Universidad Autónoma de Querétaro
Facultad de Filosofía

Sistemas de símbolos en inteligencia artificial

Tesis
Que como parte de los requisitos para obtener el grado de

Maestro en
Filosofía

Presenta
José Eduardo García Mendiola

Querétaro, Qro.
Abril, 2008

Universidad Autónoma de Querétaro
Facultad de Filosofía
Maestría en Filosofía

SISTEMAS DE SÍMBOLOS EN INTELIGENCIA ARTIFICIAL

TESIS

Que como parte de los requisitos para obtener el grado de
Maestro en Filosofía

Presenta:
José Eduardo García Mendiola

Dirigido por:
Dr. José Luis González Carbajal

SINODALES

Dr. José Luis González Carbajal
Presidente — Firma

Dr. Fernando González Vega
Secretario — Firma

Mtro. Gabriel Corral Basurto
Vocal — Firma

Dr. Bernardo Romero Vázquez
Suplente — Firma

Mtro. Gonzalo Guajardo González
Suplente — Firma

Firma
Director de la Facultad
Mtro. Gabriel Corral Basurto

Firma
Director de Investigación y Posgrado
Dr. Luis Gerardo Hernández Sandoval

Centro Universitario
Querétaro, Qro.
Abril, 2008
México

RESUMEN

El propósito de este trabajo es mostrar la importancia de los sistemas de símbolos dentro de las investigaciones en Inteligencia Artificial (IA). Al mismo tiempo, sostener la tesis de que, si bien tales sistemas son necesarios para reproducir una actividad cercana a la inteligencia humana, por medios computacionales, éstos no son suficientes para tales efectos; no para ir más allá de imitaciones en gran medida convincentes. El enfoque sintáctico-semántico de las estructuras generadoras de actividad inteligente es el núcleo de la investigación. La información y los puntos de vista que se ofrecen giran en torno a una de las escuelas de IA, la corriente clásica, que visualiza a la mente como una estructura esencialmente simbólica, basada en la lógica simbólica, especialmente en la lógica de predicados. La metodología se finca en este enfoque simbólico que parte de la premisa, fundamental para esta tesis, de la hipótesis de los sistemas de símbolos físicos, la cual establece que un sistema de símbolos físicos (tal como una computadora digital, por ejemplo) posee los medios necesarios y suficientes para la actividad inteligente. Se consideran los métodos heurísticos aplicados a los sistemas de inteligencia artificial como aproximación a la capacidad del ser humano para resolver problemas. Tratando de sistemas de símbolos, se han considerado algunas conceptualizaciones filosóficas y antropológicas acerca del carácter simbólico del ser humano. Se han considerado nociones fenomenológicas básicas sólo a nivel de guía fáctica en cuanto a producir sistemas que actúen al modo como lo haría la conciencia inteligente. Sin embargo, la perspectiva fenomenológica proporciona una idea de la complejidad de la conciencia y, por ende, de la inteligencia. Se trata también la representación de la mente/cerebro como un modelo constituido por varios niveles de formalización; la importancia de cuestiones metamatemáticas tales como la recursividad y la algoritmicidad; así como los teoremas limitativos fundamentales a la Inteligencia Artificial. Se concluye que no parece plausible que algún sistema de símbolos, sobre algún sustrato físico, agote en su dinámica una actividad equivalente a la inteligencia. Es factible decir, en cambio, que tales sistemas de símbolos físicos participan en la exhibición de tal actividad.

(**Palabras clave**: símbolo, inteligencia, computación, algoritmicidad)

SUMMARY

The intention of this work is to show the importance of the systems of symbols inside the investigations in Artificial Intelligence (AI). At the same time, to hold the thesis of which, though such systems are necessary to reproduce an activity near to the human intelligence, by computational means, these are not sufficient for such effects; not to go beyond imitations to a great extent convincing. The syntactic - semantic approach of the generating structures of intelligent activity is the core of this work. The information and the points of view concern one of AI's schools, the classic current, which it visualizes to the mind as an essentially symbolic structure, based on the symbolic logic, specially in the logic of predicates. The methodology is based on this symbolic approach arising from the premise, fundamental for this thesis, of the physical symbols systems hypothesis (PSSH), which establishes that a system of physical symbols possesses the necessary and sufficient means for the intelligent activity. The heuristic methods applied to the systems of artificial intelligence have been considered as an approximation to the aptitude of the human being to solve problems. Related to symbols systems, some philosophical and anthropological conceptualizations have been considered over the symbolic character of the human being. Some basic notions of phenomenology have been considered only as a guide as for producing systems that act to the way like it intelligent conscience would do. Nevertheless, the phenomenological perspective provides an idea of the complexity of the conscience and, therefore, of the intelligence. This work treats also the representation of the mind / brain as a model constituted by several levels of formalization; the importance of metamathematical questions such as recursivity and algorithmicity; as well as the limitative fundamental theorems to the Artificial Intelligence. One concludes that it does not seem to be commendable that any system of symbols, on any physical substratum, exhausts in its dynamics an activity equivalent to the intelligence. It is feasible to say, on the other hand, that such systems of physical symbols take part in the exhibition of intelligence.

(**Key words**: symbol, intelligence, computation, algorithmicity)

A mis padres.

Que aunque insuficiente, necesaria para su honra.

AGRADECIMIENTOS

Para la elaboración de este trabajo se ha contado con el apoyo y buen consejo del personal académico de la Maestría en Filosofía. Un agradecimiento especial para quienes revisaron esta tesis.

I N D I C E

INDICE DE FIGURAS

I. INTRODUCCIÓN

El propósito de este trabajo es mostrar la importancia de los sistemas de símbolos dentro de las investigaciones en Inteligencia Artificial (IA). Al mismo tiempo, sostener la tesis de que, si bien tales sistemas son necesarios para reproducir una actividad cercana a la inteligencia humana, por medios computacionales, éstos no son suficientes para tales efectos; no para ir más allá de imitaciones en gran medida convincentes.

En el capítulo II se exponen los fundamentos de los sistemas de símbolos físicos, como estructuras generadoras de actividad inteligente. El enfoque sintáctico-semántico de estas estructuras se muestra como el núcleo de la investigación, no sólo por parte de la Inteligencia Artificial, sino también de la Psicología Cognitiva.

Dado que tratamos de sistemas de símbolos, es oportuno considerar algunas conceptualizaciones filosóficas y antropológicas acerca del carácter simbólico del ser humano. Pero es importante mostrar también, en una aproximación analógica, las relaciones que pueden establecerse, sobre este punto, entre el hombre y la computadora y, por ende, con la tecnología. Ésta es la materia esbozada en el capítulo III. La filosofía y la ciencia cognitiva han mantenido, desde los albores de la IA, una relación muy estrecha. Esto se debe principalmente a la metodología conceptual que ambas perspectivas comparten. Las consideraciones sobre la primacía entre la filosofía y la cibernética son, en este trabajo, un pretexto para mostrar dicha relación así como la importancia de mantener, en toda investigación en el ámbito de IA, una actitud tanto prudente como motivante. La especulación filosófica en conjunción con la aplicación técnica es producto de un mismo foco de creatividad. No es dañino levantar periódicamente la mirada y abrirla al horizonte, mientras la investigación se concentra en un tramo del camino. Bajo este ambiente se muestra el enfoque fenomenológico a las investigaciones de la ciencia cognitiva en general, así como de la IA. Desde un punto de vista particular, las aportaciones de la descripción fenomenológica permitirían a la investigación, especialmente en estos temas, contextualizarse y evaluarse a sí misma.

La información y los puntos de vista que se ofrecen giran en torno a una de las escuelas de IA, la corriente clásica, basada en la lógica simbólica, especialmente en la lógica de predicados. De ahí la importancia de implementar lenguajes computacionales aptos para traducir, a un lenguaje interpretable por una computadora, las proposiciones más simples de esta lógica. La inteligencia como función simbólica se presenta en el capítulo IV.

Dentro de la función operativa de la inteligencia (capítulo V), la capacidad del intelecto humano para resolver problemas se manifiesta básicamente en la actividad encaminada a buscar soluciones en el menor tiempo y con la mayor precisión posibles. Los métodos heurísticos aplicados a los sistemas de inteligencia artificial pretenden aproximarse a esta capacidad de varias maneras, que van desde búsquedas al azar, hasta búsquedas en profundidad, pasando por métodos recursivos.

En este mismo capítulo se muestran las limitaciones así como las aproximaciones hacia el razonamiento humano de la lógica de predicados. Se muestran algunos ejemplos de traducción de proposiciones simples a un lenguaje simbólico que, en su momento, sirve de fundamento a los lenguajes computacionales aplicados a la IA. Además, algunas aplicaciones concretas de sistemas de símbolos se ejemplifican con mecanismos de aprendizaje y razonamiento, así como de planeación. La problemática involucrada en la manipulación de símbolos, aunada a la generación de algoritmos de razonamiento y búsqueda de soluciones se concretiza hasta cierto nivel técnico. En este nivel es posible darse una idea general de las implicaciones que, en lo concreto, tiene el trabajo teórico de las ciencias del conocimiento, la filosofía de la IA y la lógica-matemática. Las implicaciones de orden fisiológico y neurolingüístico están fuera del ámbito de este trabajo, y quedan reservadas para la corriente conexionista, que no es la clásica en IA.

La metáfora ha dejado de ser un mero utensilio de comunicación con vistas a una mayor claridad y una mejor comprensión de alguna temática. Actualmente, uno de los enfoques fundamentales a los sistemas de aprendizaje y razonamiento tienen sus bases en la analogía. Esta aplicación testifica, de algún modo, que nuestra comprensión del mundo –incluso el de la Ciencia- y de la conciencia humana no se ajusta a estructuras demasiado rígidas. Las

aportaciones de la filosofía de la IA y de la fenomenología a la cibernética y a las ciencias del conocimiento, en lo que a la inteligencia se refiere, proveen algunas dificultades en torno a la suficiencia de los sistemas de símbolos físicos como mecanismos generadores de inteligencia, en base a una estructura que, desde un punto de vista particular, tiene su columna vertebral en la analogía. Una perspectiva analógica aplicada a los sistemas de símbolos físicos merecería investigaciones posteriores; no obstante, se muestra el enfoque conexionista y neurocientífico en sus perspectivas, en todo caso, complementarias al enfoque clásico de la IA.

Sin embargo, para el propósito de este trabajo, los términos bajo los cuales se ha desarrollado la tesis corresponden al enfoque proposicional-discreto de la corriente clásica de la IA.

Nociones fenomenológicas tales como horizonte de significatividad, intencionalidad de la conciencia o inmediatez de los fenómenos han sido considerados, dentro de la investigación en IA, sólo a nivel de guía fáctica – aunque subjetiva - en cuanto a producir sistemas que actúen al modo como lo haría la conciencia inteligente. Sin embargo, la perspectiva fenomenológica proporciona una idea de la complejidad de la conciencia y, por ende, de la inteligencia. Esto nos brinda la oportunidad de valorar los métodos conceptuales de la filosofía que, en general, constituyen la base sobre la que descansa todo el trabajo realizado hasta hoy en cuanto a la IA.

Finalmente, en el capítulo VI se trata de establecer los principales puntos de discusión en torno a la metáfora computacional-simbólica: la representación de la mente/cerebro como un modelo constituido por varios niveles de formalización; la importancia de cuestiones metamatemáticas tales como la recursividad y la algoritmicidad; y se analizan las principales limitaciones al modelo clásico en base a los teoremas limitativos fundamentales y, consecuentemente, a la Inteligencia Artificial.

II. ESTRUCTURA METODOLÓGICA Y MARCO TEÓRICO

Antecedentes de la Inteligencia Artificial.

La Inteligencia Artificial puede definirse como la parte de la ciencia de la computación que trata sobre el diseño de sistemas de computación inteligentes, es decir, sistemas que exhiban características asociadas con el comportamiento humano inteligente, tales como entendimiento, aprendizaje, razonamiento, lenguaje o resolución de problemas.

Esta caracterización, así como su futuro tecnológico, son producto de las décadas recientes y están estrechamente vinculados con el nacimiento y desarrollo de la computadora digital. Pero la gama de aspiraciones y perspectivas de la Inteligencia Artificial (IA) remontan su origen a tiempos muy anteriores. La IA ha perseguido siempre el entendimiento de la mente a través de la explicitación y construcción de sus mecanismos subyacentes. Y esta idea proviene de la separación cartesiana entre mente y materia. Descartes estudió el sistema nervioso y propuso una teoría de la actividad nerviosa basada en los principios de la hidráulica, en base al supuesto de que todo cuerpo, sea humano o animal, no sería distinto de una máquina cuidadosamente construida que exhibiera cierto nivel de autonomía, tal como un mecanismo de relojería.[1] De acuerdo con Descartes, la vida estaría completamente vinculada con el funcionamiento del cuerpo físico (*res extensa*) – de la máquina - mientras que, por otra parte, la máquina no tendría relación alguna con lo mental (*res cogitans*). La actividad mental no requeriría corporeizarse y, además, ningún tipo de funcionalidad física, por muy compleja que fuera, sería suficiente para dar cuenta de la existencia de lo mental. En este tenor cartesiano, la IA se ceñiría a una investigación en términos puramente de comportamientos y de las capacidades que los generan. Cualquier posibilidad de construir una mente real, no meramente artificial, sería una quimera. Es en base a esta dualidad

[1] René Descartes, *The Philosophical Writings of René Descartes*, trans. J. Cottingham, R. Stoothof, and D. Murdoch, Vol. 3 (Cambridge, UK: Cambridge UP, 1991) 214, citado por Güven Güzeldere y Stefano Franchi: *SEHR, volume 4, issue 2: Constructions of the Mind.*

mente-cuerpo que, como señala Allen Newell, se pone en tela de juicio el éxito que pudiera realmente lograr la IA.[2]

Independientemente de la concepción de lo que se llama mente, la idea de construir mecanismos que exhibieran algún tipo de comportamiento que emulara algunas actividades específicamente humanas – hablar, cantar, escribir, jugar ajedrez – ha ocupado la imaginación desde hace muchos años, especialmente a partir del siglo XVIII. Sin embargo, el acceso a los elementos físicos y conceptuales como los que hoy utiliza la investigación en IA ha sido una posibilidad exclusiva de los diseñadores y constructores desde hace unas cuantas décadas atrás. Las computadoras digitales han favorecido las tareas de simulación metafóricamente basadas en el procesamiento en paralelo. El bagaje intelectual y tecnológico marca una notable diferencia entre el pasado y el presente del paradigma de la IA; entre quienes han investigado cuestiones tales como la mente, los mecanismos y las relaciones entre ellos. Actualmente, le investigación en IA es abanderada únicamente por los expertos del mundo de las computadoras (*hardware* y *software*), dentro de un marco tecnológico que resulta inaccesible para el lego.

Sin embargo, históricamente la situación ha sido diferente. Quienes han concebido la mente como una máquina hidráulica no se limitaron a desarrollar teorías con la colaboración exclusiva de ingenieros expertos en dinámica de fluidos. La empresa de la IA actual bien podría ser planteada en términos de contribuciones de agentes especializados en áreas diversas – no únicamente de las ciencias de la computación – tales como filósofos, artistas, sociólogos, antropólogos, etc. La posibilidad se antoja altamente probable dado que la IA actual tiene que hacer frente a múltiples controversias generadas por el trato con nociones básicas tales como símbolo, conciencia, proceso o comprensión, entre otros, y con metodologías surgidas de las neurociencias y de la física moderna en general. Es muy posible que los conflictos de interpretaciones y conceptualizaciones fundamentales para la investigación en las cuestiones de la mente hayan surgido por la excesiva homogeneización y la restrictiva composición profesional de la comunidad de investigadores en IA.

[2] Allen Newell, "Intellectual Issues in the History of Artificial Intelligence," *The Study of Information: Interdisciplinary Messages*, ed. Fritz Machlup y Uma Mansfield (New York: Wiley, 1983) 4.

Hacia 1960 era común, dentro de la comunidad de investigadores en IA, pensar que dentro de 10 años las computadoras serían tan listas como los humanos. Se creía incluso un deber de los científicos el proclamar tal predicción con el fin, se pensaba, de preparar al público para afrontar una inminente situación, y para evitar un posible "*shock*" psicológico. Pero, más allá de este deber social, resulta muy relevante el hecho mismo de la creencia, por parte de los científicos, en la inminente capacidad de las próximas computadoras; el hecho de la plausibilidad de la visión a corto plazo de la IA, que entonces se tenía.

Es muy probable que tal visión – que a la fecha no se ha hecho realidad – se debiera a la carencia de suficientes análisis históricos, filosóficos y sociológicos aplicados a la realidad del momento desorbitadamente optimista que se vivía al interior de la comunidad investigadora en IA. Las humanidades y las ciencias sociales no fueron parte, dentro de las primeras generaciones de la IA, de las áreas profesionales consideradas exclusivamente básicas, tales como las ciencias computacionales, la lógica y las matemáticas. Siempre se consideró a la IA como una empresa totalmente ingenieril o, cuando mucho, una empresa que no requería de ningunas disciplina, salvo las ciencias naturales, para colaborar con la ingeniería. Se consideraba que las teorías provenientes de la filosofía, la lingüística, la psicología y de disciplinas relacionadas eran insuficientemente completas y precisas para ser directamente implementadas en la investigación en IA. Aun concediendo tal imprecisión e incompletud de estas teorías, no se explica su exclusión del campo de la IA y, más aún, es muy probable que las teorías provenientes de las disciplinas consideradas básicas de la IA estén intentando modelar de una manera ásperamente precisa y estricta un complejo fenómeno tal como la inteligencia.

IA y humanidades.

John McCarthy ha señalado que la IA no debiera alejarse de la filosofía, si no quiere terminar haciendo una mala filosofía en lugar de ninguna. En su

artículo "What has AI in Common with Philosophy?"[3], el autor muestra esta idea: “la IA necesita muchas ideas que han sido estudiadas sólo por filósofos. Esto es porque un robot, si se pretende que tenga un nivel de inteligencia como la humana y que sea capaz de aprender desde su experiencia, necesita una visión general del mundo en el cual pueda reconocer hechos.”

Otros autores han apuntado la necesidad de ampliar la composición profesional de la IA, así como de reexaminar los presupuestos fundamentales acerca de la naturaleza humana. Así, en 1980, Phil Hayes, cuando dirigía un proyecto sobre la comprensión del lenguaje natural, en Carnegie-Mellon University, establecía que “… una investigador en IA debe aprender a aplicar las perspectivas de otras áreas a su propio trabajo de construcción de sistemas inteligentes. Inversamente, las otras áreas tienen la oportunidad de poner a prueba computacionalmente sus especulaciones básicas a partir de las cuales elaboran sus teorías, aprovechando la IA.[4]

El establecimiento de puentes entre la IA y otras áreas – música, filosofía, historia y estudios sociales – requiere esfuerzos de índole intelectual, académico, personal y político. Sin embargo, como afirman Güven Güzeldere & Stefano Franchi, aunque arduo, este camino es el único que realmente conduciría al Santo Grial de la IA.[5]

Proyección de la IA.

La gran meta de la IA ha sido referida al llamado “sueño de Turing”, según se constata en el famoso artículo (escrito por Alan Turing) "*Computing Machinery and Intelligence*": construir un mecanismo digital que pueda realizar tareas que, a la vista de cualquiera, requieran ciertas cualidades específicas de la mente humana: inteligencia, flexibilidad, comunicabilidad, etc. Este sueño no tiene nada que ver con la construcción de aplicaciones computacionales que funjan como asistentes en tareas de búsqueda mediante bases de datos,

[3] John McCarthy, **http://www-formal.stanford.edu/jmc/**, 1996
[4] *Sigart Newsletter*, special issue on Knowledge Representation, guest ed. Ronald J. Brachman and Brian C. Smith, 70 (February, 1980) 109.
[5] Güven Güzeldere & Stefano Franchi**,** *mindless mechanisms, mindful constructions*, *SEHR, volume 4, issue 2: Constructions of the Mind,* p.

reservaciones de transportación aérea, y actividades de ese tipo, además de aquéllas que requieren la fuerza bruta de muchos cálculos numéricos que resultan muy laboriosos para los humanos. Es por esto que ya no llaman la atención mecanismos que juegan ajedrez (que puedan analizar más de 100,000 jugadas posibles por segundo).

A propósito del juego de ajedrez, hay un punto que es importante considerar: el hecho de que los mecanismos que realizan esta actividad lo hacen dentro de un ambiente cerrado, perfectamente definido por las reglas del juego. La máquina no necesita comprensión alguna del ámbito humano del juego; la máquina no se preocupa por evitar cometer algún "estúpido" error; tampoco se sentirá la reina del universo tras haber derrotado a un oponente incauto. Nada de estas cosas son importantes para la máquina.

Más allá de este tipo de sentimientos típicamente humanos, lo cierto es que, desde su formulación en 1950 por Turing, el propósito de construir un mecanismo capaz de sostener una conversación con un humano se ha convertido en una de las principales barreras a franquear para la IA. Esta es la llamada Prueba de Turing.[6] Lograr superar la prueba de Turing ha pasado a ser una meta que está más allá de la actual investigación en IA que, en su agenda actual, tiene proyectos bastante menos ambiciosos tales como el diseño y construcción de mecanismos de toma de decisiones, llamados "sistemas expertos", los cuales han mostrado ser bastante útiles para el trabajo de abogados o médicos que tienen que lidiar con grandes cantidades de datos para emitir un fallo o un diagnóstico suficientemente acertados. Otros proyectos que han ya dado resultados son los múltiples y variados tipos de robots que son ahora imprescindibles en las grandes líneas de producción industriales. Sin embargo, tales progresos se ubican en una categoría de "mecanismos asistenciales" que requieren de la guía y el control humanos. Este tipo de progresos en IA han sido resultado de la creciente interacción entre el humano y la máquina computarizada, y es un resultado que no estaba previsto en la historia de la IA.

