstante, un tema que especialmente me interesa. En la segunda parte del segundo capítulo se vio que el experimento mental ideado por Searle, la Habitación China, le permitió mostrar que la tesis computacionalista –en la versión mantenida por la Inteligencia Artificial Fuerte– parece ser falsa. Y si no falsa, al menos sí que tiene problemas teóricos o conceptuales por considerar. ¿De qué modo, pregunto, un experimento mental, es decir, un experimento que se da en la imaginación, puede negar tesis en el mundo fáctico? ¿Los experimentos mentales, tan comunes en filosofía y en ciencia, pueden ser usados como un recurso argumentativo fuerte o son simples "bombas de intuición" a las que no hay que tomar muy en serio? La pregunta por la validez científica de los experimentos mentales es, sin duda, un camino posible y tal vez incluso sea uno necesario. Por cierto, la apreciación que tengamos al respecto modificará sustancialmente nuestra forma de ver y de hacer la filosofía.

[76] HAUGELAND. Having Thought... Óp. cit.

BIBLIOGRAFÍA

AUSTIN, John. How to do things with words. London: Oxford University Press. 1962.

BRENTANO, Franz. Psychology from an Empirical Standpoint. London: Routledge. 2009.

CHISHOLM, Roderick. Sentences about Believing. En: Minnesota Studies in the Philosophy of Science, núm. 1 (1957). pp. 510-520.

CORNMAN, James. Intentionality and Intensionality. En: Philosophical Quarterly, XII (1962). pp. 44-52.

CRANE, Tim. Intentionality as the mark of the mental. En: O'Hear, A. (Ed.). Contemporary Issue in the Philosophy of Mind. Cambridge: Cambridge University Pres, 1998.

_____The Objects of Thought. Oxford: Oxford University Press. 2013

DENNETT, Daniel. The Intentional Stance. Cambridge: MIT Press. 1987.

DRETSKE, Fred. Knowledge and the Flow of Information. Cambridge: MIT Press. 1981.

HAUGELAND, John. Having Thought. Cambridge: Harvard University Press. 1998.

_____The Artificial Intelligence: The Very Idea. Cambridge: The MIT Press. 1989.

HOBBES, Thomas. Tratado sobre el cuerpo. Madrid: Universidad Nacional de Educación a Distancia. 2009.

KNALE,William. Intentionality and Intensionality. En: Aristotelian Society, LXII (1968).

MALLY, Ernst. Gegenstandtheoretische Grundlagen der Logik und Logistik. Leipzig: Barth. 1912.

MCGINN, Colin. The Character of Mind. New York: Oxford University Press. 1982.

MEINONG, Alexius. The theory of objects. En: CHISHOLM, Roderick (Ed.). Realism and the Background of Phenomenology. Glencoe: The Free Press, 1960.

PARSONS, Terence. Nonexistent Objects. New Haven: Yale University Press. 1980.

PUTNAM, Hilary. The nature of mental states. Cambridge: Cambridge University Press. 1975.

RUSSELL, Bertrand. The philosophy of logical atomism. En: The Monist. (1918) pp. 177-281.

SEARLE, John. Actos de habla. España: Ediciones Cátedra. 1994.

_____ Expression and Meaning. Cambridge: Cambridge University Press. 1979.

_____ Intencionalidad. Madrid: Editorial Tecnos. 1992.

_____ Intentionality. Cambridge: Cambridge University Press. 1983.

_____ Mentes, cerebros y ciencia. España: Ediciones Cátedra. 1985.

_____ Mind: A brief introduction. New York: Oxford University Press. pp. 66-72.

_____ Minds, Brains, and Programs. En: Behavioral and Brains Science, núm. 3 (1980). pp. 417-457.

_____ Mind, Language and Society. New York: Masterminds. 1999.

STRAWSON, Galen. Mental reality. Cambridge: MIT Press. 1994.

TURING, Alan. Computing Machinery and Intelligence. En: Mind, 49 (1950). 433-460.

URMSON, James. Criteria of Intensionality. En: Aristotelian Society, LXII (1968).

ZALTA, Edward. Intensional Logic and the Metaphysics of Intentionality. Cambridge: MIT. 1988.