El gran problema de la IA ha sido el encontrar algún paralelismo entre el micro-mundo cerrado y definido en el que se han venido desarrollando sus

[6] v. cap. IV – La función simbólica de la inteligencia

productos y, por otra parte, el macro-mundo abierto y no determinable humano. Este problema puede traducirse en la imposibilidad de establecer una generalización, desde un conjunto consistente de variables controlables o micro-mundos, hacia la comprensión y modelaje de un mundo en gran medida inconsistente de variables prácticamente irrestrictas.

Actualmente, el problema de la imposibilidad de generalización ha sido afrontado mediante la adición indiscriminada de más micro-mundos (como pretendiendo que las partes sumen al todo) dando por resultado una escenario de gran complejidad.

En resumen, los progresos de la IA no han sido los que se esperarían en vistas a la gran meta: la construcción de la mente. Ante este panorama, la falta de credibilidad en la IA proviene de un círculo de críticos que sostienen sus posturas desde diversas perspectivas; desde el énfasis en la carencia de corporeidad que le de al mecanismo computarizado un real "estar en el mundo" (Dreyfus)[7], hasta el énfasis en la carencia del contexto social que el lenguaje crea y aporta a la actividad mental (Winograd y Flores)[8]. Desde la perspectiva industrial también se hace patente la crítica: la IA se ha convertido en la disciplina de la computación que, para fines industriales y comerciales, es cada vez más eficaz y avanzada cuanto necesaria. Sin embargo, en vistas a sus metas originales, la IA se encuentra en una etapa bastante primitiva.

No obstante, existen actualmente esfuerzos serios por dirigir las investigaciones en IA hacia sus propósitos originales, motivados por el sueño de Turing. Tales investigaciones buscan ofrecer una concepción más amplia de la IA que aporte las bases suficientes para hacer efectiva la materialización tecnológica y, al mismo tiempo, se encamine a la comprensión de la mente humana. Es así que el "proyecto Cog"[9], encabezado por el roboticista del *MIT*, Rodney Brooks y un equipo interdisciplinario de investigadores, integra en sus investigaciones las consideraciones metodológicas y teóricas provenientes de las ciencias cognitivas, las teorías de la evolución, la neuropsicología, la etología y la filosofía.

---

[7] *Cfr.* Hubert Dreyfus, *What Computers Still Can't Do* (Cambridge, MA: MIT Press, 1992), ix

[8] *Cfr.* Terry Winograd y Fernando Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Reading, MA: Addison, 1986) 11-12

[9] *Cfr.* Daniel C. Dennet, "The Practical Requirements for Making a Conscious Robot", publicado en *Philosophical Transactions of the Royal Society*, A, 349, 1994, 133-46.

IA y filosofía.

No ha habido en general, a lo largo de la historia de la IA, mayor vinculación entre ésta y las humanidades con sus tradicionales investigaciones acerca de la mente y el lenguaje. Sin embargo, las fronteras entre ambas clases de disciplinas han tendido en la última década a diluirse. Así, Philip Agre, en un artículo publicado en la revista *Constructions of the Mind* establece que "la Inteligencia Artificial es Filosofía subyacente"[10], esto es, la IA sería vista como un esfuerzo por generar y desarrollar, con sus medios técnicos característicos, los sistemas filosóficos inherentes. La estrecha vinculación entre ambas disciplinas estriba en el origen filosófico de las dificultades fundamentales con que se topa la IA. Según Agre, la metodología formal de la IA haría explícitas tales dificultades, ocultas e inherentes a los sistemas filosóficos implícitos en las construcciones de la propia IA. (El paradigma tradicional de la IA, por ejemplo, habría estado implícitamente apoyado en el sistema causal cartesiano que presupone la separación mente-cuerpo). Se requeriría entonces una disciplina teórica, científica y tecnológica que permitiera una simbiosis entre la investigación en IA y el análisis humanístico-filosófico de los sistemas y las ideas subyacentes a los enfoques, metas y objetivos.

La posibilidad de colaboración entre ambas áreas, IA y humanidades, es también subrayada por Serge Sharoff, quien la establece en el marco de una confrontación entre la tradición fenomenológica – en filosofía -, y el trabajo de investigación clásico en IA.[11] La postura de Sharoff consiste en visualizar a la IA no como una serie de intentos de construir máquinas pensantes sino, más bien, como realizaciones computarizadas de cierto tipo de filosofía. Su interpretación de la IA y la ciencia cognitiva es la de una "filosofía estricta", que trae a recuerdo el proyecto husserliano fenomenológico, el esfuerzo por esclarecer la estructura de la conciencia y que pueda orientar el trato con el mundo y, más aún, el ser en el mundo. La postura de Sharoff apunta a la interpretación, desde el enfoque de la IA, de las nociones fenomenológicas

[10] Philip E. Agre, the soul gained and lost: artificial intelligence as a philosophical project, *SEHR, volume 4, issue 2: Constructions of the Mind*
[11] *v. infra*. Cap. III, obstáculos a una fenomenología de la computación.

básicas – intencionalidad, horizonte, tiempo conciente interno, etc. – de tal forma que puedan ser efectivamente explotados y aplicados por los programas de la IA. Este enfoque constituiría una posibilidad de ampliación, tanto en área como en profundidad, del alcance actual de la IA.

Al interior de la filosofía – y este es el núcleo de la presente investigación – se ha discutido acerca de la capacidad que pudiera tener una computadora para manifestar un comportamiento inteligente, sobre la base del análisis teórico de las relaciones humanas con el mundo circundante. A este respecto, la postura de Dreyfus representa una posición en contra de la posibilidad de construir un mecanismo con tal capacidad de comportamiento inteligente, sobre la base de un reduccionismo implícito en el enfoque operatorio-simbólico para dar cuenta del ambiente en el que se desenvolverían tales mecanismos.

Es el paradigma tradicional de la IA, el enfoque simbólico sustentado en la llamada metáfora de la computación, que visualiza a la mente como una estructura esencialmente simbólica, el que centra la atención de este trabajo. La metodología se finca en este enfoque simbólico que parte de la premisa, fundamental para esta tesis, de la hipótesis de los sistemas de símbolos físicos.

La hipótesis de los sistemas de símbolos físicos.

La hipótesis de los sistemas de símbolos físicos (*PSSH*, por sus siglas en inglés), formulada por primera vez por Newell y Simon, establece que un sistema de símbolos físicos (tal como una computadora digital, por ejemplo) posee los medios necesarios y suficientes para la actividad inteligente. Esta hipótesis implica que las computadoras, cuando son provistas con los programas de procesamiento simbólico adecuados, son capaces de operar inteligentemente. Esta hipótesis también implica que el comportamiento simbólico humano se da gracias a que el hombre posee las características de un sistema de símbolos físicos.

Newell y Simon han admitido que la hipótesis de los sistemas de símbolos físicos podría resultar falsa. El comportamiento inteligente no es tan fácil de producirse como para que cualquier sistema pueda exhibirlo de alguna manera. Hay opiniones que apuntan a la conclusión de que, sobre bases

filosóficas o científicas, tal hipótesis es falsa. Desde un punto de vista estrictamente científico es posible atacar o defender dicha hipótesis únicamente sobre evidencias empíricas.[12]

Los argumentos en contra de la *PSSH* versan sobre cuatro temas principalmente. Una primera temática apunta a la presuposición de que las computadoras únicamente pueden manipular símbolos significativos. La inteligencia, opinan algunos, requeriría algo más que la mera manipulación de símbolos formales; se requiere cierta conexión con el medio ambiente a través de la percepción y la acción que le den sentido a los símbolos y que, de alguna forma, puedan "arraigarse" en dicho ambiente. Tal conectividad sólo sería posible mediante una suerte de "corporización". Es decir, la inteligencia requiere un cuerpo físico que sienta, actúe y tenga experiencias. Algunos afirman incluso que la inteligencia humana, para ser exhibida, necesita de un cuerpo de tipo humano. En general, y más allá del "tipo de cuerpo" exigido, la argumentación se enfocaría a la necesidad de que los símbolos se arraiguen en el ambiente, cualquiera que éste sea, para que se dé un comportamiento inteligente. Tal ambiente podría ser del tipo físico, como el humano, o bien, simulado: un mundo artificial que contenga otro tipo de agentes distintos al humano.

Una segunda temática de argumentos en contra de la *PSSH* se sostiene en la suposición de que lo subyacente a toda actividad inteligente, especialmente la percepción, involucra procesos no-simbólicos, esto es, señales de tipo analógico.

Un tercer tipo de argumentación – relacionado con el segundo – se sostiene sobre la afirmación de que la computación, en su acepción común, no puede proporcionar un modelo adecuado de la inteligencia; el cerebro no es una computadora. La actividad inteligente requiere mecanismos de "tipo cerebral" (no computacional).

Finalmente, una cuarta temática se basa en la observación de que todo comportamiento calificado como inteligente es, en realidad, "mecánicamente absurdo". Así, los insectos y las plantas se desenvuelven eficientemente en

[12] Allen Newell and Herbert A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM*. vol. 19, No. 3, pp. 113-126, March, 1976.

ambientes complejos y sus respuestas a las diferentes situaciones reflejan un cierto tipo de inteligencia, aún cuando – por lo menos aparentemente – ellos no manejan símbolos.

Hasta hoy no ha sido posible mecanizar la inteligencia, por lo menos a nivel humano. Refiriéndonos al estado actual de la hipótesis de los sistemas de símbolos físicos, y de acuerdo a los cuatro tipos de argumentaciones que, en su contra, se han establecido, tenemos el siguiente panorama.

En primer lugar, respecto al tema de la manipulación de símbolos meramente formales (sin significado), es preciso aclarar que, de acuerdo a Newell y Simon, un sistema de símbolos físicos es una máquina que produce, al paso del tiempo, ciertos conjuntos de estructuras simbólicas. Es decir, el mundo en el que existe tal sistema es un mundo de objetos más amplio que el que se pudiera inferir a partir, únicamente, de las propias expresiones simbólicas. La PSSH establece que un sistema de símbolos físicos tiene la capacidad de designar, es decir, de que los símbolos se "apeguen" a los objetos. Se dice que una expresión simbólica designa un objeto si el sistema puede, ya sea, afectar al propio objeto al que se refiere, o bien si el sistema adquiere un comportamiento que depende de dicho objeto.

Un punto importante a establecer en esta temática es que no está clara la necesidad de la existencia del mencionado "apego" o "conexión" de los símbolos con los objetos para que pueda exhibirse algún grado de actividad inteligente.

En relación a que la actividad inteligente requiere una manipulación no-simbólica, se ha afirmado que la actividad humana inteligente, en mayor grado, está basada en nuestra capacidad de realizar juicios espontáneos mediante reconocimiento de ciertos patrones. Dado que no está al alcance el realizar los ejercicios de introspección suficientemente profundos y precisos para comprender situaciones tales como el reconocimiento de voz, de rostros familiares o de lenguaje corporal, así como de aprehensión sintética de un estado de jugadas en ajedrez y otras situaciones, resulta muy difícil diseñar reglas basadas en la manipulación simbólica que permitan "programar", en una computadora, tales tareas. Es por ello que, en relación a este tipo de actividades, se ha optado por el uso de métodos dinámicos, estadísticos y

basados en redes neuronales que implican un tratamiento analógico, no-simbólico, en lugar de uno simbólico-discreto.

El punto aquí es que el uso de procesos no-simbólicos no necesariamente reemplazaría sino, tal vez, complementaría, el procesamiento simbólico-discreto.

El tercer tema de argumentación puede expresarse de una sola mirada mediante la expresión: "el cerebro no es una computadora". Quienes afirman esto se basan en algunas distinciones entre el cerebro y una computadora tales como el procesamiento en paralelo, y no serial, humano, así como en la distinción entre programación y aprendizaje, además del grado de tolerancia a los fallos.

En este punto resulta crucial el esclarecimiento de las nociones de "cerebro" y "computadora". El concepto de computación no necesariamente ha de ser entendido como un tipo de procesamiento simbólico en un cierto nivel. También puede entenderse como un conjunto de operaciones recursivas que operen sobre estructuras de listas simbólicas. En forma alternativa, la computación puede ser entendida como el funcionamiento en paralelo de ciertas fuentes de conocimiento a partir de, - o transformando y produciendo - complejas expresiones simbólicas. El término "computación" constantemente se expande o modifica. Así pues, ¿qué clase de computadora no es el cerebro?

Por último, acerca de la argumentación basada en la idea de que la inteligencia es "absurda", es decir, que no se le puede dar una explicación con sentido, uno de los más claros ejemplos es el de Jordan Pollack[13], quien establece que la Inteligencia Artificial se ha fundado sobre un objetivo equivocado. Los intentos de crear inteligencia de nivel humano mediante la simulación de la arquitectura cognoscitiva, a menudo descubierta vía protocolos introspectivos, son prácticamente imposibles debido a las limitaciones de la ingeniería de software. Sin embargo, afirma Pollack, muchos procesos en la naturaleza son mucho más poderosos que el pensamiento humano simbólico. Estos sistemas exquisitamente iterativos, como la evolución y embriogénesis, no requieren lógica, gramática, ni otros atavíos de

[13] Jordan B. Pollack, "Mindless Intelligence," *IEEE Intelligent Systems*, p. 55, May/June 2006.

arquitectura antropomorfa cognoscitiva. Los procesos dinámicos son generados por cierta acumulación de datos obtenidos mediante procesos iterativos y aleatorios, los cuales resultarían más inteligentes que una mente humana brillante, además de que se trata de procesos que operan sobre principios que no están fincados en ningún tipo de razonamiento lógico simbólico. Pollack propone la investigación sobre la inteligencia tal como surge fuera de la mente humana y, sobre estos estudios, basar el futuro de la IA. No obstante, es evidente que tales procesos dinámicos no suelen probar teoremas, diseñar y ejecutar planes ni compilar historias – como lo hace la inteligencia humana. *La afirmación de Pollack* de que "la mayor parte de lo que nuestras mentes realizan involucran procesos químicos ininteligibles (absurdos)" *no ayuda a esclarecer el conocimiento inteligente en mayor medida en que se puede afirmar, paralelamente, que los procesos que realiza un sistema de reservaciones aéreas involucran corrientes electrónicas absurdas. O, en otros términos, los procesos entendidos en tanto se consideran como absurdos, únicamente generan comportamientos "igualmente" absurdos*.

A la luz de estas argumentaciones en contra de la hipótesis de los sistemas de símbolos físicos, se puede establecer que la importancia de la investigación acerca de los procesos simbólicos para generar inteligencia es muy alta. Incluso quienes estudian el cerebro humano y afirman que éste no es una computadora utilizan conceptos computacionales para explicar las funciones cerebrales.

Por otra parte, se podría establecer la necesidad de los sistemas no-simbólicos para entender la inteligencia, contra la afirmación de la suficiencia de la *PSSH* para este efecto. Es probable que la hipótesis no sea suficiente. Sin embargo, hasta ahora no hay evidencias en contra de su necesidad para el diseño, generación y explicación de la actividad inteligente.

III. INTELIGENCIA Y TECNOLOGÍA

La inteligencia como sistema operativo y simbólico.

Generalmente se piensa que la inteligencia artificial (IA) trata de la manera en que las computadoras pueden hacer ciertas cosas de las que la mente es capaz. Esas cosas son las que requieren inteligencia, tales como la demostración de teoremas, el proporcionar asesorías o el realizar diagnósticos en diversos temas. Se trata de actividades que se distinguen de otras que no implican control consciente alguno, tales como hablar la lengua materna, o que atienden simplemente al sentido común.

Esta forma de considerar a la inteligencia artificial presupone, sin embargo, que las computadoras tienen la capacidad para hacer las cosas que la mente puede hacer, es decir, asesorar, diagnosticar, inferir y comprender. Una definición tal se enfoca a los modos en que las computadoras podrían hacer esas cosas. En cambio, al definir a la IA como "la creación de computadoras cuyo desempeño observable tiene características que, en los seres humanos, atribuiríamos a procesos mentales", se evita aquella presuposición problemática, así como las comparaciones entre los modos de proceder de la computadora y de la mente.

Otra definición, más controvertida, considera a la IA como la "ciencia de la inteligencia en general". Su objetivo es sistematizar una teoría que pudiese explicar las categorías generales de la intencionalidad así como las diversas capacidades psicológicas que en ellas se fundamentan. Considerada de este modo, la IA abarcaría no solo la psicología terrestre sino toda la gama de mentes posibles, investigando las estructuras básicas en las que podría incorporarse la inteligencia. Esta definición habría de hacer patente también la importancia que las computadoras tienen para esta ciencia. La problemática filosófica surge del cuestionamiento en torno a las posibilidades de estas metas y del modo en que podrían lograrse, así como de la concepción misma de esta ciencia.

Considerada así la IA, como ciencia de la inteligencia en general, su filosofía tiene estrecha relación con las filosofías de la mente y del lenguaje y con la epistemología; con las ciencias cognitivas en general y, particularmente,

con la psicología de la computación. Esta última sostiene cuatro supuestos filosóficos:

a) Un enfoque funcionalista en el estudio de la mente y de la inteligencia, percibiendo los procesos mentales como estrictamente especificables y definiendo los estados mentales en relación causal con la información sensorial recibida, el comportamiento motor y otros estados mentales.
b) La psicología de la computación es el estudio de los procesos de cómputo mediante los cuales se construyen, interpretan y transforman las representaciones mentales.
c) El cerebro es un sistema de computación que discierne en cuanto al tipo de relaciones funcionales que incorpora, sin importar la fisiología de esta incorporación o la intervención celular.
d) Los conceptos de la IA deben ser parte del contenido esencial de la teoría psicológica.

La explicación de la inteligencia mediante conceptos significativamente similares a los de la IA ha engendrado teorías metafísicas, especificaciones formales y modelos exploratorios del proceso mental, como lo atestiguan los trabajos de Hobbes, Leibniz y Babbage. Los recursos intelectuales de la IA se han derivado de los avances en la teoría formal de la computación, el diseño de máquinas que aplican cómputos formalmente especificados y el descubrimiento de la neurona.

Aunque las investigaciones recurren en mayor medida a alguna de estas tendencias, actualmente se distinguen dos tipos de investigación en IA: la variante “tradicional” y el “conexionismo”. Ambas parten de un antecedente común que es el artículo escrito por el psiquiatra y neurólogo Warren McCulloch en coautoría con el matemático Walter Pitts, “Un cálculo lógico de las ideas inmanentes en la actividad nerviosa”.

En efecto, este “calculo lógico” influyó en Von Neumann para el diseño de su computadora digital y en la creación de modelos formales de pensamiento por parte de los pioneros de la IA; la exposición de la “actividad nerviosa” contribuyó a la teoría psicofisiológica de Hebb sobre agrupaciones

celulares y derivó modelos de redes neurales, primeros precursores de los sistemas conexionistas.

La determinación y diseño de redes neurales capaces de ejecutar ciertas actividades, en la construcción de modelos funcionales, es labor de la IA que, al enfocarse a redes reales y posibles, se particulariza en la psicología humana. Las "redes", concebidas como aproximaciones a conexiones neurales reales, dirigen una investigación conexionista. En cambio, interpretadas como idealizaciones de la actividad neural, enfocadas principalmente a la lógica binaria y no a la naturaleza conectiva real de la célula y del umbral celular, devienen en el proceso de información digital característico de la IA tradicional. Ambos tipos de investigación surgieron a partir del artículo de McCulloch y Pitts.

El artículo de Alan Turing, "La maquinaria de computación y la inteligencia" (1936) sobre números computables, en el que concebía la computación como el manejo formal de símbolos (no interpretados) mediante la aplicación de reglas formales, es el fundamento teórico de ambos puntos de vista en IA. Turing da origen a la noción de "procedimiento eficiente", un proceso de computación estrictamente definible que mostró con ejemplos de cálculo matemático. La explicación de la inteligencia mediante este tipo de procedimientos implicaba que ésta podría simularse con la máquina universal de Turing o algún otro mecanismo semejante. Para 1950, Turing y otros ya habían construido computadoras digitales con las que podían simularse algunos aspectos de la inteligencia. Turing se preguntaba acerca de la posibilidad de que esas máquinas pudieran pensar.

Según este autor, la pregunta no estribaba fundamentalmente en una definición acerca del "pensar", sino en si una computadora podría participar en el "juego de las imitaciones". Esta pregunta, a su vez, constaría de tres aspectos: ¿Podría alguna computadora responder de la manera como lo haría un ser humano a un examinador? ¿Existen procedimientos eficientes, capaces de generar esa respuesta? ¿Bastaría ese resultado para poder atribuirle inteligencia a la computadora? Turing responde afirmativamente a cada cuestión.

Esta misma confianza la han compartido Allen Newell y Herbert Simon y han tratado de llevar esas posibilidades a la práctica, tanto en el análisis de

tareas abstractas como en observaciones experimentales detalladas. Ellos han evidenciado las implicaciones que tendría la IA para la filosofía de la mente: la mente es un sistema de cómputo, el cerebro ejecuta literalmente cómputos (suficientes para la inteligencia) e idénticos a los que podrían realizar las computadoras. La inteligencia humana puede explicarse mediante procedimientos de entradas y salidas de datos que controlan el comportamiento y el procesamiento de información interna. Las computadoras también pueden ser inteligentes, ya que, como el cerebro, es un sistema de símbolos físicos, el cual proporciona los medios necesarios y suficientes para realizar actividades de inteligencia general.

Newell y Simon vinculan una teoría sintáctico-formal del simbolismo con una teoría causal de la semántica. Según ellos, la identidad de un símbolo o de un cómputo es de carácter puramente formal y su importancia radica en su historia causal y sus efectos. Los símbolos son modelos físicos que se relacionan entre sí por diversas formas (como la yuxtaposición, por ejemplo) construyendo "expresiones" compuestas. Los procesos de cómputo comparan y modifican modelos, en los que la entrada y salida de datos son modos de expresión. Tales procesos se efectúan por medios físicos. La capacidad para manejar símbolos, en cualquier sustrato físico, radica en la de almacenar y transformar sistemáticamente los modos de expresión. Así, la comprensión de la inteligencia se basa en la descripción de sistemas de símbolos físicos, en términos de designación e interpretación, en el nivel de procesamiento de información. Atendiendo a una definición causal de estos términos semánticos, el conjunto de cambios que un símbolo permite al sistema realizar, en respuesta a un determinado estadio –interno o externo- constituye la noción de significado. Las dependencias causales son, en sentido estricto, arbitrarias ya que cualquier símbolo (simple) puede designar lo que sea. Caso aparte son las representaciones analógicas, en las que existe semejanza significativa entre el representante y lo representado.

Procesos formales y pensamiento.

El esclarecimiento de los múltiples procesos que constituyen lo que se denomina pensamiento es uno de los objetivos fundamentales de la

Inteligencia Artificial. En la complejidad del pensamiento pueden identificarse procesos que van desde la asociación de imágenes e ideas que están íntimamente vinculados con aspectos psicológicos en general pero también dependientes de particularidades propias de la persona que realiza tales procesos, hasta procesos que pueden ser formalizados y que constituyen el ámbito del razonamiento.

El análisis de lo que sucede cuando la mente discrimina entre una multitud de alternativas, eligiendo la opción más adecuada o decidiendo la que tenga un mayor sentido para ser aplicada en una situación compleja, es una tarea básica para las ciencias cognitivas, en general, incluyendo a la Inteligencia Artificial. La trivialidad aparente en la toma de decisiones en circunstancias de la vida diaria se convierte en una complejidad que se hace patente a la mirada introspectiva. En tales circunstancias, la visualización de los mecanismos de decisión mediante procesos deductivos resulta inadecuada ya que, aunque pudieran formularse hasta su agotamiento todas las posibilidades de decisión, la opción elegida se establece en base a una cierta pertinencia. Es decir, la enumeración de las múltiples posibilidades que se abren al paso de las diversas circunstancias de la vida real no aporta, por sí misma, dicha pertinencia. Son muchas las cosas que hay que considerar - aparentemente en forma simultánea - para que el solo razonamiento sea suficiente. Virtudes asociadas a los sentidos de simplicidad, buen juicio y belleza son factores cruciales en la toma de la decisión más trivial y cotidiana.

La caracterización de la mente/cerebro como una estructura multinivel - al modo análogo en que se caracteriza el ámbito del *software* en una computadora - es un modelo que busca conectar los niveles de mera manipulación formal de símbolos - como el lenguaje en su aspecto sintáctico - con los niveles inferiores asociados con la naturaleza “fija” de la materia cerebral o lo que sería el *hardware* de las computadoras. Esta conexión entre niveles tendría lugar a través de niveles intermedios donde tendrían lugar los procesos de interpretación y asociación de símbolos.

En este sentido, la caracterización del pensamiento como una suerte de procesos formales es una premisa fundamental para el propósito de modelar las funciones mentales/cerebrales, de tal manera que pudieran ser reproducidas en un sistema computacional. Ejerce una especial atracción a los

investigadores de las ciencias cognitivas el poder modelar los procesos del pensamiento al modo como se diseñan las estructuras matemáticas. Citando a Stanislaw Ulam[14] , este modelo trataría con entidades del pensamiento tales como nociones, símbolos, clases de símbolos, clases de clases y así sucesivamente. Estas entidades funcionarían a través de procesos matemáticos recursivos que darían cuenta del camino recorrido casi instantáneamente que va desde la excitación neuronal hasta el manejo simbólico mediante patrones iterativos de formación. La herramienta computacional - sugiere Ulam - apoyaría en gran medida la descripción de procesos tales como, por ejemplo, la "libre" asociación de ideas. (El entrecomillado indica que una formulación así tendría que hacer algunas precisiones acerca de una genuina libertad de asociación, dado que se haría patente una forma de necesidad - cuando menos bajo ciertas condiciones - en la formación de estructuras mentales elaboradas y en la toma de decisiones pertinentes.)

Dado que el lenguaje y el pensamiento estén sujetos a ciertas reglas formales, como se asumiría en el modelo multinivel, tales reglas sustentarían tanto los niveles superiores como los inferiores. De esta forma, los procesos mentales serían descritos, dentro del esquema computacional y como lo ha señalado Boden (v. *supra*), como procesos que se manejan desde un nivel de instrucciones - caracterizado por la flexibilidad que dan los diferentes lenguajes de programación - en el ámbito del *software*, hasta un nivel de operación o ejecución - caracterizado por la estructura rígida del sustrato físico - en el ámbito del *hardware*, análogo al nivel neuronal Este modelo sería una concatenación de niveles de reglas en el que el nivel de manejo formal de símbolos, independientemente de la arbitrariedad manifiesta en el nivel superior, tiene una conexión necesaria con el nivel inferior pasando por lo niveles intermedios en los que cada uno refleja su metanivel.

Ahora bien, la elaboración de este modelo requiere la difícil tarea de esclarecer lo que parece ser el hecho, como lo señala Hofstadter[15*] , de que cuando pensamos tenemos acceso a los propios pensamientos, de tal forma

[14] Stanislaw Ulam, *Adventures of a mathematician*, citado por D. Hofstadter, Gödel, Escher y Bach: una eterna trenza dorada, Conacyt, México, 1982, p. 661
[15] D. Hofstadter, *op. cit*. pp. 813-814

que podemos cambiarlos o modificar los estilos de pensamiento, si bien no tenemos acceso al nivel neuronal. Esto significa, en términos de este modelo mente/cerebro, que podemos modificar las "reglas *software*", aunque no podamos modificar las reglas hardware: las neuronas funcionan siempre de la misma manera. "¡La flexibilidad del *software* se debe a la rigidez del *hardware*!"

Siguiendo a Hofstadter, el aislamiento que tenemos respecto a los niveles inferiores produce la sensación de sentirnos "autoprogramados". Más aún, no podemos sentirnos de otra manera. Este hecho incide directamente en el modelo multinivel: el modelaje de nuestra propia estructura nos tendría que incluir dentro de él ¡con todo y nuestra autoprogramación! El modelo de nuestra propia estructura constituiría la esencia de la comprensión.

El proceso de la comprensión se presenta así como un proceso no sólo recursivo sino autoreferencial, de tal forma que decir la verdad sobre la comprensión objetivamente, como podría serlo mediante un modelo formalizado, tendría que sustentar dicho modelo sobre las bases mismas de la lógica o del razonamiento, que son las bases de la verdad y validez inferencial. Pero estas bases resultan ser la materia misma del problema que se intenta describir: la comprensión.

La ciencia cognitiva como filosofía.

A medida que la filosofía va desapareciendo, según Heidegger, la cibernética se convierte en la filosofía del siglo XX.[16] Sin embargo, esta tesis puede revertirse; esto es, los sistemas filosóficos pueden interpretarse desde el punto de vista de la ciencia de la computación. Las diferentes escuelas de la ciencia cognitiva, entonces, representan las interpretaciones o realizaciones de las correspondientes escuelas filosóficas.

La historia de la inteligencia artificial comienza con los esfuerzos para crear máquinas pensantes diseñadas para abarcar lo más ampliamente posible el dominio de la actividad intelectual humana, y el objetivo explícito de tal desarrollo fue ir más allá del razonamiento humano común, al menos en ciertas

[16] Heidegger, Martin, "'Only A God Can Save Us': Der Spiegel Interview with Martin Heidegger," The Heidegger Controversy: A Critical Reader, ed. Richard Wolin (Cambridge, MA: MIT Press, 1993).

áreas. La programación de sistemas sirvió como base de esas investigaciones. En términos prácticos, la falla relativa de esos intentos iniciales –la discrepancia entre sus intenciones y sus éxitos reales- ha propiciado la creación de programas efectivos que operen exitosamente dentro de dominios de problemas cuidadosamente delimitados. Conceptualmente, la meta actual de la investigación teórica en IA es investigar la mente humana, en donde se unen tanto la investigación psicológica como la programación pura.

Pero muchos problemas de IA tienen también una relación directa con los problemas de la antigua filosofía: determinación de las categorías mentales, problemas del círculo hermenéutico, el balance entre conocimiento empírico y *a priori*, las interrelaciones entre conocimiento abstracto y concreto, entre otros. Si consideramos las investigaciones teóricas en IA, nos daremos cuenta de que están siempre basadas en algún trasfondo filosófico; este marco ayuda en gran parte a determinar la estructura de los modelos de IA a medida que son elaborados.

Dreyfus y Dreyfus aportan un intento de interrelación entre la filosofía clásica y la ciencia cognitiva y una interpretación de la primera desde el punto de vista de la segunda. A medida que ellos describen la historia de la filosofía europea, identifican una secuencia de los predecesores de la IA: Platón, Galileo, Descartes, Leibniz, Kant y Husserl. Así por ejemplo, escriben:

Kant tenía una idea nueva acerca de cómo trabajaba la mente. Él sostenía que todos los conceptos eran en realidad reglas. Por ejemplo, el concepto de perro es algo así como esta regla: Si tiene cuatro patas, ladra y menea la cola, entonces es un perro… Husserl, que puede considerarse como el padre del modelo del proceso de información de la mente, [ampliando las ideas de Kant] argumentaba que los conceptos eran jerarquías de reglas que contenían otras reglas dentro de ellas. Por ejemplo, la regla para el reconocimiento de perros contenía una subregla para el reconocimiento de colas. Husserl también consideró que tales reglas no tendrían que decirnos algo acerca de algún perro en particular, o de los perros en general, sino acerca de un perro típico. Todas

las ideas básicas utilizadas por Minsky y sus estudiantes de inteligencia artificial concordaban con esto.[17]

Pero aún cuando muchas de las ideas básicas de la IA pueden concordar con la filosofía clásica, los investigadores de IA tienen que desarrollar los sistemas filosóficos particulares que utilizan: tienen que aclarar proposiciones oscuras y desarrollar varias líneas de investigación dejando aparte el esquema de los sistemas filosóficos originales. Por ejemplo, en "*On the Art of Combinations*" (1666), Leibniz propuso que todo el razonamiento puede reducirse a una combinación ordenada de elementos. Si pudiéramos definir semejante álgebra del pensamiento, sería posible el razonamiento en una máquina. Tal máquina podría resolver cualquier controversia filosófica, así como realizar descubrimientos por sí misma. La tesis de Leibniz viene a ser una teoría de la inteligencia artificial del siglo XVII. Sin embargo, Leibniz no tuvo que desarrollar demasiadas cuestiones concretas acerca de la correlación entre sus elementos, acerca de problemas de suficiencia o sobre la inferencia de conclusiones correctas a partir de premisas correctas.; es decir, nunca tuvo que someter a examen su programa. El "*General Problem Solver*" (*GPS*) desarrollado principalmente por Allen Newell y Herbert Simon es uno de los primeros y más generales enfoques en la ciencia cognitiva.[18] El modo en el que *GPS* describe el razonamiento (en términos de simples símbolos algebraicos y operaciones que combinan tales símbolos en expresiones) se deriva directamente de los pensamientos de Leibniz y del examen de ellos. Al parecer, los desarrolladores de sistemas de IA nunca han enfatizado cuánto de su trabajo descansa sobre las teorías filosóficas y se desarrolla en relación a ellas. Así, por ejemplo, las discusión acerca de las relaciones entre las representaciones de *GPS* y las "combinaciones" de Leibniz es más bien sugestiva –e inusual.

Una realización de la filosofía.

[17] Hubert L. y Stuart E. Dreyfus, Mind Over Machine (New York: Free, 1986) 4.
[18] V. *Supra*, "La Búsqueda Heurística".

Otro ejemplo de la filosofía clásica puede servir como metáfora para la interpretación de las investigaciones de IA como filosofía. Esbozando una distinción entre la cosa-en-sí y el fenómeno que aparece ante nosotros, Kant escribió en su Crítica de la Razón Pura:

No puedo explorar mi alma como una cosa-en-sí por medio del razonamiento teórico (y menos por medio de la observación empírica); entonces, no puedo explorar el libre albedrío como una característica del ser… Sin embargo, puedo pensar sobre la libertad, es decir, al menos sobre la representación de ella sin contradicciones.[19]

Trasladando este ejemplo kantiano hacia el dominio de la IA: los investigadores, como seres conscientes, probablemente no puedan crear una conciencia artificial, pero sí pueden pensar sobre su propia conciencia y expresar sus pensamientos en algún lenguaje –en el lenguaje conceptual (en el caso de la filosofía), o mediante un lenguaje de programación (en el caso de los investigadores de IA).

Para desarrollar la ciencia cognitiva como una filosofía rigurosa, es necesario adoptar la premisa de que una descripción de estados de conciencia como estados realmente representativos puede ser consistente. Los estados de conciencia en sí mismos, junto con aptitudes, emociones y todo eso, no son representaciones en sí mismos y no pertenecen al ámbito del lenguaje; sin embargo, el hecho de que esos estados puedan ser expresados en formas verbales demuestra que es posible algún tipo de representación simbólica. Además, los estados de conciencia tienen una necesidad inherente de algún tipo de expresión, de tal modo que puedan ser comprendidos, y el lenguaje es *el* medio para simbolizar los estados internos. Schütz se refiere a este proceso como explicación.[20] Sin duda, la explicación es posible sólo para alguna parte de la conciencia, y no puede hacerse desde el "cero absoluto" a la enésima potencia. Pero la interpretación de situaciones es una de las principales

[19] Kant, Immanuel, "Preface to the Second Edition," Critique of Pure Reason, trans. Kemp Smith (1787; New York: St. Martin's, 1965) 28 (B XXVIII), citado por Sharoff, Serge. "Constructions of the Mind". SEHR, vol. 4, núm. 2.

[20] Schütz , Alfred y Luckman, Thomas. The Structures of the Life-World, Northwestern UP, Evanston, IL, 1973.

actividades de la conciencia, y explicarlas a través del lenguaje es necesario para la socialización y expansión del "almacén de conocimiento" de la conciencia. Schütz utiliza la frase "dado por hecho" para describir la postura común y corriente que uno adopta todos los días en torno a los fenómenos tales como características del mundo y de otros seres concientes. En efecto, lo que esta actitud "natural" da por hecho es precisamente la posibilidad de describir la conciencia. Podemos retomar una cita de Pascal que Dreyfus y Dreyfus usan como título al prólogo de su libro: "El corazón tiene sus razones que la razón no entiende". Sin duda, hay una razón por la que la tradición filosófica europea ha intentado explicar, desde hace mucho tiempo, los procesos de la conciencia. No hay razón para afirmar que este intento carece ya de validez.

Conceptos básicos de la fenomenología de la computación.

Muchas de las ideas fundamentales de la fenomenología de Husserl y del Heidegger inicial se prestan a una interpretación desde el punto de vista de la ciencia cognitiva; nociones como fenómeno, la constitución del significado, tener-a-la-mano, intencionalidad, horizonte y conciencia del tiempo-interno.

Para los propósitos de este tema, la fenomenología puede describirse como la filosofía de las representaciones dinámicas. En *Verdad y Método*, Hans-Georg Gadamer cita las palabras de Schleiermacher como slogan para su filosofía: "El florecimiento es la verdadera madurez. Una fruta madura es tan sólo una superficie caótica que no pertenece a la planta orgánica." El objetivo de la descripción fenomenológica es examinar en detalle la vida pensante oculta dentro de nosotros. En contraste con la filosofía analítica que sustituye con construcciones simplificadas lo inmediatamente dado en toda su complejidad, y aplica la "navaja de Ockham", la fenomenología se resiste a todas las reinterpretaciones transformantes de lo dado, analizándolo por lo que es en sí mismo y sobre sus propios términos.

El concepto clave de la fenomenología es la noción de constitución, una descripción de la dinámica creativa del fenómeno de la conciencia. Como Husserl escribió, "es necesario mostrar en cada acto constitutivo concreto

cómo se crea el sentido del fenómeno."[21] La fenomenología da una descripción compleja del fenómeno como "aquello que muestra su mismidad a través de sí mismo." Para nuestro propósito –el de describir una fenomenología de la computación- basta considerar un fenómeno como un constructo mental que tiene lugar en la conciencia, junto con otros fenómenos, y que tiene la habilidad de revelarse a sí mismo.

El solipsismo metodológico de Husserl corresponde muy aproximadamente a la naturaleza de las representaciones computacionales. Sus descripciones se refieren exclusivamente a fenómenos subjetivos. El mundo externo está fuera de contexto; como Husserl dice, la *epoché* está comprometida. Un acto mental, como la fenomenología lo describe, no concierne a las cosas materiales sino a sí mismo. Husserl usa la noción de intencionalidad, la dirección de la conciencia hacia un objeto percibido, para describir la interacción existente entre la conciencia y los objetos en el mundo externo. A través de la intencionalidad, la conciencia viene a representar el objeto como un fenómeno.

La intencionalidad expresa la característica fundamental de la conciencia: siempre es *conciencia de algo*. La conciencia no es un mecanismo abstracto que procesa datos; su estructura interna se correlaciona con y, por tanto, depende del fenómeno comprendido. Esto asegura la imposibilidad de una descripción de la conciencia que esté separada de los objetos percibidos. Husserl escribió:

En todas las experiencias psíquicas (en la percepción de algo, el juicio sobre algo, la voluntad de algo, el goce de algo, la esperanza de algo, etc.) se encuentra inherentemente un estar-dirigido-a… Las experiencias son intencionales. Este estar-dirigido-a no está simplemente unido a la experiencia como una mera adición y, ocasionalmente, como una reacción accidental, como si las experiencias pudieran ser lo que son sin una relación intencional.

[21] Husserl, E. "The Paris Lectures", Husserliana. (The Hague: Nijhoff, 1975), 1:3-39, citado por Sharoff, Serge. "Constructions of the Mind". SEHR, vol. 4, *issue* 2.

Con la intencionalidad de las experiencias se evidencia, más bien, la estructura esencial de lo puramente psíquico.[22]

La noción de intencionalidad se hizo popular en el mundo de la IA por J. R. Searle, quien la describió como una característica de muchos estados mentales y sucesos, mediante la cual los actos conscientes son dirigidos a objetos y estados acerca de asuntos del mundo externo. Searle afirmaba que él quería remover algunas de las peculiaridades de ciertas tradiciones filosóficas antiguas. Todavía las definiciones de Husserl y Searle sobre la intencionalidad son muy similares. Como el mismo Searle admitía, la principal diferencia estriba en los modos de usar esas nociones. La idea de intencionalidad puede ser interpretada a partir de dos puntos de vista diferentes: es tanto la dirección de los actos de conciencia hacia objetos en el mundo externo, como el modo en el que los fenómenos existen dentro de la conciencia. Pero en relación a la primera interpretación, Husserl escribió:

La invención de la intencionalidad concebida por Brentano no resultó ser un naturalismo, el cual, digamos, capturaba las experiencias intencionales y cerraba el camino a las verdaderas tareas de la investigación de la intencionalidad.[23]

Es precisamente la segunda interpretación, y el trabajo consiguiente en la descripción de los fenómenos de la conciencia, lo que Husserl consideraba como la verdadera tarea de las investigaciones sobre la intencionalidad. El método Husserliano de análisis de la conciencia es puramente descriptivo. El hecho de que el mundo externo esté fuera de contexto, de que la *epoché* esté comprometida, no niega este mundo; el mundo externo mantiene su existencia. Un filósofo comprometido con la *epoché* se rehusa a tratar con el mundo externo antes de su entrada en la conciencia. La diferencia entre lo "imaginado" o carente de sentido (*e.g.* un centauro fumando pipa) y lo "real" u

[22] Searle, John R. "The Nature of Intentional States," Intentionality: An Essay in the Philosophy of Mind Cambridge UP, Cambridge, 1983, 1-29, citado por Sharoff, Serge. "Constructions of the Mind". SEHR, vol. 4, *issue* 2.
[23] Husserl, E. "Amsterdam Reports: Phenomenological Psychology," Husserliana, 9, citado por Sharoff, Serge. "Constructions of the Mind". SEHR, *volume 4, issue 2*.

objetos sensibles radica solamente en el modo en que se dan los fenómenos; ambos objetos están intencionalmente representados o actualizados dentro de la conciencia. La dirección de la conciencia hacia un objeto resulta ser el acto de dotar de sentido a algo. En palabras de Husserl, cuando hablamos acerca del sentido, "hablamos acerca de una entidad ideal que puede ser algo implícito en la infinidad abierta de experiencias reales y posibles que dan sentido".[24]

Husserl usa la noción de acto (es decir, un acto de dar sentido o un acto de percibir el tiempo) solamente para indicar la síntesis pasiva en la conciencia de lo que algo significa. La noción de acto no significa una acción conciente, así como la intencionalidad no significa un deseo, por lo que siempre estamos ya situados dentro de la conciencia, aún cuando la analicemos. De otra forma, si asignamos un acto a un hipotético actor, tenemos que describir las funciones de una conciencia interna, una conciencia dentro de la conciencia: en términos de programación, nos encontramos atrapados sin esperanza alguna dentro de un ciclo sin fin.

La intencionalidad concierne a los fenómenos en el centro de la conciencia, en su foco. En la periferia de la conciencia está lo que Husserl llamaba el "horizonte", el trasfondo que proporciona las condiciones para la comprensión de los fenómenos. En otras palabras, lo que el horizonte aporta es una pre-comprensión (*Vorverständnis*). Así, entendemos el significado de las palabras en el contexto de un horizonte constituido por nuestro entendimiento de otras palabras y sus relaciones. Al describir las relaciones entre horizonte e intencionalidad, Husserl señaló:

La conciencia –donde el objeto dado se dirige hacia su realización- no es como una caja con datos adentro. Un determinado estado de conciencia está constituído de tal manera que cada objeto muestre su mismidad.[25]

Heidegger usa una noción de horizonte similar a la de Husserl: lo que está a la mano (*Zuhanden*). La palabra *Zuhanden* –a la mano- enfatiza el hecho de que los objetos relevantes están sostenidos en el centro de la

[24] *Idem.*
[25] Edmund Husserl, *The Idea of Phenomenology: Lectures for an Introduction to Phenomenology*, trans. William A. Alston and George Nakhnikian (The Hague: Nijhoff, 1964). Citado por S. Sharoff. *Op.Cit.*

conciencia. Tanto el horizonte como la intencionalidad están en continuo cambio y, un fenómeno puesto en el horizonte, en el trasfondo, puede ser repentinamente llevado al centro por la conciencia. Recíprocamente, los fenómenos concentrados en el campo de la intencionalidad forman parte del horizonte del próximo campo de intencionalidad. A medida que pasan del centro a la periferia, pasan del presente al pasado inmediato; se sumergen en el horizonte, se hunden en el tiempo.

Conciencia-tiempo interno.

Para describir la constitución de los fenómenos mentales, necesitamos alguna posibilidad de representar el tiempo en la conciencia. La fuente original de la fenomenología del tiempo puede tal vez encontrarse en los escritos de San Agustín. En sus lecturas sobre la fenomenología del tiempo, Husserl cita una descripción del tiempo tomada del décimo primer volumen de las Confesiones: “Mido el tiempo en mi alma.” Como San Agustín, Husserl rechaza cualquier noción objetiva del tiempo. Todos los fenómenos están representados en la conciencia, y la conciencia trabaja con significados creados a través de la intencionalidad. Además, los fenómenos temporales están constituídos por la conciencia, pero se refieren a un estado de cosas acerca del mundo externo.

Husserl describe el tiempo a través de una estructura tripartita constituída por la protención, el momento actual y la retención. La protención es una anticipación al futuro, las diferentes expectativas que constituyen y condicionan “lo que viene”. En el momento actual de Husserl se encuentran el horizonte actual y la intencionalidad constituida dentro de este horizonte. La retención es la cadena formada por el pasado, las reflexiones (*Abschatungen*) sobre fenómenos anteriores guardados en la conciencia.

Para Husserl, la conciencia es un flujo de fenómenos moviéndose a lo largo de una línea común o girando alrededor de un pivote común: la revelación o el desdoblamiento de las experiencias en el tiempo. El proceso de revelación determina la estructura de dichas experiencias. En otras palabras, el tiempo-conciencia proporciona la base organizativa para todas las demás actividades de la conciencia. Como otros fenómenos, el tiempo se constituye en la

conciencia, pero también proporciona una base para constituir otros fenómenos, porque cada fenómeno comienza, se transforma y concluye a lo largo de la vida integral de la conciencia. Cada instante de este tiempo se manifiesta en una gradación continua de "sensaciones temporales", por decirlo así; cada fase actual del curso de la conciencia – en cuanto se manifiesta en ella todo un horizonte temporal de dicho curso – posee una forma, que abraza todo su contenido y permanece idéntica continuamente, mientras su contenido cambia sin cesar.

Es posible distinguir entre dos tipos de intencionalidad: longitudinal y transversal. La transversal significa considerar un objeto desde varias perspectivas en el tiempo: cómo comienza, cambia y termina; proporciona la posibilidad de considerar un objeto temporal. La intencionalidad longitudinal, por otra parte, nos permite considerar el flujo de la conciencia misma. La conciencia puede así convertirse en el objeto de análisis debido a la retención, la cadena formada a partir del pasado.[26] Lógica y prácticamente, la introducción del tiempo nos permite escapar de otro ciclo sin fin: la regresión infinita de la conciencia reflexiva, de una conciencia que analiza su propia conciencia. La cuestión para la IA es, sin embargo, si podemos analizar nuestra propia conciencia.

Obstáculos a la fenomenología de la computación.

Una de las piedras angulares de la filosofía de Heidegger es su rechazo a aislar el lenguaje y la mente del ámbito social y corporal. Según él, la metafísica de la Nueva Era reemplaza el mundo por una representación y a un hombre por un sujeto. La representación de conceptos fenomenológicos y el análisis de la conciencia y el lenguaje sin incorporar el sistema completo, significa que perdamos toda conexión real con una situación; la respuesta de Husserl al problema de la incorporación de la conciencia, la fenomenología de la conciencia, tenía ciertos problemas. Un intento de encerrarse en la pura subjetividad podría fácilmente fallar. Cuando se describe una situación parece difícil negar la influencia directa, más que interiorizada, de las fuerzas externas.

[26] Husserl, E. Husserliana, 10:119.

La fuente del tiempo –el hecho de durar- es también muy difícil de ubicar exclusivamente dentro de nosotros mismos. Como resultado, una tesis referente a la unión entre la conciencia y el cuerpo, el ser-en-el-mundo, vino a ser un sujeto clave en las versiones fenomenológicas de Heidegger y Merleau-Ponty.

Heidegger cambió el énfasis de Husserl y pasó del análisis de la conciencia al análisis del ser (*Dasein-analytik*); por esta razón, sus problemas difieren de los de la ciencia cognitiva. Aunque la exploración de Dreyfus en torno al trasfondo filosófico de la IA, basado en la ontología de Heidegger,[27] nos lleva a una conclusión adecuada acerca de la imposibilidad de tal proyecto, su imposibilidad surge de la fenomenología del *ser* de Heidegger, más que del análisis de la conciencia de Husserl. En contraste, el análisis descriptivo que hace Husserl de la conciencia a través de la reflexión, corresponde muy aproximadamente a las metas de la ciencia cognitiva. En cuanto a la cuestión acerca de un método para la psicología fenomenológica, Husserl escribió:

La reflexión debe hacerse de tal manera que la variable y fluctuante vida del ego, la vida de la conciencia, no se aprecie en su superficie pero, en cambio, sea explicada en la contemplación, de acuerdo a sus partes constitutivas esenciales.[28]

El intento de Husserl de explorar la constitución de los fenómenos en la conciencia –su búsqueda de modelos de generación de los fenómenos- va de acuerdo con la esencia de la ciencia cognitiva. Sin embargo, Husserl impuso ciertas limitaciones al alcance de su análisis cuando escribía acerca del campo de la pre-intencionalidad, el campo de la pura posibilidad de la intención, que constituye la corriente principal y que limita la posibilidad del análisis mediante el lenguaje. Esta corriente principal se convierte en el problema crucial cuando consideramos la posibilidad de continuar la tradición filosófica europea a través de la ciencia cognitiva. Hay dos posibilidades. Si la ciencia cognitiva puede usarse para interpretar la estructura de la corriente principal, entonces se trasciende la filosofía tradicional, ya que va más allá del enfoque lingüístico de

[27] Dreyfus, Hubert L. What Computers Can't Do, Harper, New York, 1972.
[28] Husserl, "Amsterdam Reports.", citado por Sharoff, Serge, *op. cit*

un pensador humano; es decir, tal modelo estaría en un nivel abajo de la descripción tradicional. La segunda posibilidad es que la corriente principal es tan inalcanzable al análisis basado en la ciencia cognitiva, como lo eran las descripciones antiguas basadas en la filosofía tradicional.

Según Serge Sharoff, la fenomenología impone algunas restricciones en su aplicación a la ciencia de la computación. ¿Cuán significativa podría ser una fenomenología de la computación bajo tales restricciones? Al parecer, la fenomenología no aporta conceptos interesantes para el desarrollo del paradigma clásico de la IA.[29] Ésta ya no puede intentar crear "máquinas pensantes"; el único proyecto de computación posible es el desarrollo de alguna versión filosófica. Podemos elaborar nuestros conceptos, pero no podemos elaborarnos a nosotros mismos. Aún tal suposición es injustificable si no consideramos cómo vamos a concretizar nuestra postura, si no clarificamos nuestros conceptos mediante el uso de la computadora. Lo siguiente es un intento de aplicación de los conceptos fenomenológicos a los campos tradicionales de la IA.

Una interpretación de los conceptos fenomenológicos para la ciencia cognitiva.

El horizonte de Husserl y el tiempo interno de la conciencia pueden aportar ideas clave para una versión computarizada de la fenomenología, una versión que utilice representaciones continuamente cambiantes de información con objeto de asegurar el tener-a-la-mano los hechos relevantes. Esta situación se constituye a sí misma en la base de la historia de las etapas anteriores.

La principal distinción entre los enfoques clásico y fenomenológico de la ciencia cognitiva radica en el uso de la información almacenada, por ejemplo, vocabulario, textos y esquemas. Esencialmente, la distinción recae entre el recuerdo dinámico y la memoria constante. Las representaciones dinámicas suponen que los textos o el significado de las palabras no son simplemente seleccionados a partir de un vocabulario; más bien, son creadas y constituídas durante el proceso de análisis. El problema de la polisemia no surge durante el

[29] Sharoff, Serge, *op. cit*

proceso de entendimiento, ya que cada situación se constituye a sí misma de tal manera que una "palabra significativa" *tiene que* ser integrada en su contexto; el significado de una palabra es su significado en cierta situación. Es la situación la que integra el significado a una palabra.

También se puede aplicar el esquema de la intencionalidad de Husserl al campo de la comprensión del lenguaje natural. Una palabra u oración puede servir como objeto de intencionalidad; el curso de interpretación de palabras y oraciones produce el curso de actos de intencionalidad. El acto de dar sentido a una palabra tiene lugar en el contexto de otras palabras y de una comprensión general de la situación. El horizonte se forma del momento actual, combinando el curso de actos intencionales con los supuestos que integran el trasfondo del entendimiento. En cierto sentido, la noción de horizonte corresponde a la noción tradicional de contexto. La descripción de Husserl, sin embargo, se resiste a la separación tradicional entre foco y contexto, ya que la intencionalidad presupone que la estructura de un objeto intencional corresponde a la estructura del horizonte.

El significado se representa como un fenómeno que se revela. Por un lado, tenemos una determinación un tanto vaga de la palabra, una aparente unidad de significado: la forma externa del fenómeno. Y por otra parte, el contenido interno de la palabra. Éste último no incluye todos los significados, por ejemplo, de la palabra "agua" (agua1, agua2), o la serie de situaciones en las que se aplica tal o cual significado. Más bien se trata de una sola estructura significativa, digamos, "empacada", que no tiene significado sino en la forma en que es asumida o actualizada. Esta actualización se lleva a cabo en el marco del horizonte actual, el cual provee una base para la pre-comprensión, y que construye las formas de significado correspondientes a una situación dada. En términos fenomenológicos, el resultado es una intención de significado, un acto de dar cierto sentido. Las características esenciales del significado aparecen en primer plano cuando la forma interna se concretiza en determinada situación.

La descripción del significado en la tradición analítica es radicalmente diferente. Ésta enfatiza el significado resultante como una "objetificación"; el significado es separado en la intencionalidad que lo constituye, de tal modo que se convierte en un objeto analizable en forma independiente. En sus último

escritos, Wittgenstein afirmó la idea de que estamos atados a la idea errónea de que el significado de una oración resulta de ella y la sigue permanentemente. Una cita muy distinta refleja los pensamientos del primer Wittgenstein: "Una oración es entendida si sus partes constitutivas son entendidas."[30] Un ejemplo de análisis situacional: Wittgenstein, Moore y Malcolm han analizado la expresión "Yo sé". El significado de esta expresión en la vida real depende de la situación: afirmación, convicción, evidencia, certeza. Un invidente diría "Yo sé que ésto es un árbol", en lugar de "Yo veo."

Podemos aplicar la distinción entre descripciones objetificadas e interpretadas para diferenciar entre planes de acción (scripts usados en ciencia cognitiva) y actividades de manejo-situacional. Mientras que la descripción de la visita a un restaurant, como llegar, llamar al mesero, ordenar y todo eso, esté basada en nuestras visitas reales a un restaurant, no puede ser el resultado de una objetificación del significado. Tal descripción resulta evidentemente inadecuada si consideramos la necesidad de dividir esas escenas en sub-escenas, la posible aparición de circunstancias no consideradas y, la importancia de aprender el comportamiento apropiado en tal situación. Por ejemplo, podemos dividir la primera escena en abrir las puertas del restaurant, poner el abrigo en el lugar adecuado, encontrar un lugar libre. Pero este escenario podría desviarse en varias formas: tal vez un portero nos abriría la puerta; podría no ser necesario colocar el abrigo en ningún lugar (si fuera un día caluroso); podría no haber lugares libres. La situación se constituye a sí misma. Cuando nos sujetamos a *scripts* nos vemos en la necesidad de especificar muchas restricciones que van surgiendo de nuestra comprensión de una determinada situación. El número de restricciones se incrementa tanto que termina por nublar nuestro entendimiento.

El principio de evidencia constituye el núcleo de la metodología de Husserl. Las situaciones simples son evidentes. Cuando estamos inmersos en una cierta situación, algunos eventos nos parecen evidentes; tales eventos los tenemos a la mano. Cada paso emerge de otros previos, en un contexto particular de la situación; cada paso está determinado por su contexto. Utilizando la noción de horizonte de Husserl, la cual tiene la propiedad de

[30] Wittgenstein, Ludwig . *Tractatus Logico-Philosophicus*, ed. C.K. Ogden (1922; London: Kegan Paul, 1933) 4.024, citado por Sharoff, Serge, *op. cit.*

aparecer ante nosotros o de estar a la mano, podemos hacer referencia a un "espacio de solución del problema." Este espacio aporta datos esenciales y métodos adecuados para resolver un problema determinado. Pero, ciertamente, en el caso de un problema complejo o una mala interpretación, necesitaríamos hacer un esfuerzo por recuperar cualquier evidencia perdida.

En "*A Framework for Representing Knowledge*", Minsky trató acerca del proceso de construcción de este espacio de solución durante el análisis del problema.[31] Los marcos estructurales han sido muy populares como medios de representación del conocimiento.

En contraste con la aceptación general del marco estructurado, al análisis de la dinámica de la percepción de Minsky –la dinámica del desarrollo del marco durante el proceso de entendimiento de una situación- fue, en general, ignorada o simplemente mal interpretada. En la introducción, Minsky describe los principios básicos de utilización de un sistema basado en marcos estructurales:

Los diferentes marcos de un sistema describen la escena desde distintos puntos de vista, y las transformaciones entre un marco y otro representan los efectos del movimiento de un lugar a otro… Los diferentes marcos corresponden a diferentes vistas, y los nombres de los apuntadores entre los marcos corresponden a los movimientos o acciones que cambian el punto de vista.[32]

También usa esta idea en su discusión de la extendida noción del "espacio de solución del problema":

El principal objetivo en la solución de un problema sería tratar de entender el espacio del problema, el encontrar representaciones dentro de las cuales el problema fuera más fácil de resolver. El objetivo de esta búsqueda es obtener

[31] Minsky, Marvin. "A Framework for Representing Knowledge", The Psychology of Computer Vision, ed. Patrick Henry Winston (New York: McGraw, 1975).
[32] *Idem.* 212, 218.

información para una reformulación, y no –como comúnmente se asume- para encontrar soluciones…[33]

Minsky da una descripción acerca de cómo se percibe un cubo, que corresponde directamente al esquema fenomenológico; igualmente, Husserl usa esta misma percepción como el modelo más simple:

En la percepción de un cubo intervienen varios actos de intencionalidad; el cubo es representado desde diferentes puntos de vista y desde distintos ángulos. Las partes visibles del cubo se relacionan con las partes invisibles pero conocidas. Así que la percepción de una serie de vistas y el modo cómo son sintetizadas revelan la presencia de una conciencia única e indivisible que está dirigida hacia algo.[34]

En cuanto al análisis fenomenológico de los procesos de memorización y recuerdo –esto es, la actualización de la información memorizada-, Sharoff apunta lo siguiente. Esos procesos pueden relacionarse con la conciencia del tiempo interno y la característica de "estar-ahí", a la mano. Cuando tratamos de entender cualquier descripción, las representaciones que construimos durante este proceso son colocadas en capas. Cada representación está dispuesta sobre la anterior y la modifica. Husserl utiliza la siguiente metáfora para describir el modo en el que retenemos una imágen de un objeto: podemos ver su imágen previa como si lo hiciéramos a través de una capa transparente de agua. El proceso de establecer capas y de modificación de representaciones es un resultado del "empaquetamiento" de una respresentación bajo la presión de y junto con las representaciones que le siguen. Las últimas representaciones se ubican en la memoria de largo plazo; podemos decir que "el agua de Husserl" las oculta. Cuando leemos un texto con palabras desconocidas o combinaciones inesperadas de palabras conocidas, tales palabras y los contextos en los que se usan son sobrecubiertas o sedimentadas. Cuando consideramos una situación o una

[33] *Idem.* 259.
[34] Husserl, "*The Paris Lectures*", citado por Sharoff, Serge. "*Constructions of the Mind*". SEHR, *volume 4, issue 2*.

descripción de una situación, el conjunto de fenómenos correspondientes a la representación de la situación es también sedimentado. Solamente el núcleo de las representaciones permanece. Éste puede ser "desarrollado" posteriormente, en diversas formas de acuerdo a las diferentes situaciones.

Tales procesos de desarrollo, de "desempaquetamiento" del núcleo de acuerdo tanto al horizonte actual como al estado actual de conciencia, constituye la esencia del recuerdo. Este desempaquetamiento será diferente cuando el horizonte o el estado de intencionalidad –el modo de conciencia- sea diferente; la conciencia revelará otro aspecto de alguna palabra, el aspecto que está-a-la-mano. La estructura de la conciencia corresponde al objeto comprehendido debido a la intencionalidad. Lo que es recordado como una parte de la estructura de la conciencia corresponde al objeto comprehendido y a otras partes de la estructura de la conciencia. El recordar no funciona como si cortáramos y apartáramos una representación de la memoria, la cual está estructurada (de acuerdo a esta perspectiva) como un arreglo lineal o vector unidimensional de datos que son accesibles según su posición. Tal vector podría ser un medio insensible que podría no depender de las estructuras de datos contenidas en él. Sin embargo, recordar es el *proceso* de constituir la representación necesaria para la situación. La metáfora del río puede ayudar a clarificar el proceso de recordar. El agua del río no recuerda la dirección de la corriente. Moviéndose en el lecho del río, la corriente provee la dirección necesaria del movimiento. El río corresponde aquí a un acto intencional de recuerdo, y su lecho corresponde al horizonte. Una vez que el núcleo de algún objeto recordado es desarrollado, puede volver a ser sedimentado después de haber sido enriquecido por los contextos situacionales. La sedimentación subsecuente no es, empero, siempre necesaria: para las palabras bien conocidas o las situaciones familiares, por ejemplo. En este caso, el desarrollo del significado prototípico no altera mucho la estructura íntegra de la conciencia ni el campo protencional. Esas palabras o situaciones devienen fenómenos con una correspondencia a sus tipos y a la estructura de conciencia actual. Con el tiempo se hundirán también, pero sin mucha sedimentación adicional sobre sus características prototípicas.

Desde el punto de vista de la ciencia cognitiva, la representación de la estructura conciencia-tiempo interno abre la posibilidad de que una aplicación

pueda usar su propia historia –significando aquí simplemente una secuencia de estados previos. Si un sistema de IA entiende su historia, este entendimiento permite la posibilidad de que pueda reflejarse en sus propias acciones o representaciones. La restricción de la lógica del sentido común y la limitación de recursos están también relacionadas con la conciencia-tiempo interno.

La posibilidad de usar la estructura interna del flujo de tiempo provee un mecanismo para establecer restricciones en el proceso del entendimiento, de tal manera que el entendimiento venga a ser un proceso con recursos limitados. Muchas veces la inferencia lógica ocasiona el problema de una infinidad "superficial", pero un modelaje natural de un ámbito de recursos limitados nos permite evitar paradojas como la siguiente: Ai está junto a Ai+1, Ai+1 está junto a Ai+2, …Ai+99 está junto a Ai+100. El fijar la longitud de la cadena del razonamiento y el restringir la lógica mediante el "sentido común" pueden implementarse de una manera relativamente simple, usando el modelo de la conciencia-tiempo interno.

Al mismo tiempo, la representación de la secuencia temporal del razonamiento puede evitarnos caer en un recorrido hacia atrás, que es posible en algunos lenguajes de programación de lógica como Prolog. Un problema que surge de tal recorrido hacia atrás es el siguiente: si en el razonamiento se alcanza algún punto muerto –es decir, si alguna cláusula de Prolog no puede ser satisfecha- entonces las asignaciones que el intérprete de Prolog hubiese hecho a las variables libres, a las estructuras de datos, son reiniciadas hacia atrás. En tal caso, el mecanismo de identificación de patrones de Prolog trata de unificar variables por otros medios, de acuerdo a una cláusula alternativa. Ésta última trata de continuar su inferencia y no usa los resultados de la inferencia hecha por la cláusula anterior. Cuando hacemos un análisis dentro del flujo de actos temporales en el que nuestros pasos previos son accesibles casi inmediatamente (como a través de una capa de agua transparente), mantenemos la posibilidad de re-utilizar tales pasos, de usar pruebas correctas parcialmente.

El movimiento fenomenológico puede sugerir conceptos nuevos al enfoque de programación orientado a objetos. Los fenómenos son entidades autoreveladas introducidas en la conciencia y desarrolladas en el marco del horizonte actual. Este esquema nos recuerda los conceptos de orientación a

objetos, los cuales suponen que un objeto es dispuesto en el marco estructural formado por el ámbito de programación actual, y que le es asignado un determinado comportamiento conforme a la clase con la que se le inicializó. Posteriormente, el objeto recibe mensajes externos y reacciona a ellos de acuerdo a las propiedades de su clase. El flujo de la conciencia puede también considerarse como un fenómeno. En una interpretación orientada a objetos de la fenomenología, la conciencia puede ser considerada un objeto de la clase "continente", la cual puede contener otros objetos –fenómenos. Además de su estatus que comparten como fenómenos, tales objetos tienen una reacción específica a los mensajes comunes en virtud de su contenido, de la transición de representaciones desde un estado potencial a uno articulado, desde un sistema interno de imágenes (por ejemplo, la forma interna de una palabra) a una representación lingüística externa accesible a un observador externo. Las clases de objetos deben ser fenómenos también, ya que las estructuras noemáticas corresponden a estructuras noéticas y tienen que ser explicadas reflexivamente. Este hecho no es nuevo para la programación orientada a objetos. El protocolo de meta-objetos nos brinda la posibilidad de considerar las clases como objetos de algún tipo.

Las nociones de programación orientada a objetos pueden interpretarse, entonces, fenomenológicamente. La descripción fenomenológica es, sin embargo, mucho más compleja y no se ajusta completamente a este esquema de orientación a objetos. Primero que nada, después de revelar el fenómeno de la conciencia durante el proceso de desarrollo, la explicación puede implementarse bajo el protocolo de meta-objetos. Pero un fenómeno no simplemente desarrolla sus propiedades; más bien, éstas son creadas y constituídas sobre la base de un objeto. Este tipo de "programación constitutiva" es inusual o, cuando menos, no es una práctica regular.

Conclusiones.

Algunas intenciones y nociones fenomenológicas encajan dentro del paradigma de la IA actual, al menos dentro de una interpretación más o menos amplia. Dentro de ellas se incluye el *stress* como modelo de producción de la conciencia, el horizonte (que corresponde al contexto), y la intencionalidad de

Husserl (la cual, en términos de IA, significa la correspondencia entre un algoritmo y una estructura de datos). Otras nociones no se ajustan a este paradigma. La noción de conciencia-tiempo interno, por ejemplo, es una de ellas, aunque representa el núcleo que puede iluminar con nueva luz al primer grupo de conceptos. Sin embargo, hay que cuestionarse acerca de las posibles implicaciones de este análisis. ¿Es posible que la fenomenología aporte una base más adecuada para las aplicaciones de IA? Para responder a esta pregunta es necesario separar tres afirmaciones relacionadas entre sí: la común, la débil y la fuerte. La afirmación común es la de que las investigaciones filosóficas aportan una base adecuada para las investigaciones en la ciencia cognitiva. La afirmación débil dice que una interpretación computacional del movimiento fenomenológico es posible al menos en cierto sentido. La afirmación fuerte sostiene que las aplicaciones de la IA constituídas sobre la base de las conceptos fenomenológicos será más efectiva y poderosa.

Ya sea que aceptemos o no la afirmación fuerte, la postura del investigador, finalmente, es una cuestión de creencia. La historia de las aplicaciones de la IA clásica implícitamente basada en la filosofía analítica, muestra el gran número de problemas relacionados con la descripción de la conciencia por tales medios. Posiblemente las descripciones fenomenológicas puedan considerarse como estrategias efectivas en torno a algunos problemas de IA. No es seguro, sin embargo, de que esta tesis sea absolutamente correcta. La afirmación débil, en un significado más general, podría ser interesante como un desarrollo de la fenomenología. Cuando menos, los problemas de la fenomenología como investigadora de “la vida oculta de los pensamientos”, tienen una relación directa con los problemas de la ciencia cognitiva. La afirmación común tiene una significación general también, y puede aportar el potencial para interpretar las investigaciones actuales de la ciencia cognitiva desde un punto de vista filosófico, y no desde el punto de vista de ciencias tales como la biología o la psicología. Por supuesto, la interpretación de la ciencia cognitiva como filosofía es una metáfora que puede ayudar a dilucidar la posición de la ciencia cognitiva en la perspectiva histórica de la investigación de la conciencia. Los métodos de la ciencia cognitiva y de la filosofía son diferentes: los textos filosóficos son interpretados por humanos,

quienes son portadores de conciencia. En cambio, los programas computacionales operan (en principio) sobre una sustancia inconciente, y la ciencia cognitiva está dirigida hacia la construcción de modelos de producción. Pero esta metáfora resulta fructífera ya que permite el enriquecimiento de las líneas de investigación actuales en ciencia cognitiva con un gran número de ideas filosóficas clásicas. Dreyfus y Dreyfus son seguidores de la opinión de Heidegger sobre el fin de la filosofía y el surgimiento de la realidad cibernética. Esto no deja espacio para el desarrollo de la filosofía, mucho menos para su desarrollo por medio de la ciencia cognitiva. Sin embargo, una inversión de la tesis de Heidegger sobre la sustitución de la filosofía por la cibernética apunta a una metáfora interesante. Con una interpretación tal, podemos eludir la inquietante pregunta, "¿Pueden pensar las computadoras?" y formular una cuestión más productiva, "¿Cómo pueden los conceptos filosóficos ser interpretados por una computadora?".

Conceptos de tecnología y sistemas inteligentes

Se entiende por *técnica* un saber práctico acumulativo sin suficiente apoyo teórico. Es un saber operativo pero carece de fundamentación científica. En este sentido técnica se utiliza hasta que no aparece una fundamentación científica del quehacer operativo. Técnica viene a ser tecnología antigua; anterior a la constitución de la tecnología moderna.

La *tecnología* es aquel tipo de saber operativo y transformador que cuenta con apoyo teórico, que cuenta con el *know how*, el *saber cómo hacer.* Transformación creativa de la realidad. Tras la tecnología las cosas nunca quedan igual. Desde el punto de vista de los objetivos, la ciencia busca conocer, informarse, sin embargo los objetivos de la tecnología son transformar la realidad.

La ciencia no aspira a modificar la realidad en cuanto tal ciencia; pero la tecnología sí busca introducir la información en ámbitos reales naturales o artificiales (afecten a la Naturaleza, a la sociedad y/o a los artefactos).

De modo que la tecnología tanto desde el punto de vista de los objetivos, como de los procesos y como de los resultados, es distinta de la ciencia. El tecnólogo lo que busca es intervenir en el curso de las acontecimientos, o bien,

propiciar otros. La tecnología está en función de ciertos objetivos, pensados a tenor de los valores culturales, ideológicos y sociales.

Estos sistemas de valores dictaminan qué ha de hacerse y qué ha de evitarse. Muchas veces los valores son económicos. El objetivo tecnológico tiene que ver con la eficacia y con la eficiencia. Al valorar una tecnología, se valora en términos de eficacia. En segundo lugar se trata de conseguir el objetivo con el menor número posible de medios, de esto trata la eficiencia.

La Inteligencia Artificial es el resultado de implementar en un objeto inanimado las facultades humanas que configuran la inteligencia. Nuestros tiempos, técnicamente más avanzados, son el marco ideal para el florecimiento de la IA en las computadoras. La (relativamente) nueva ciencia cognitiva precisa un estudio sistemático de todos aquellos factores que moldean nuestras facultades, pues en el caso contrario sería imposible implementarlos correctamente. A grandes rasgos estos factores son los siguientes:

Conocimientos generales. Para una correcta emulación de la inteligencia, el sistema debe disponer de un conocimiento general, que abarque todos los campos, equivalente a la cultura adquirida por un humano.

Uso del lenguaje: Todo aquello que persiga ostentar la etiqueta de IA debe ser capaz de comunicarse de forma lógica en un lenguaje comprensible y humano. Esto implica un perfecto dominio de la expresión escrita y una completa capacidad de entendimiento y síntesis de voz. Un ejemplo: "los principales mandos europeos se reunieron en la cumbre de la OTAN". Un sistema IA ha de ser capaz de distinguir el significado ambiguo de "cumbre", para no confundir a la OTAN con una montaña.

Procesamiento visual. La percepción visual del entorno es el sistema principal de los humanos para conocer e interpretar su medio. En un sistema de IA completo, esta característica tiene que estar disponible y sin limitaciones: Se ha de lograr, además de la visión, la comprensión de lo visto.

Capacidad para tomar decisiones. Esta característica que ha de ser automática y flexible; quiere decir que, ante dos situaciones iguales, el sistema ha de tomar la decisión que considere más apropiada, en ambas.

Soluciones por experiencia. Las personas trabajamos por heurística. Esto significa que, ante un problema de características similares a otro anteriormente experimentado, podemos aplicar el conocimiento adquirido en ese momento para la solución del problema. Es una facultad complementaria a la anterior. La IA ha de crear equipos capaces de enriquecerse por medio de la experiencia. Todo lo anteriormente visto es realizado de una forma continua por los humanos, sin que esto disminuya la eficacia del resto de sus tareas; son facultades para nosotros tan triviales, que suelen ser agrupadas bajo el denominativo de "sentido común". Sin embargo, estos procesos, aparentemente sencillos, constituyen -por su dificultad de implementación en una máquina- las principales barreras ante la creación de un producto de IA. Un sector de pensamiento más radical exige que, para que un producto sea considerado IA, los métodos que utilice han de ser propiamente humanos; por otra parte, el área moderada de la ciencia, alude al fin del producto para justificar, de esta manera, los medios de desarrollo.

Tecnología para el aprendizaje.

El aprendizaje supone un proceso inteligente, repetitivo, eficaz y eficiente. Es interesante porque permite:

- resolver problemas cambiantes
- detectar y corregir conocimiento que se ha introducido inicialmente y es incorrecto
- resolver problemas en entornos inaccesibles
- resolver problemas desconocidos: por ejemplo, reconocer patrones, aun cuando no sabemos cuáles estamos buscando.

IA y Aprendizaje

En la Inteligencia Artificial (IA) se pueden observar, a grandes rasgos, dos enfoques diferentes:

a) La concepción de IA como el intento de desarrollar una tecnología capaz de suministrar al ordenador capacidades de razonamiento o discernimiento similares, o aparentemente similares a las de la inteligencia humana.

b) La concepción de IA como investigación relativa a los mecanismos de inteligencia humana (por extensión, investigación relativa a la vida y al universo), que emplea el ordenador como herramienta de simulación para la validación de teorías.

El primer enfoque es por lo general el más práctico, se centra en los resultados obtenidos, en la utilidad, y no tanto en el método. En este enfoque se encuadran, por ejemplo, los Sistemas Expertos. Son temas claves en esta dirección la representación y la gestión del conocimiento. Algunos autores representativos de este enfoque podrían ser McCarthy y Minsky, del MIT.

El segundo enfoque está orientado a la creación de un sistema artificial que sea capaz de realizar los procesos cognitivos humanos. Desde este punto de vista no es tan importante la utilidad del sistema creado (qué hace), como lo es método empleado (cómo lo hace). Como aspectos fundamentales de este enfoque se pueden señalar el aprendizaje y la adaptabilidad. Ambos presentan gran dificultad para ser incluidos en un sistema cognitivo artificial. Esta orientación es propia de Newell y Simon, de la *Carnegie Mellon University*.

Es importante indicar que frecuentemente ambas posturas no se pueden distinguir, ni siquiera en muchos trabajos de los autores mencionados como significativos en cada una de ellas.

En los dos planteamientos, pero especialmente en el segundo enfoque, uno de los mayores deseos es el poder contar con una arquitectura que soporte todo tipo de proceso inteligente.

En la ciencia cognitiva, el concepto "Arquitectura" se refiere a la estructura no flexible subyacente al dominio flexible del proceso cognitivo, es decir, a la estructura que soporta los procesos cognitivos superiores.

Las arquitecturas propuestas como bases de la cognición humana se denominan Arquitecturas Cognitivas, mientras que las correspondientes para la cognición artificial son llamadas Arquitecturas para Sistemas Inteligentes Integrados, o Arquitecturas para Agentes Inteligentes, o Arquitecturas Generales de Inteligencia.

Los intentos de construcción de sistemas cognitivos artificiales se basan en la hipótesis de Newell y Simon según la cual "un sistema físico de símbolos constituye el medio necesario y suficiente para una acción inteligente general".

Si se pudieran explicar los procesos cognitivos superiores de una manera intrínseca, es decir, si se pudiera demostrar que los procesos mentales inteligentes que realiza el hombre se producen a un nivel superior (o intermedio) con independencia de las capas subyacentes que existen hasta la constitución física del ente inteligente, se demostraría que es posible crear -mediante un sistema de símbolos físicos-, una estructura artificial que imite perfectamente la mente humana mediante una arquitectura de niveles, ya que se podría construir dicho nivel superior mediante la combinación de elementos que no necesariamente han de ser los que forman el nivel inferior en los humanos (que por ejemplo, podemos suponer que son las neuronas).

En cambio, si sólo se pudieran explicar los procesos cognitivos superiores mediante una descripción al más bajo nivel (comportamiento neuronal), sólo se podría imitar la inteligencia humana mediante la construcción de neuronas artificiales. Para ser exactos, esta afirmación está condicionada por la certeza de la suposición (bastante común) según la cual el neuronal es el más bajo de los niveles relevantes para la formación de los procesos cognitivos. Arbitrariamente, se podría haber elegido otro nivel aún mas bajo (moléculas, átomos). Llevado al extremo, se podría rescribir la afirmación, sustituyendo "neuronas" por "la más pequeña partícula de nuestro universo", si este fuera discreto (no infinitamente divisible).

Las denominaciones "nivel superior" y "nivel inferior" son arbitrarias en cuanto a que parece que se puede encontrar con facilidad un nivel que esté aún más bajo que el que hemos llamado "nivel inferior" -el nivel atómico es inferior al neuronal- y lo simétrico respecto al nivel superior -la conciencia colectiva es superior a la individual-. La existencia de una conciencia colectiva capaz de comunicarse a un nivel superior al del individuo parece evidente en los estudios sobre el comportamiento de algunos insectos, siempre que hagamos el esfuerzo de no interpretar el término "conciencia colectiva" desde nuestro punto de vista subjetivo como individuos. ¿Cómo conseguir esto? No es difícil, si se usa una analogía bajando un nivel. Imaginemos dos células (concretamente, dos neuronas) de nuestro cuerpo charlando amistosamente acerca de la posibilidad de que el conjunto de células forme una "conciencia colectiva". Las neuronas podrían hablar sobre esta "conciencia colectiva", ponerla en duda o intentar argumentar su existencia, pero difícilmente podrían

llegar a comprenderla, no puede ser un concepto familiar para ellas. E.O. Wilson, en "*The insect societies*" define la comunicación masiva como la transmisión de información, dentro de grupos, que un individuo particular no podría transmitir a otros.

El hecho de suponer que el comportamiento inteligente en el hombre se produce a un nivel superior con independencia de los niveles inferiores está íntimamente relacionado con el debate entre holismo o creencia en que "el todo es más que la suma de sus partes" y el reduccionismo, o creencia en que "un todo puede ser comprendido completamente si se entienden sus partes, y la naturaleza de su suma."

Los esfuerzos desarrollados en Arquitecturas Generales de Inteligencia son puramente reduccionistas. Por el contrario, el holismo subyacente en los modelos conexionistas como las Redes Neuronales Artificiales, sugiere el aspecto de la interdependencia entre algunos niveles, o lo que es lo mismo, la imposibilidad de sustituir un nivel (las conexiones neuronales, como sistema sub-simbólico) por otro que realice sus mismas funciones (sistema simbólico). Sin embargo, también las Redes Neuronales Artificiales pueden ser consideradas reduccionistas si tenemos en cuenta otros niveles aún más bajos.

IV. LA FUNCIÓN SIMBÓLICA DE LA INTELIGENCIA

Símbolos e inteligencia.

No existe una sola cosa elemental que explique la inteligencia en todas sus manifestaciones. No existe un "principio de la inteligencia" como no hay un "principio vital" del que por sí mismo emane la esencia de la vida. Sin embargo, esto no implica que no existan requisitos estructurales para la inteligencia. Uno de ellos es la capacidad para almacenar y manipular símbolos. Parafraseando las palabras de McCulloch, ¿qué es un símbolo, en tanto que puede ser usado por la inteligencia, y qué es la inteligencia, en tanto que puede utilizar un símbolo?

La naturaleza esencial de un sistema depende de la ciencia que lo estudia o utiliza. Los símbolos se caracterizan invariablemente de manera cualitativa, pues estipulan los términos en los que se busca lograr un conocimiento más detallado. La historia de la ciencia muestra que estas estructuras cualitativas, captadas en enunciados simples tales como las leyes, han sido de la mayor importancia. Así, la doctrina celular en biología establece que la célula es el componente básico de todos los organismos vivos. En geología, por otra parte, las placas tectónicas constituyen los cimientos de la superficie del globo terrestre y explican las formas y la localización relativa de los continentes y océanos y, en general, la configuración de las diversas zonas geológicas del planeta. Pruebas de esta teoría parecen encontrarse por doquier, pues vemos el mundo en sus términos y sobre sus símbolos. Dentro de las ciencias de la salud, la teoría de los gérmenes de las enfermedades, que ha revolucionado la medicina, establece que la mayoría de las enfermedades es causada por la presencia y multiplicación de organismos unicelulares. En fin, las leyes de estructura cualitativa establecen los términos en los que opera toda una ciencia.

Sistemas de símbolos físicos.

Un sistema de símbolos físicos consiste en un conjunto de entidades, llamadas símbolos, que son modelos físicos que pueden presentarse como componentes

de otro tipo de entidad llamada expresión (o estructura del símbolo).[35] La estructura del símbolo se compone de instancias (o signos) de símbolos relacionados de alguna manera física. El sistema contiene un conjunto de estas estructuras de símbolos. También contiene, además, un conjunto de procesos que operan en las expresiones y que dan lugar a otras. Existen procesos de creación, modificación, reproducción y destrucción. El sistema, propiamente, existe en un mundo de objetos más extenso que el de sus expresiones simbólicas.

Esta estructura de expresiones, símbolos y objetos tiene dos nociones centrales: la designación y la interpretación.

Designación: Una expresión designa un objeto si, dada la expresión, el sistema puede afectar al propio objeto o comportarse de una manera que dependa del objeto.[36]

Interpretación: El sistema puede interpretar una expresión si ésta designa un proceso y si, dada la expresión, el sistema puede llevar a cabo el proceso.[37]

El sistema puede realizar el proceso indicado, o bien ejecutar sus propios procesos de acuerdo a las expresiones que los designan.

Además de tener designación e interpretación, un sistema debe satisfacer los requisitos de integridad y de conclusión.

1) Un símbolo puede utilizarse para designar cualquier expresión. Esta arbitrariedad pertenece sólo a los símbolos. Los signos de los símbolos y sus relaciones mutuas determinan al objeto designado por una expresión compleja.
2) Existen expresiones que designan todos los procesos de los que es capaz la máquina.
3) Existen procesos para crear o modificar cualquier expresión arbitrariamente.

[35] Newell y Simon, La ciencia de la computación, en Boden, Margaret Filosofía de la inteligencia artificial, F.C.E., México, 1994, p.127.
[36] *Idem*
[37] *Idem*

4) Las expresiones, una vez creadas, seguirán existiendo hasta que sean suprimidas o modificadas explícitamente.
5) Un sistema puede contener, esencialmente, un número indefinido de expresiones.

La hipótesis de los sistemas de símbolos físicos.

Esta hipótesis afirma que un sistema de símbolos físicos cuenta con los medios necesarios y suficientes para realizar actos de inteligencia general.

Un acto de "inteligencia general" expresa el mismo nivel de inteligencia que se observa en la actividad humana: en cualquier situación real se cuenta con un comportamiento adecuado para los fines del sistema y para las exigencias ambientales dentro de ciertos límites de complejidad y rapidez.

La hipótesis de los sistemas de símbolos físicos es una ley de estructura cualitativa que especifica una clase general de sistemas entre los cuales pueden encontrarse los que son capaces de actos inteligentes.

El desarrollo de la hipótesis de los sistemas de símbolos.

La hipótesis tiene su origen en los trabajos de Frege, y de Whitehead y Russell sobre la formalización de la lógica, en la que ésta asume los conceptos fundamentales de las matemáticas y donde se establecieron las bases para las nociones de prueba y deducción. Surgió así la lógica matemática. De ella se derivó un punto de vista denominado "juego de símbolos", es decir, un juego basado en signos sin sentido de acuerdo a reglas meramente sintácticas, carentes de todo significado. Se trataba de un sistema mecánico que se había apartado de toda relación con los significados y símbolos humanos. Es la etapa de manipulación de símbolos formales.

Esta actitud se refleja en el desarrollo de la teoría de la información, en la que Shannon había establecido un sistema que, según se decía, nada tenía que ver con el significado y sólo era útil para la comunicación y la selección.

El desarrollo de las primeras computadoras digitales y la teoría de los autómatas concuerdan, básicamente, en una misma perspectiva. Así, podemos abordar el modelo de Turing para mostrarla.

La máquina de Turing consiste en dos memorias: una cinta infinita y un control de estados finitos. La cinta contiene datos, constituidos por unos y ceros. Posee además un conjunto de operaciones propias: lectura, escritura y búsqueda. Aunque no es una operación de datos, la operación de lectura dirige el control hacia un estado determinado en función de los datos bajo la cabeza lectora. Fundamentalmente, este modelo contiene lo que cualquier computadora en virtud de lo que puede hacer, aunque existen otras computadoras con distintas memorias y operaciones y que trabajan en distintas condiciones de espacio y tiempo. Particularmente, el modelo de la máquina de Turing implica tanto las nociones de lo que no puede computarse como de las máquinas universales (computadoras que pueden hacer todo lo que puede realizar una máquina).

Tanto el trabajo de Alan Turing como el desarrollo de la lógica matemática son de las concepciones más profundas dentro del procesamiento de información y se realizaron en los años treinta, antes del surgimiento de las computadoras modernas. Simultáneamente a la obra de Turing aparecieron, independientemente, los trabajos de Emil Post (las producciones) y de Alonzo Church (las funciones recursivas). Ambos sistemas logísticos llegaron a resultados análogos acerca de la irresolubilidad y la universalidad, los cuales implicaron la equivalencia de los tres sistemas. En efecto, se trataba de intentos independientes de definir la clase más general de sistemas de procesamiento de información.

No existe en estos sistemas, a primera vista, un concepto del símbolo como algo que *designa*. Los datos son simplemente hileras de unos y ceros. Este hecho es fundamental para reducir la computación a un proceso físico. En esta etapa se logró la mitad del principio de interpretación: una máquina podía funcionar a partir de una descripción. Se trata de la manipulación automática de símbolos formales.

A mediados de los años cuarenta, con la evolución de la segunda generación de máquinas electrónicas (después de la Eniac), surgió la noción de *programa almacenado*. Ahora los programas podían ser datos y operarse como tales. Aunque esta capacidad ya estaba implícita en el modelo de Turing, la idea se comprendió a partir del crecimiento de la memoria, lo cual hizo practicable la localización de programas en lugares internos.

Este concepto de programa almacenado concretiza la segunda parte del principio de interpretación, la cual afirma que pueden interpretarse los propios datos del sistema. Sin embargo, no contiene aún la noción de designación, esto es, de las relaciones físicas que sustentan el significado.

En 1956 surge el *procesamiento de listas*. En este concepto, los contenidos de las estructuras de datos eran ahora símbolos, es decir, patrones que designaban, que tenían referentes. Las listas contenían direcciones de acceso a otras listas; de ahí la noción de estructuras de listas.

El procesamiento de listas implica ser una estructura de memoria dinámica, en contraste con las estructuras fijas de las máquinas precedentes. El conjunto de operaciones se expandió al incorporar aquéllas que construían y modificaban la estructura además de las que cambiaban su contenido.

Además, con el procesamiento de listas se hizo patente la abstracción básica de que una computadora consiste en un conjunto de tipos de datos y de operaciones adecuadas para manipularlos de acuerdo a una determinada aplicación, independientemente de la máquina implícita.

Finalmente, el modelo de listas implicó un modelo de designación, tal como se utiliza actualmente en la ciencia de la computación. A partir de entonces la designación y la estructura simbólica dinámica fueron características definitorias dentro de todo un marco conceptual más amplio, conformado por la unión de la computabilidad, la factibilidad física (por múltiples tecnologías), la universalidad y la representación simbólica de procesos (interpretabilidad). Cabe mencionar que con la creación del LISP (*LISt Processing language interactiv*), por McCarthy, en 1959-1960, se completó el acto de abstracción, ya que representaba un nuevo sistema formal general que no requería de su incorporación en máquinas concretas.

Pruebas de la hipótesis.

Las pruebas de la hipótesis que afirma que los sistemas de símbolos físicos son capaces de actos inteligentes, y que la actividad inteligente general demanda un sistema de símbolos físicos se ha de basar en la experiencia. En efecto, tratándose, esta hipótesis, de una generalización empírica y no de un teorema, no se conoce una demostración puramente lógica que haga patente la relación existente entre sistemas de símbolos físicos e inteligencia.

Con el surgimiento de la noción del sistema de símbolos físicos se da inicio al desarrollo de la inteligencia artificial como un campo de la ciencia de la computación. A partir de entonces se han hecho evidentes dos tipos de pruebas empíricas que pueden ilustrar la hipótesis de los sistemas de símbolos físicos. El primer tipo se refiere a la *suficiencia* de los sistemas de símbolos físicos para generar inteligencia; el segundo, a la *necesidad* de contar con un sistema de símbolos físicos donde se exhibe inteligencia. La primera se considera, en general, dentro del campo de la inteligencia artificial; la segunda, dentro de la psicología cognitiva.

La construcción de sistemas inteligentes.

Identificar una enfermedad y después buscar el germen, fue el paradigma básico para el surgimiento de la teoría de los gérmenes de las enfermedades. Análogamente, dentro de la investigación en inteligencia artificial, identificar tareas que requerían inteligencia motivó la construcción de programas para computadoras digitales que pudieran realizar esas tareas. Se consideraron tareas sencillas y bien estructuradas tales como rompecabezas, juegos, programación y asignación de recursos y tareas de inducción simple.

Gradualmente se fue enfocando la investigación hacia niveles superiores de desempeño en ciertos dominios de tareas. Ilustran esto los diferentes niveles abarcados por distintos programas de ajedrez, en el dominio de juegos.

Otro dominio, en constante expansión, dentro del cual se ha intentado lograr actos inteligentes se refiere a la construcción de sistemas que manejan y comprenden lenguajes naturales, sistemas para interpretar escenas visuales, para la coordinación de manos y ojos, para el diseño, para escribir programas de computadora, para entender mensajes hablados, etc. Hasta el momento no se ha vislumbrado límite a la hipótesis.

Más allá de una simple acumulación de ejemplos de sistemas inteligentes en dominios de tareas específicos, estos programas tienen en común el buscar mecanismos que posean generalidad y componentes comunes a los programas en base a las tareas que realizan. La búsqueda se ha dirigido hacia una caracterización más completa que tipifica los sistemas particulares de símbolos físicos. Se trata de un segundo nivel de especificidad: la hipótesis de la búsqueda heurística.

La búsqueda de generalidad motivó el diseño de programas tendientes a separar los mecanismos para resolver problemas generales de los requisitos de los dominios de tareas específicas. El *General Problem Solver* (GPS) fue probablemente el primero de ellos, y de éste se han derivado otros, tales como PLANNER Y CONNIVER. Por otra parte, la búsqueda de componentes comunes según los dominios de tareas ha dirigido las investigaciones al diseño de esquemas generalizados de representación de metas y planes, a métodos para construir redes de distribución, a procedimientos para el control de árboles de búsqueda, a mecanismos de igualación de patrones, a sistemas de análisis gramatical de lenguajes y a mecanismos para representar secuencias de tiempo cronológico y verbal, movimiento y causalidad. Se trata de componentes básicos que han permitido ensamblar modularmente sistemas inteligentes.

En síntesis, la construcción de sistemas para desempeñar tareas en dominios específicos ha originado investigaciones tendientes a la comprensión de los mecanismos comunes de esos sistemas.

La construcción de modelos del comportamiento simbólico humano.

Según la hipótesis del sistema de símbolos, el comportamiento simbólico del hombre obedece a que éste posee las características de un sistema de símbolos físicos. Debido a esto, la modelación del comportamiento humano mediante sistemas de símbolos se dirige a la prueba de la hipótesis y las investigaciones en inteligencia artificial van de la mano con las de la psicología del procesamiento de información.

La explicación del comportamiento inteligente del hombre en términos de sistemas de símbolos ha privilegiado a la teoría del procesamiento de información como uno de los principales enfoques de la psicología cognitiva. Específicamente, los trabajos se han dirigido al diseño de modelos de manipulación de símbolos principalmente en las áreas de resolución de problemas, adquisición de conceptos y memoria de largo plazo.

Las investigaciones en psicología del procesamiento de información se conducen en dos vertientes principales. La primera se refiere a la observación de la conducta y la experimentación acerca del comportamiento humano inteligente. La segunda se dirige a la modelación del comportamiento humano

observado mediante la programación de sistemas de símbolos. Esta última en relación paralela a los trabajos en inteligencia artificial. Las observaciones y experimentos psicológicos en torno al análisis de los protocolos que los seres humanos siguen en la resolución de problemas, han permitido formular hipótesis sobre los procesos simbólicos que utilizan los sujetos y han aportado ideas importantes para la estructuración de los mecanismos básicos del GPS.

En esta alianza con la psicología se ha hecho patente el carácter empírico de la ciencia de la computación. Ambas ciencias se retroalimentan tanto para la verificación de los modelos de simulación del comportamiento humano, como en la aportación de nuevos esquemas de construcción de sistemas de símbolos físicos a partir de la experimentación.

No existen suposiciones específicas rivales de la hipótesis de los sistemas de símbolos que expliquen la actividad inteligente. La mayor parte de estos intentos han surgido de la psicología, desde el conductismo a la teoría de la *Gestalt*. Sin embargo, no han demostrado la suficiencia de sus mecanismos para explicar el comportamiento inteligente en tareas complejas. La vaguedad de sus teorías las hace fácilmente interpretables desde el punto de vista de la hipótesis de los sistemas de símbolos.

Simbolismo: el hombre y la computadora.

Cassirer, en su Antropología Filosófica[38], expone el núcleo teórico de su concepción del hombre. El animal vive adaptado y coordinado a su medio ambiente. Por un "sistema receptor" recibe los estímulos externos y, mediante un "sistema efector" reacciona ante ellos. El hombre, en cambio, descubrió un nuevo modo de adaptarse al ambiente. Entre los sistemas receptor y efector, intercala un eslabón que Cassirer llama "sistema simbólico", el cual transforma la totalidad de la vida humana. Comparado con el animal, el hombre vive en una nueva dimensión de la realidad. En las reacciones orgánicas la respuesta es directa e inmediata; en el hombre, la respuesta es demorada por un proceso lento y complicado del pensamiento.

[38] Cassirer, E., Antropología Filosófica. Introducción a una Filosofía de la Cultura, F.C.E, México 1963. 55-60

El hombre no vive, como el animal, en un puro universo físico, sino en un universo simbólico que él mismo ha creado, y del que forman parte el mito, el arte, la religión, la ciencia y la cultura en general. No se enfrenta a la realidad en forma directa, cara a cara, sino a través de sus propias creaciones. Está inmerso en sus propios símbolos, de tal modo que no puede ver nada, sino a través de esta urdimbre símbólica. Esto se da tanto en el orden teórico como en el práctico. En el orden práctico, los hechos no le afectan, sino sus propios temores, esperanzas e ilusiones.[39]

Si bien Cassirer considera que la definición del hombre como "animal racional" no ha perdido toda su fuerza, la racionalidad no caracteriza todas las formas de la actividad humana en toda su riqueza y diversidad. Todas estas formas son, más bien, simbólicas. De ahí que defina al hombre como "animal simbólico" o "simbolizante".

Cassirer distingue entre "signo" y "símbolo". En la conducta animal se da una serie de signos y señales, de hechos anunciadores de otros, pero carentes de sentido conceptual, que caracteriza al símbolo. Sostiene que una señal forma parte del mundo físico, mientras que el símbolo pertenece al mundo humano del sentido. Las señales son "operadores" y los símbolos son "designadores".

Sólo el hombre tiene inteligencia e imaginación simbólicas, que se distingue de la "inteligencia práctica", característica de los animales, y que está basada en signos. El símbolo humano se caracteriza por su variabilidad, mientras que el signo está relacionado con el objeto de un modo fijo y único. El símbolo no es rígido e inflexible sino móvil. En un cierto estado avanzado de desarrollo intelectual y cultural, el hombre se da cuenta de esa movilidad del símbolo que, en un estado primitivo, es considerado como propiedad de la cosa.

Cassirer distingue entre el pensamiento relacional y el simbólico. El primero presupone la existencia previa de un sistema de símbolos. El mero darse cuenta de que existen relaciones, no configura un acto intelectual de pensamiento lógico o abstracto, puesto que aún en la percepción simple se captan relaciones estructurales y no objetos aislados. El hombre, además,

[39] *Idem.* 57-59.

posee la capacidad de aislar relaciones y de considerarlas en su sentido abstracto. La geometría es el mejor ejemplo de esta capacidad. Sin el simbolismo, la vida del hombre estaría reducida a sus necesidades biológicas e intereses prácticos. El espacio geométrico hace abstracción de la variedad de datos de los sentidos; es un espacio abstracto, homogéneo y universal. De aquí surge la concepción de un orden cósmico único, sistemático.

Mientras que el espacio está relacionado al mundo físico, el tiempo lo está a nuestra experiencia interna y, por ende, a la memoria. Más allá de una simple reproducción de hechos anteriores, la memoria ordena lo recordado dentro de un espacio abstracto y serial. Se trata de un proceso creador que reordena, organiza y sintetiza los recuerdos por parte de la "memoria simbólica" del hombre. Por otra parte, la relación del hombre con el futuro es también de tipo simbólico y se corresponde con su pasado y presente. De aquí surge la capacidad del hombre para distinguir entre realidad y posibilidad. Gracias a ella el hombre determina su lugar en la cadena general del ser[40]. Para el pensamiento humano, que es simbólico, resulta indispensable distinguir entre lo real y lo posible, entre cosas reales e ideales. Un símbolo no posee existencia real, sino sentido. De la relación entre esta distinción realidad-posibilidad y el pensamiento simbólico del hombre, se genera el pensamiento hipotético.

Los hechos hipotéticos son fundamentales para el desarrollo de la ciencia. Aquellos en que se apoya un principio científico son anticipados como posibles antes de ser reales, pues la ciencia no se constituye por acumulación de hechos tomados al azar. Ese elemento teórico anticipado es un símbolo o una relación de símbolos. Sólo el hombre, dado su pensamiento simbólico, es capaz de construir una teoría científica. Así, la matemática, por ejemplo, no es una teoría de las cosas, sino de símbolos.

En el ámbito de la Inteligencia Artificial, no parece difícil entender que ésta, a través de los dispositivos creados por el hombre, posea la característica de mediar su relación con el mundo externo a través de una estructura simbólica. Así, la computadora atiende a mensajes externos y reacciona ante ellos de acuerdo al sistema de símbolos con el que ha sido dotada.

[40] Cassirer, E. *Op.Cit.*, p. 111.

Mientras que el hombre es capaz de formular hipótesis gracias a su capacidad de distinguir entre realidad y posibilidad –siguiendo a Cassirer-, un sistema de símbolos aplicado a la IA contiene en sí mismo toda la serie de posibilidades anticipadamente visualizadas, dentro de un cierto margen espacio-temporal. Los símbolos son la realidad inmediata y única de un programa de IA *desde el punto de vista humano*. Esto es, desde el punto de vista del dispositivo artificialmente inteligente -suponiendo que estuviera dotado de alguna suerte de conciencia-, tales símbolos se confunden con las cosas; los símbolos son las cosas. Esto significa que, al cosificarse, los símbolos devienen signos y, por tanto, dejan de tener contenido conceptual. El dispositivo trata con los signos –o símbolos cosificados- como su materia concreta y ésta constituye el ámbito de funcionalidad del sistema. Es con los signos, con ese tipo de objetos, con los que trata un programa de IA.

Un sistema de IA consiste también en una estructura relacional que integra las interdependencias de los objetos. Esta estructura determina el orden y la organización de funciones y objetos, de la que proviene la sistematicidad del dispositivo. Los sistemas de aprendizaje inductivo requieren de información previamente estructurada referente a objetos y relaciones entre ellos (*training set*), a partir de la cual el sistema pueda generar reglas de clasificación que, posteriormemente, le permitan generalizar los patrones determinados hacia situaciones desconocidas y posibles.[41] Estas funciones proveen al sistema con una cierta capacidad de anticipación, de antelación a la posibilidad, de construcción de elementos apropiados que le permitan asimilar sucesos que, desde el punto de vista humano, estarían dentro de lo hipotético. Esta capacidad de asimilar nuevos sucesos no previstos dentro de la base de conocimiento del sistema, es la forma que tiene el sistema de *dar sentido* a los hechos, a los fenómenos.[42] Notemos que en esta preparación para asimilar hechos posibles por parte del sistema, tienen lugar dos funciones que, de algún modo, hacen referencia a dos funcionalidades de la conciencia: a) la construcción de esquemas de recepción mediante generalización, habilita al sistema el anticiparse a situaciones posibles y a la vez reales. Es decir, esta función es una suerte de acto de dar sentido, según lo expone Husserl. Y, b) la

[41] V. *Supra*. "Aprendizaje y Razonamiento".
[42] V. *Infra*. "Conceptos básicos de la fenomenología de la computación".

dirección de la conciencia hacia lo posible, lo hipotéticamente realizable, tiene un paralelismo en el sistema en cuanto a la clasificación, a la agrupación por similaridad mediante la que construye nuevos esquemas de asimilación de información. En particular, según se ha visto más arriba, un sistema de IA puede enfocarse a la similaridad con situaciones previas para hacer predicciones útiles.[43]

[43] V. *Supra.* "Razonamiento".

V. LA FUNCIÓN OPERATORIA DE LA INTELIGENCIA

La búsqueda heurística.

Los sistemas de símbolos físicos resuelven problemas utilizando procesos de búsqueda heurística. La relación entre este tipo de procesos y la actividad inteligente puede enunciarse mediante la siguiente hipótesis.

Hipótesis de la búsqueda heurística.
Las soluciones a problemas se representan como estructuras de símbolos. Un sistema de símbolos físicos ejerce su inteligencia en la resolución de problemas mediante la búsqueda, es decir, generando y modificando progresivamente las estructuras de símbolos hasta producir una estructura de solución.[44]

La utilización de procesos de búsqueda heurística responde a las limitaciones de recursos de los sistemas de símbolos físicos: ejecutan un número finito de procesos en un número finito de pasos y en un intervalo finito. Esta limitación es relativa a la complejidad de las situaciones a las que se enfrentan los sistemas, trátese de computadoras o de seres humanos. En general, se puede considerar a un sistema de símbolos como un mecanismo serial que ejecuta una tarea a la vez aunque, en realidad, ejecuta una pequeña cantidad de procesos en un intervalo breve de tiempo.

La resolución de problemas.
La capacidad para resolver problemas se considera, en general, como un primer indicador de que un sistema tiene inteligencia. De ahí los esfuerzos por construir y comprender sistemas de resolución de problemas en la historia de la inteligencia artificial.

Enunciar un problema consiste en designar:

1) una *prueba* para una clase de estructuras de símbolos, que significan las soluciones al problema, y

[44] Boden, Margaret *Op.Cit.* 137.

2) un *generador* de estructuras de símbolos, que serían las soluciones potenciales.

De esta manera, resolver un problema es generar una estructura (2), que satisfaga una prueba (1).

Tener un problema significa saber lo que necesitamos hacer (la prueba), y no saber de inmediato cómo hacerlo –el generador de estructuras de símbolos no produce alguna que satisfaga la prueba. Cuando un sistema de símbolos resuelve problemas es porque puede generar y probar.

Para que un generador de estructuras para un problema pueda operar se requiere que exista un espacio del problema, es decir, un espacio de estructuras de símbolos en el que puedan representarse las situaciones del problema. El generador puede entonces transformar una situación en otra, esto es, generar situaciones posibles, hasta encontrar aquélla que satisfaga la prueba de definición del problema.

La extracción de información a partir del espacio del problema.

Considérese algún conjunto de estructuras de símbolos, así como algún subconjunto que contiene soluciones a un problema dado. Supóngase que las soluciones están distribuidas aleatoriamente. Esto significaría que el generador de posibles soluciones únicamente podría funcionar aleatoriamente. Generalizando esta situación, ningún sistema de símbolos exhibiría ni más ni menos inteligencia que cualquier otro para resolver el problema, aunque algunos darían más rápido con la solución de manera fortuita.

De aquí se desprenden algunas condiciones para que exista inteligencia en el proceso de generación de soluciones posibles. Primeramente, el espacio de estructuras de símbolos ha de mostrar cierto grado de orden y algún patrón. En segundo lugar, dicho patrón tendría que ser más o menos detectable. Finalmente, el generador debería comportarse según el patrón detectado. Esto significa que debe haber información en el espacio del problema y que el sistema de símbolos, particularmente el generador de soluciones potenciales, debe ser capaz de extraerla y utilizarla en el proceso de búsqueda.

Por ejemplo, sea el problema de resolver la siguiente ecuación:

$$Ax + B = Cx + D$$

La prueba del problema consiste en cualquier expresión de la forma x = E, tal que

AE + B = CE + D. Un generador podría ser cualquier proceso que sustituyera números en la segunda ecuación hasta conseguir una identidad. Éste no sería un generador inteligente.

Podrían utilizarse, en cambio, generadores que pudieran modificar la ecuación original mediante operaciones algebraicas que lo condujeran hacia la obtención de la forma de la solución, según la definición de prueba. Un procedimiento tal mostraría inteligencia:

$$Ax - Cx = D - B$$

$$(A - C)x = D - B$$

$$x = (D - B)/(A - C)$$

$$x = E$$

Se puede observar que cada una de las expresiones sucesivas no es generada independientemente, sino cuando se ha modificado la antecedente. Por otra parte, las modificaciones, lejos de ser azarosas, dependen de dos tipos de información. Durante el proceso de modificación algebraica se mantiene la información constante (A, B, C y D) y ésta es integrada a la estructura del propio generador. Por otra parte, la información que cambia en cada paso está dada en las estructuras cuya forma se mantiene entre la expresión original y la buscada. El primer tipo de información (las constantes) garantiza que sólo se genere un subconjunto pequeño de soluciones posibles, sin perder la forma de la solución a partir de ese subconjunto. Por su parte, el segundo tipo de información permite lograr la solución buscada en base a sucesivas aproximaciones, recurriendo a un análisis de medios y fines para dirigir la búsqueda.

Árboles de búsqueda.

Generalmente, los procesos de búsqueda en la solución de problemas se refieren a estructuras de ramas de múltiples posibilidades de soluciones parciales que han de probarse antes de llegar a la solución o al conjunto de soluciones reales. El ejemplo algebraico anterior consistía de un número de ramas igual a uno. En cambio, los programas de ajedrez, por ejemplo, producen árboles de búsqueda que constan de varios miles de ramas, aunque

no se trata, en estos casos, de generar soluciones propuestas, sino de evaluarlas (probarlas).

Los procesos de búsqueda, la generación sucesiva de estructuras de soluciones potenciales, muestra que la cantidad de búsquedas no es una medida de la cantidad de inteligencia que exhibe un sistema. Ésta se muestra en razón del nivel de selectividad y, por ende, de información. Si el sistema simbólico tiene suficiente información acerca de lo que pretende hacer, procede entonces directamente hacia su objetivo; pero si la información es poca o inadecuada, el número de búsquedas se incrementará exponencialmente.

Las formas de la inteligencia.

La inteligencia, por tanto, previene la explosión exponencial de una búsqueda en la medida en que el sistema de símbolos posea la selectividad que le permite generar soluciones "verosímiles". La consecuencia de un proceso de generación selectivo es la disminución en la velocidad de ramificación sin evitarla por completo. La orientación de la búsqueda se agudiza complementando la selectividad con técnicas de utilización de información. Algunas de esas técnicas son las siguientes.

En la búsqueda heurística en serie, la pregunta fundamental es: ¿qué se hace a continuación? En el caso de búsqueda en árbol la pregunta tiene dos componentes: ¿a partir de qué nodo hay que emprender la búsqueda? y, ¿qué dirección hay que tomar a partir de ese nodo? Para responder a la primera pregunta, en la búsqueda en árbol, es útil la información referente a las distancias relativas entre los nodos y la meta. Así, la técnica de la primera mejor búsqueda exige continuar la búsqueda a partir del nodo más próximo a la meta. En cuanto a la pregunta acerca de la dirección de búsqueda, es útil la información que permite detectar diferencias entre la estructura actual, en un nodo determinado, y la estructura de la meta según indica la prueba de la solución y, consecuentemente, reduciendo gradualmente las diferencias. Esta técnica se conoce como análisis de medios y fines, y es de capital relevancia en la estructura del GPS. Los esfuerzos encaminados al perfeccionamiento de estas técnicas han aportado importantes ideas generales en la investigación en IA.

Métodos "fuertes" y "débiles".

Estas técnicas son llamadas débiles dado que se dedican más a controlar la expansión que a prevenirla. Son técnicas propias de los sistemas de símbolos cuya información y estructura de su espacio de problema son inadecuados para evitar por completo la búsqueda. Es útil observar la diferencia entre una situación sumamente estructurada, como un problema de programación lineal, con una mucho menos estructurada de problemas combinatorios como el de un agente viajero. (La falta de estructuración se refiere a la insuficiencia de una teoría pertinente acerca de la estructura del espacio del problema).

La resolución de problemas de programación lineal requiere una cantidad considerable de cómputos, aunque no se ramifica, ya que cada paso dado lleva, tarde o temprano, a la solución. Es decir, la generación de soluciones se da linealmente. En cambio, la resolución de problemas combinatorios o la demostración de teoremas rara vez evitan la búsqueda en árboles y requieren utilizar alguna de las técnicas ya descritas.

Algunas corrientes en la investigación de resolución de problemas en IA han seguido líneas distintas a las que aquí se han mencionado. Tal es el caso del trabajo realizado sobre sistemas de demostración de teoremas. Las ideas aportadas por las matemáticas y la lógica han sido determinantes. Así, la oposición al uso de la heurística, en este trabajo, estaba fundamentada en la imposibilidad de probar las propiedades de integridad en muchos tipos de generadores selectivos. Sin embargo, comenzó a usarse la heurística selectiva cuando la explosión combinatoria de sus árboles de búsqueda lo exigieron.

Hay diferentes opiniones en cuanto a la eficiencia que ha mostrado la búsqueda heurística como mecanismo de resolución de problemas, dependiendo de los dominios de tareas que se consideren y de los criterios de suficiencia adoptados. En general, puede garantizarse el éxito tratándose de niveles básicos de aspiración; lo contrario, tratándose de altas aspiraciones. Existen mecanismos de alta eficiencia en materia de búsqueda heurística aplicados a problemas de investigación de operaciones, al igual que en materia de juegos como en ajedrez, donde estos mecanismos se comportan a nivel de aficionados competentes, en algunos ámbitos de demostración de teoremas y en muchas clases de acertijos. Aunque no se han alcanzado, ni remotamente, los niveles humanos, aún por los programas de reconocimiento visual, los que

comprenden mensajes hablados, o los robots que operan en espacio y tiempo reales, se ha acumulado una gran experiencia en torno a estas tareas.

A partir de la experiencia obtenida acerca del desempeño humano experto en tareas como el ajedrez, probablemente, un sistema que pueda igualar tal desempeño habrá de tener acceso a una gran cantidad de información semántica en su memoria. Por otra parte, la demostrada superioridad humana en tareas que exigen un gran componente perceptual puede atribuirse a que los ojos y oídos humanos procesan la información sensorial en forma paralela e integrada.

En todo caso, la calidad del desempeño depende tanto del contexto del problema a resolver como de los sistemas de símbolos utilizados para abordar tales contextos. Hasta ahora no ha sido posible formular teoremas acerca de la complejidad de los contextos reales, que muestren, en términos no empíricos, la magnitud del mundo real con relación a las habilidades de los sistemas de símbolos, dado que estos contextos no son lo suficientemente simples. Mientras esta situación prevalezca, la exploración tiene que ser empírica en torno a las características de dificultad que los problemas reales entrañan. Las áreas más estructuradas como la de la programación lineal han aportado elementos teóricos para perfeccionar los mecanismos heurísticos, más que para proporcionar un análisis más estructurado de la complejidad.

El análisis efectuado hasta aquí en torno a la inteligencia, permite equipararla con la capacidad de extraer y utilizar información sobre la estructura del espacio del problema, con el objeto de generar la solución en el menor tiempo posible. Para mejorar la capacidad de resolver problemas se han intentado tres modos básicamente.

El uso no local de la información.

Las investigaciones en este campo han arrojado algunas observaciones en cuanto a la acumulación de información durante los procesos de búsqueda en árboles. Por lo general, esta información sólo se utiliza *localmente*, esto es, en el nodo específico donde se generó. Así, la información referente a una posición determinada en el ajedrez suele ser útil únicamente para evaluar dicha posición y no otras con características similares. En consecuencia, los mismos hechos son redescubiertos en diferentes nodos del árbol de búsqueda. La

solución a esta deficiencia no estaría en el uso generalizado de la información extraída en un contexto específico, tal como en un nodo, ya que, probablemente, ésta sólo sea válida dentro de un rango limitado de contextos con características similares. Aunque aún no es oportuno hacer una evaluación sobre los intentos que ya se han realizado para transportar información de su contexto de origen a otros apropiados, el proyecto resulta prometedor. La línea de investigación de Berliner (1975) consiste en un análisis causal para determinar el rango de validez de un determinado bloque de información. De acuerdo con este objetivo, si pudiera remontarse la vulnerabilidad de una posición de ajedrez a la posición que la generó, puede esperarse la misma vulnerabilidad en otras posiciones derivadas de la misma jugada.

Una tendencia para lograr que la información esté globalmente disponible es la que ha incorporado el sistema para entender mensajes hablados HEARSAY. Mediante este sistema se busca el reconocimiento, en paralelo, de cadenas de habla en distintos niveles: fonemático, léxico, sintáctico y semántico. La información proporcionada por cada búsqueda es transcrita a una especie de pizarra común que puede ser consultada por todas las fuentes. Tal información puede utilizarse para eliminar hipótesis, evitando así tener que ser buscadas por alguno de los procesos. Así, el uso no local de información pretende eficientar los sistemas de resolución de problemas.

Sistemas de reconocimiento semántico.

Tendiente a este mismo objetivo es el diseño de mecanismos para abastecer al sistema de un *corpus* de información semántica referente a la esfera de acción de las tareas que desempeña. Un ejemplo de ello lo aportan los maestros del ajedrez que son capaces de reconocer diversos patrones en el tablero, lo cual les permite realizar las jugadas adecuadas a una determinada situación reconocida.

La posibilidad de sustituir la búsqueda por el reconocimiento es motivada por el hecho de que un determinado modelo que tenga suficiente relación con la estructura del espacio del problema, puede contener una cantidad de información altamente relevante. Cuando las estructuras son irregulares, de tal manera que no pueden ajustarse a una descripción matemática sencilla,

entonces el nivel de inteligencia del sistema puede elevarse mediante el conocimiento de una suficiente cantidad de modelos pertinentes. Los sistemas actuales aún son deficientes en cuanto al reconocimiento de patrones, incluso contando con suficiente información semántica.

Selección de representaciones adecuadas.

La selección de un espacio de problema adecuado, es otra línea de investigación que puede evitar o, cuando menos, reducir la búsqueda en árbol. Un ejemplo de selección de representaciones es el problema del tablero de damas recortado (fig. 1). Se trata de un tablero de damas donde, en vértices opuestos, se han suprimido dos casillas (una en cada vértice). Se pretende cubrir exactamente las demás casillas, utilizando para ello fichas de dominó. Cada ficha de dominó cubre dos casillas. ¿Tiene solución este problema? Intentando todos los arreglos posibles puede demostrarse que no hay solución. Una alternativa de demostración más rápida consiste en notar que las esquinas suprimidas son del mismo color, de modo que el tablero mutilado contiene dos cuadros menos de un color que del otro. Como cada ficha de dominó cubre dos cuadros de distinto color, y cualquier grupo de fichas debe cubrir la misma cantidad de cuadros de cada color, el problema no tiene solución. Un sistema que llegara a esta conclusión exhibiría un alto grado de inteligencia.

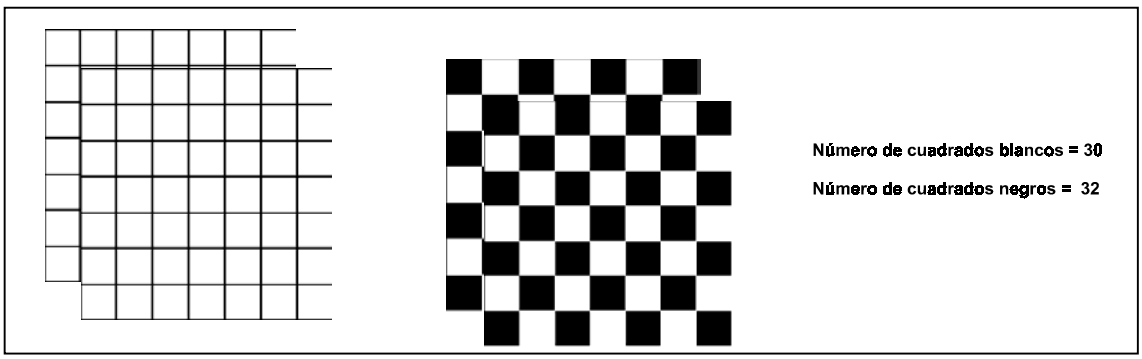

Fig.1 Tres representaciones de un tablero de damas recortado.

En realidad, la selección de representaciones adecuadas equivale a desplazar la búsqueda de un espacio de soluciones posibles a un espacio de representaciones posibles. De aquí surge una amplia línea de investigación referente al descubrimiento de las leyes de estructura cualitativa que regulan las representaciones.

Mecanismos de representación.

La complejidad de los problemas con los que se enfrenta la inteligencia artificial requiere de una gran cantidad de conocimiento así como de mecanismos que permitan manipularlo a fin de obtener soluciones a los nuevos problemas. En todas las representaciones se manejan dos tipos de entidades:

- Hechos: verdades en cierto mundo. Lo que se pretende representar.
- Representaciones de hechos en un cierto formalismo. Las entidades que los sistemas realmente son capaces de manejar.

Entre los hechos y las representaciones debe existir correspondencia (fig. 2). Una representación hacia delante significa una correspondencia desde los hechos y sus representaciones, mientras que la representación inversa, hacia atrás, es una correspondencia que va de las representaciones a los hechos.

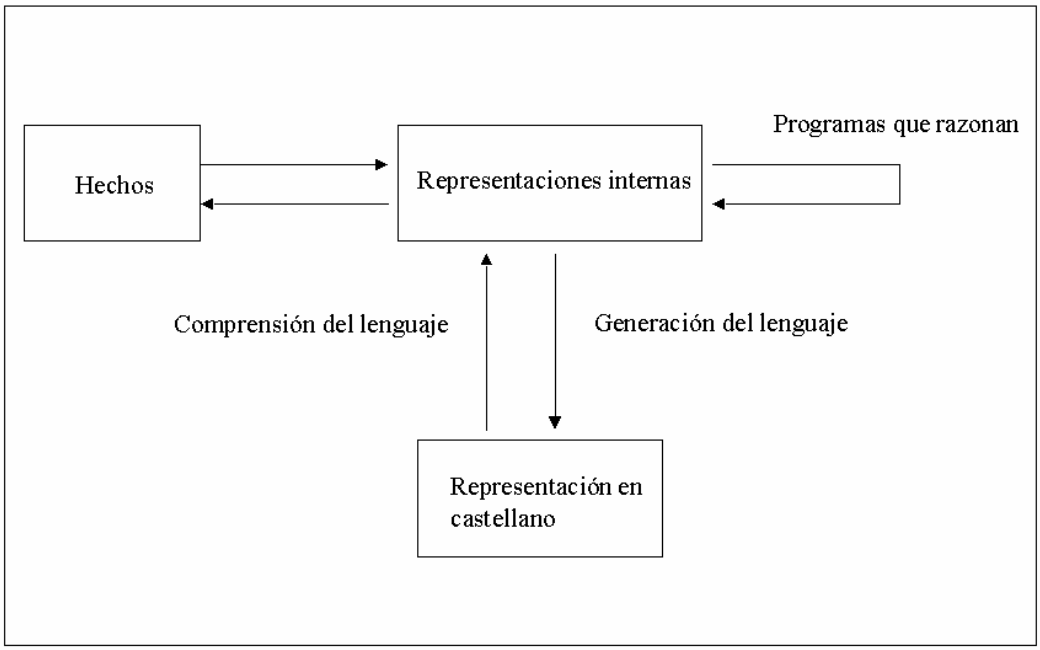

Fig. 2 Correspondencia entre los hechos y las representaciones.

Las frases del lenguaje natural son también una representación de hechos. El uso del lenguaje natural como medio de intercambio de información con el programa hace necesarias ciertas funciones de correspondencia, de modo que transformen dichas frases en representaciones internas y viceversa. En la fig. 2 se muestran las relaciones entre estos tres tipos de entidades.

Un sistema de representación del conocimiento debe poseer las siguientes características:

- Suficiencia en la representación: la capacidad de representar todos los tipos de conocimiento necesarios en el dominio particular de que se trate.
- Suficiencia deductiva: capacidad para manipular las estructuras de la representación para obtener nuevas estructuras que tengan correspondencia con la nueva información deducida.
- Eficiencia deductiva: capacidad de incorporar información adicional en las estructuras de conocimiento con objeto de que los mecanismos de inferencia sigan las direcciones con mayor probabilidad de éxito.

- Eficiencia en la adquisición: capacidad de adquirir nueva información con facilidad. Un ejemplo de esto es una base de datos en la que un usuario inserta directamente el conocimiento. Idealmente, un programa podría ser capaz de incorporar la información por sí mismo.

Aún no hay un sistema que optimice estos requisitos y que sea aplicable a cualquier tipo de conocimiento. Algunas técnicas de representación del conocimiento se describen a continuación.

Conocimiento relacional simple.

Es el modo más sencillo de representar los hechos declarativos. Se elabora mediante un conjunto de relaciones. Su simplicidad radica en la escasa información deductiva que proporciona. Sin embargo, este tipo de conocimiento puede servir como entrada a otros mecanismos de inferencia más elaborados. Esta representación puede ejemplificarse en la fig. 3.

Jugador	**Altura**	**Peso**	**Batea-Lanza**
J1	1.80	90	Derecha-Derecha
J2	1.68	70	Derecha-Derecha
J3	1.83	88	Izquierda-Izquierda
J4	1.75	78	Izquierda-Izquierda

Fig. 3 Conocimiento relacional simple

Si se especifica un conjunto de reglas para determinar qué bateador se debe colocar frente a un lanzador dado (basándose, por ejemplo, en el hecho de que sea zurdo o diestro), esta relación podrá proporcionar parte de la información para estructurar esas reglas.

Conocimiento heredable.

El conocimiento acerca de los objetos, sus atributos y sus valores no necesariamente es tan simple como el mostrado en el ejemplo. La representación básica puede ser extendida mediante algunos mecanismos de inferencia que operen sobre la estructura de la representación. Tal estructura debe entonces diseñarse de acuerdo con el mecanismo de inferencia de que se trate. La *herencia de propiedades* es una forma de inferencia donde los

elementos de una clase heredan los atributos y valores de otras clases más generales que los incluyen.

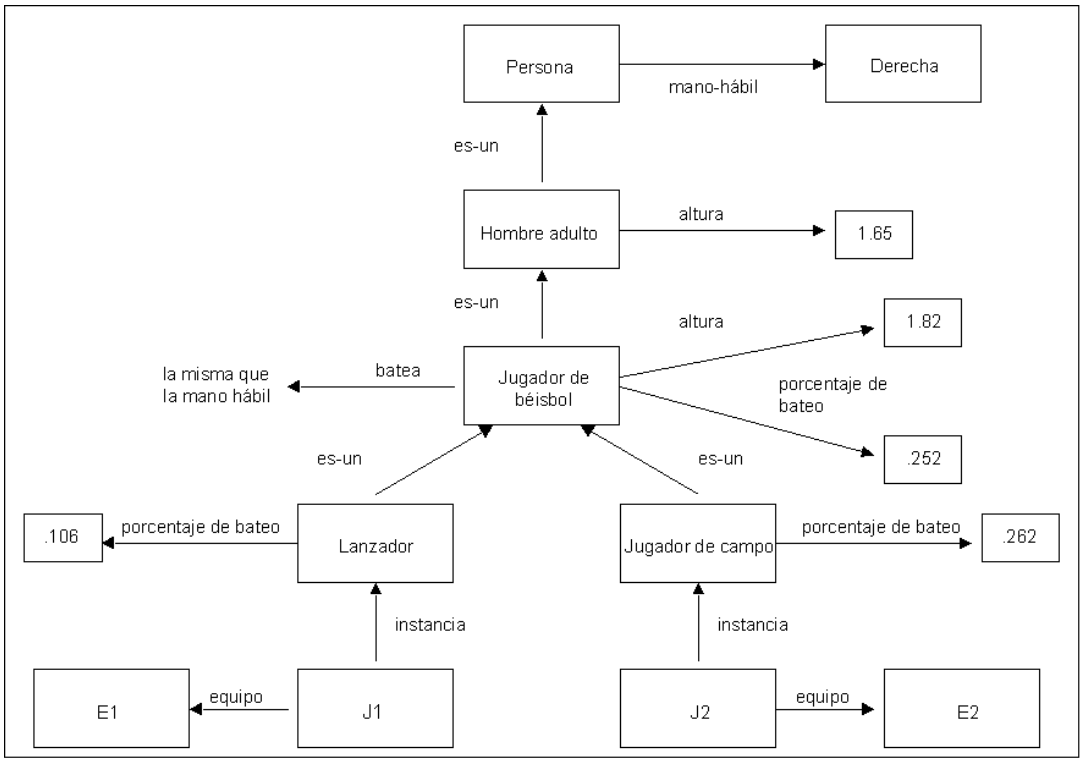

Fig. 4 Conocimiento heredable

Una estructura organizada en clases puede ser la que se muestra en la fig. 4, la cual se refiere a información acerca de béisbol. Las líneas representan atributos; los nodos recuadrados representan objetos y valores de los atributos de los objetos.

Estos valores, a su vez, también pueden verse como objetos con atributos y valores, y así sucesivamente. Las flechas conectan los objetos con sus valores a través de los atributos correspondientes. La estructura de la figura se denomina del tipo *de ranura y relleno (slot-and-filler)*[45]

[45] Rich, Elaine y Knight, Kevin. Inteligencia Artificial, p.124, McGraw-Hill/Interamericana de España, 1994

Lógica de predicados.

El lenguaje de la lógica es uno de los mecanismos concretos para representar el conocimiento. El formalismo lógico proporciona un método poderoso para mostrar información a partir de otra previa: la deducción matemática. Es un medio de obtener respuestas a ciertas preguntas y solucionar problemas.

La demostración automática de problemas ha sido uno de los primeros dominios de tareas en que se aplicó la IA, con la aportación de técnicas matemáticas. Pero estas técnicas se han extendido más allá del ámbito matemático. De hecho, la matemática no se distingue mucho de cualquier otra labor intelectual compleja que implique la necesidad de contar con mecanismos confiables de deducción, así como de técnicas heurísticas con la correspondiente dosis de información.

El uso de la lógica de predicados como representación del conocimiento puede mostrarse con el siguiente ejemplo. Sean las sentencias:

1. Marco era una persona.
2. Marco era un pompeyano.
3. Todos los pompeyanos eran romanos.
4. César fue un gobernante.
5. Todos los romanos o eran leales a César o lo odiaban.
6. Cada persona es leal a otra persona.
7. La gente sólo intenta asesinar a los gobernantes a los que no es leal.
8. Marco intentó asesinar a César.

Estas sentencias pueden representarse como un conjunto de fórmulas bien formadas de lógica de predicados de la siguiente forma:

1. *persona(Marco)*

   Aunque esta representación muestra el hecho fundamental de que Marco era persona, se pierde el concepto de tiempo (pasado). Esta omisión puede ser aceptable o no según el uso que se le dé a este conocimiento.

2. *pompeyano(Marco)*

3. $\forall x : pompeyano(x) \supset romano(x)$

4. $gobernante(César)$

En esta fórmula se pasa por alto el hecho de que los nombres propios no suelen referirse a un único individuo. En la mayoría de los casos, decidir a quién se refiere una frase particular de entre un grupo de personas con el mismo nombre, implica una gran cantidad de conocimiento y razonamiento.

5. $\forall x : romano(x) \supset leal(x, César) \vee odia(x, César)$

6. $\forall x : \exists y : leal(x, y)$

7. $\forall x : \forall y : persona(x) \wedge gobernante(y) \wedge intenta\_asesinar(x, y) \supset \neg leal(x, y)$

El enunciado, en lenguaje natural, es ambiguo. Puede significar que a los únicos gobernantes a los que la gente quiere asesinar son aquéllos a los que no son leales (interpretación deseada en esta caso), o bien puede significar que lo único que la gente quiere es asesinar a los gobernantes a los que no son leales.

8. $Intenta\_asesinar(Marco, César)$

Estos ejemplos muestran la dificultad de convertir enunciados en lenguaje natural a fórmulas bien formadas de la lógica de predicados.

Si se quisiera utilizar los enunciados anteriores para responder a la pregunta de si Marco era leal a César, se podría llegar al procedimiento siguiente.

$$persona(Marco)$$
$$\downarrow \quad (8)$$
$$persona(Marco) \wedge intenta\_asesinar(Marco, César)$$
$$\downarrow \quad (4)$$
$$persona(Marco) \wedge gobernante(César) \wedge intenta\_asesinar(Marco, César)$$
$$\downarrow \quad (7, \text{sustitución})$$
$$\neg leal(Marco, César)$$

Los principales problemas involucrados en la conversión de frases del lenguaje natural a sentencias lógicas son los siguientes:

- Muchos enunciados en lenguaje natural son ambiguas y no existe un procedimiento que garantice que se ha elegido la interpretación correcta.
- Normalmente hay más de una forma de representar la información. La elección de una u otra forma depende del uso que se quiera hacer de la información contenida en un enunciado.
- Por lo general, en un conjunto de enunciados dado no está contenida toda la información necesaria para razonar sobre el tema en cuestión. Normalmente se requiere un conjunto adicional de enunciados que no se especifican por ser, aparentemente, obvias.

Puede presentarse además el problema de no conocer el objetivo a deducir. Así, en el caso anterior, si no se especificara el objetivo, un sistema de símbolos no podría decidir si demostrar la proposición de que Marco no era leal a César o de que sí lo era. Podría intentarse razonar a partir de las proposiciones conocidas y ver qué respuesta se obtiene. Pero es probable que el factor de ramificación provoque el no llegar a ninguna respuesta en un lapso razonable. O bien podría intentarse usar algunas técnicas heurísticas para decidir acerca de la respuesta más probable para luego tratar de demostrarla. Pasado algún tiempo, si no se muestra solución alguna, entonces se pasaría a otra respuesta probable, y así sucesivamente. Esta técnica se conoce como "generación y prueba".

La importancia de limitar el esfuerzo radica en la alta probabilidad de que un procedimiento de prueba no se detenga al enfrentarse con situaciones distintas a un teorema.

Desde el punto de vista computacional sería muy conveniente contar con un procedimiento de demostración que simplificara, en una sola operación, los distintos procesos implicados en el razonamiento con sentencias lógicas.

Sistemas lógicos en inteligencia artificial.

Aprendizaje y Razonamiento.

Aprendizaje y razonamiento son aspectos de lo que se considera inteligencia. Los estudio de uno y otro dentro de la Inteligencia Artificial han estado separados históricamente, siendo el aprendizaje un tópico del aprendizaje en

las máquinas y de las redes neuronales y, el razonamiento, siendo tratado dentro de la línea clásica (o simbólica) de la IA. Sin embargo, aprendizaje y razonamiento son interdependientes en varios sentidos. El modelo llamado FLARE (*Framework for Learning and Reasoning*) nos muestra la naturaleza de algunas de tales interdependencias, combinando el aprendizaje inductivo en base al conocimiento previo, con el razonamiento proposicional.

Tanto la inducción como la deducción son procesos subyacentes a los agentes inteligentes. La inducción implica saltos intelectuales de lo particular a lo general. Juega un papel importante en la adquisición del conocimiento o aprendizaje. La deducción es - por otra parte - una forma de razonamiento con y sobre el conocimiento adquirido. No está relacionada con la generación de nuevos hechos, más bien establece relaciones de implicación entre hechos ya existentes. La deducción apunta hacia adelante en cuanto a las consecuencias de ciertas hipótesis existentes, o hacia atrás en cuanto a las condiciones necesarias para la aparición de ciertos sucesos o el logro de ciertas metas. La habilidad para razonar acerca de un dominio del conocimiento se basa, en general, en las reglas acerca de ese dominio que deben ser asimiladas de algún modo; y la habilidad para razonar con frecuencia guía la adquisición de nuevo conocimiento o aprendizaje.

El aprendizaje inductivo ha sido objeto de muchas investigaciones dirigidas al diseño de una variedad de algoritmos. En general, los sistemas de aprendizaje inductivo generan reglas de clasificación a partir de ejemplos. Típicamente, primero se presenta al sistema un conjunto de ejemplos (objetos, situaciones, etc.), también conocido como *training set*. Por lo general, los ejemplos se presentan en lenguaje atributo-valor y representan instancias predeterminadas de parejas de atributo-valor junto con su correspondiente clasificación. La meta del sistema es entonces descubrir conjuntos de suficientes características relevantes o reglas que clasifiquen apropiadamente los ejemplos del *training set* (convergencia), y que puedan extenderse a ejemplos no predefinidos (generalización).

Aunque las máquinas aún están bastante lejos de igualar los saltos inductivos de la calidad humana, los sistemas de aprendizaje inductivo han mostrado su utilidad en una amplia gama de aplicaciones en medicina (cáncer de mama, detección de hepatitis), otorgamiento de créditos bancarios, en la

defensa militar (discriminación entre minas y rocas), botánica (identificación de variedades de iris, detección de hongos venenosos) y otras.

El estudio del razonamiento deductivo se remonta cuando menos a los primeros filósofos griegos, tales como los estoicos y Aristóteles. Su formalización ha dado lugar a una variedad de lógicas, de la proposicional a la de predicados de primer orden, así como a varias extensiones de lógicas no monótonas. Muchas de esas lógicas han sido implementadas con éxito en sistemas artificiales (e.g., PROLOG, sistemas expertos). Éstos consisten básicamente en una base de conocimiento o reglas pre-codificadas, un conjunto dado de hechos (identificados ya sea como causas o consecuencias) y algún mecanismo de inferencia. Éste último lleva a cabo el proceso deductivo usando las reglas básicas y los hechos proporcionados. Muchos de esos sistemas han sido utilizados exitosamente en varios dominios, tales como el diagnóstico médico y la geología.

Uno de los mayores retos de los sistemas deductivos actuales es la *adquisición de conocimiento*, es decir, la construcción de las reglas básicas. Típicamente ésta se genera tal como el conocimiento de un dominio, el cual se extrae de los humanos expertos y luego se modela cuidadosamente en reglas. La adquisición del conocimiento es una tarea tediosa que presenta muchas dificultades tanto teóricas como prácticas. Si se logra obtener un adiestramiento suficientemente rico, entonces el aprendizaje inductivo puede usarse efectivamente para complementar el enfoque tradicional de la adquisición del conocimiento. En efecto, la base de conocimiento de un sistema puede construirse a partir tanto de reglas codificadas *a priori* como de reglas generadas inductivamente mediante ejemplos. En otras palabras, las reglas y los ejemplos no necesitan ser mutuamente exclusivos. La fuerza del principio del conocimiento y los trabajos realizados desde los inicios sugieren la necesidad del conocimiento previo. Las reglas *a priori* son una forma simple de conocimiento previo que ha sido utilizada con éxito en muchos sistemas inductivos. De igual forma, se ha propuesto dotar a los sistemas deductivos de capacidad de aprendizaje.

El debate y el estudio en torno a las interdependencias entre aprendizaje y razonamiento, con la subsiguiente integración de la inducción y deducción en esquemas unificados, marcan la pauta hacia el desarrollo de modelos más

poderosos. FLARE es un sistema que intenta combinar el aprendizaje inductivo en base al conocimiento previo, junto con el razonamiento. La inducción y deducción en FLARE se desenvuelven dentro de los límites de la lógica proposicional no recursiva. El aprendizaje se va desarrollando gradualmente a medida que el sistema de adapta continuamente a nueva información. El conocimiento previo está dado mediante reglas que, dentro del contexto de una tarea inductiva particular, pueden servir para generar esquemas útiles de aprendizaje.

Representación del conocimiento.

El lenguaje de FLARE es una instancia del lenguaje de atributo-valor (AVL – *attribute-value language*). Los elementos básicos de conocimiento en AVL son vectores definidos por el producto cruz de los dominios de los atributos. Los componentes de un vector especifican un valor para cada atributo. Así, si A es un atributo y D es el dominio de A, entonces A toma valores de D $\cup$ {*,?}. Los símbolos * y ? significan *no-importa* y *no-sé*, respectivamente.

Las semánticas asociadas con * y ? son diferentes. Un atributo cuyo valor es * significa ser irrelevante –sea conocido o asumido- en un determinado contexto, mientras que un atributo cuyo valor es ?, aunque puede ser relevante en un momento dado, es desconocido. El símbolo * permite la codificación de reglas, mientras que el símbolo ? significa la falta de un valor dentro de lo observable en el mundo real.

Ya que el aprendizaje y el razonamiento normalmente se expresan en el lenguaje clásico de la lógica de primer orden (FOL – *First Order Language*), FLARE requiere traducir estos enunciados en sus equivalentes en AVL. Éste no es tan expresivo como FOL, lo cual ocasiona algunas limitaciones al sistema. Si a los predicados de la forma *p(x)* y *p(x,C)* donde *C* es una constante los llamamos predicados-avl, entonces los enunciados en FOL que pueden ser traducidos a AVL son de dos tipos:

1. Hechos básicos: *p(C)* o $\neg$*p(C)* donde *C* es una constante (e.g., *piedra(A)*).

2. Implicaciones simples: $(\forall x)P(x) \Rightarrow q(x)$ donde $P(x)$ es una conjunción de predicados-avl y $q(x)$ es, sin pérdida de generalidad, un predicado-avl singular, posiblemente negado (e.g., $piedra(x) \wedge peso(x,pesado) \Rightarrow \neg sobre\text{-}la\text{-}mesa(x)$).

Todos los enunciados involucran cuando mucho una variable universalmente cuantificada y son, por tanto, enunciados proposicionales no recursivos. No obstante las restricciones de lenguaje, el sistema maneja en forma efectiva un buen rango de aplicaciones. Además, AVL es eficiente en cuanto a mecanismos de búsqueda de equivalencias y se presta a muchos problemas de aprendizaje inductivo.

Los enunciados FOL de las formas antes mencionadas se traducen de manera simple en una representación AVL equivalente de valores simbólicos, según se muestra a continuación.

1. Definición de atributo: para cada predicado-avl se crea un atributo booleano (para $p(x)$) o de valores múltiples (para $p(x,C)$) equivalente. Si hay hechos básicos, se crea un atributo de valores múltiples, llamado *label*, cuyos valores son los de las constantes.
2. Definición de vector: para cada implicación se crea un vector equivalente donde los atributos correspondientes a la premisa y a la conclusión tienen sus valores propios, mientras que todos los demás atributos son inicializados a *. Para cada hecho básico se crea un vector equivalente donde el valor del *label* es el de la constante y el atributo correspondiente al predicado tiene su propio valor. El atributo correspondiente a la conclusión se marca con un identificador.

La creación del atributo del *label* en el paso 2 proviene del hecho de que los hechos básicos de la forma $p(C)$ pueden reescribirse como implicaciones simples de la forma $label(x,C) \Rightarrow p(x)$. Como puede verse, los atributos cuyos valores son * en un vector corresponden exactamente a aquellos predicados que no aparecen en la premisa del enunciado correspondiente en FOL. El atributo de $q(x)$ tiene diferentes usos. Funciona como una conclusión en el encadenamiento hacia adelante y como clasificación de un objetivo durante el apendizaje inductivo. En algunos casos, puede ser usado también como una

meta. Para evitar confusiones, el atributo correspondiente a *q(x)* es referido como atributo-objetivo (*target-attribute*). Los valores del atributo-objetivo son identificados con el subíndice *T*. La traducción de FOL a AVL actualmente se realiza en forma manual.

Se puede observar que, a medida que el número de predicados se incrementa, lo mismo sucede con el tamaño de los vectores. Dado que los vectores son del mismo tamaño y muchos de ellos pueden tener valores asignados a un número relativamente pequeño de sus atributos, puede resultar que se requiera una memoria muy grande, así como un incremento en el tiempo de operación de los vectores. Cuando hay predicados que asignan distintos valores al mismo tipo de objeto (e.g., rojo(*x*), amarillo(*x*)), es posible limitar el tamaño de los vectores traduciéndolos en atributos únicos de valores múltiples (e.g., color(*x*,*V*) donde *V* es una constante: rojo, amarillo, etc.). Esto es particularmente útil cuando la conclusión *q(x)* corresponde a una clasificación de *x*.

La siguiente tabla muestra la transformación. Esta tabla contiene aseveraciones acerca de animales y su capacidad para volar. Se establece que normalmente los animales no vuelan, las aves son típicamente animales voladores y los pingüinos son aves que no vuelan.

FOL	***AVL***			
	Animal	Ave	Pingüino	Volador
Animal(x) $\Rightarrow \neg$Volador(x)	1	*	*	0_T
Ave(x) $\Rightarrow$ Animal(x)	1_T	1	*	*
Ave(x) $\Rightarrow$ Volador(x)	*	1	*	1_T
Pingüino(x) $\Rightarrow$ Ave(x)	*	1_T	1	*
Pingüino(x) $\Rightarrow$ $\neg$Volador(x)	*	*	*	0_T

En términos generales, el problema del aprendizaje supervisado es como sigue. Dado (1) un conjunto de categorías, (2) para cada categoría, un conjunto de instancias de "objetos" y (3) conocimiento previo opcional, producir un conjunto de reglas suficientes para ubicar a los objetos en su categoría

correcta. En AVL, las instancias consisten en conjuntos de pares de atributo-valor o vectores, describiendo características de los objetos que representan, junto con la categoría del objeto. En este contexto, la categoría es un atributo-objetivo.

Un *ejemplo* es un vector en el que a todos los atributos se les asigna ya sea ? o alguno de sus posibles valores. Una *regla* es un vector en el que algunos de los atributos tiene el valor * como resultado de la generalización durante el aprendizaje inductivo. Un *precepto* es similar a una regla pero, a diferencia de ella, no se infiere a partir de ejemplos. Los preceptos pueden ser proporcionados directamente, o bien pueden deducirse del conocimiento general relevante al dominio en cuestión. En el contexto de una regla dada o precepto, los atributos con valor * no tienen efecto en el valor de la categoría. Los preceptos y las reglas representan un número de ejemplos. Así, si hacemos que p = (*,1,0,0_T) sea un precepto, donde todos los atributos son función del conjunto {0,1,2}, entonces p representa los tres ejemplos: (0,1,0,0_T), (1,1,0,0_T) y (2,1,0,0_T).

La distinción entre reglas y preceptos está limitada al aprendizaje. En cuanto al razonamiento, todos los vectores (incluyendo ejemplos no generalizantes) son reglas. En el sistema, las reglas se forman de condiciones variables, es decir, bajo ciertas circunstancias un atributo es asignado a *. Los preceptos, por otra parte, son reglas codificadas *a priori* y reflejan un conocimiento de cierto nivel (o de sentido común) acerca del mundo real. Un precepto significa algo comunicado a modo de información y no a modo de instrucción.

FLARE es un sistema adaptable por sí mismo y expandible. Utiliza el conocimiento de un determinado dominio así como evidencias empíricas para construir y mantener su base de conocimiento. Ésta última puede interpretarse como el “último y mejor conjunto de reglas” en un momento dado para manejar una aplicación en un momento dado. En este sentido, el sistema se comporta de acuerdo al enfoque científico de la teoría de la formación/revisión: el conocimiento disponible y la experiencia producen una “teoría” que se actualiza y afina continuamente a través de evidencias nuevas.

El sistema trabaja en base a tres funciones principales que pueden presentarse en forma intuitiva como sigue a continuación.

DEFINICIÓN

- Función: Generación-de-preceptos.
  - Entrada: un conjunto de reglas generales, de hechos y un atributo-objetivo.
  - Salida: uno o más preceptos.
- Función: Razonamiento.
  - Entrada: base actual de conocimiento, un conjunto de hechos codificados en un vector *v*, un atributo-objetivo determinado y, opcionalmente, el valor de éste último.
  - Salida: un vector *v*+ igual a *v* junto con otros hechos deducidos a partir de *v*, incluyendo un valor derivado para el atributo-objetivo.
- Función: Adaptación.
  - Entrada: la base actual de conocimiento, el vector *v*+ producto de la función Razonamiento y el valor del atributo-objetivo.
  - Salida: base de conocimiento actualizada.

IMPLEMENTACIÓN

1. Pre-procesamiento: Ejecutar Generación-de-Preceptos.
2. Ciclo principal: Por cada vector generado
   (a) Ejecutar Razonamiento
   (b) Si hay un valor para el atributo-objetivo, ejecutar Adaptación.

Conceptualmente, un sistema de este tipo consiste de dos fases. La primera, de pre-procesamiento, utiliza conocimiento previo en forma de reglas generales que pueden verse como conocimiento de "sentido común" codificado. Mediante la deducción a partir de hechos dados, los preceptos de un dominio específico se generan como una instancia del conocimiento general en el dominio en cuestión. En la segunda fase, la de procesamiento normal, el sistema ejecuta, al menos conceptualmente, un ciclo infinito, dentro del cual es presentada nueva información al sistema. En un primer momento, el sistema ejecuta un proceso de razonamiento a partir de "hechos" mediante vectores de entrada y reglas de su base de conocimiento. En otro momento, dentro del mismo ciclo, el sistema actualiza su base de conocimiento. La combinación de estos dos momentos es lo que se llama *aprendizaje*.

Es factible, entonces, el razonamiento basado en el conocimiento previo disponible y en la actualización de la base de conocimiento. Aún cuando la información disponible es insuficiente y/o incompleta, los seres humanos frecuentemente intentan tomar decisiones tentativas para después corregirlas si es necesario. En cualquier momento dado, la decisión tomada representa una suerte de "mejor corazonada" en virtud de la información actual disponible. Mientras más disponible y cierta sea la información, más precisas serán las decisiones tomadas.

Razonamiento.

FLARE implementa una forma simple de razonamiento basado en reglas, en combinación con el basado en la similaridad, como lo hace el sistema CONSYDERR. Sun afirma que tal combinación disminuye notablemente la fragilidad del sistema. En particular, en ausencia de reglas aplicables o cuando la información disponible es incompleta, FLARE se enfoca a la similaridad con situaciones previas para hacer predicciones útiles. Otros autores sostienen que la analogía es una condición necesaria para el razonamiento de sentido común y la subsecuente pérdida de fragilidad.

Las reglas de aprendizaje inductivas de la forma $(\forall x)P(x) \Rightarrow q(x)$ son esencialmente reglas de clasificación o definiciones que establecen relaciones entre características, simbolizadas en $P(x)$, y conceptos, expresados por $q(x)$. Siguiendo con la suposición clásica de que lo que no es conocido por un sistema de aprendizaje es falso por *default*, las reglas generadas inductivamente se prestan naturalmente al principio de completud propuesto por Clark, en 1978. Es decir, las reglas de clasificación se convierten en enunciados de la forma "si y solo si", i.e. $P(x) \Leftrightarrow q(x)$. Entonces, bajo la completud, si se sabe que $q(x)$ es verdadera, es posible concluir que $P(x)$ es verdadera también.

Es claro que la completud no puede aplicarse a todas las reglas. Las reglas de aprendizaje inductivas son en sí mismas definicionales ya que esencialmente codifican descripciones de conceptos en términos de un conjunto de características. Otras reglas, tales como las que relacionan conceptos en el mismo nivel relativo de conocimiento, no son definicionales. Por ejemplo, dado que las aves son animales y que algún x es un animal, no se

sigue que *x* es un ave. Así pues, además de las reglas de aprendizaje inductivas, las definiciones pueden ser provistas a FLARE como conocimiento previo.

El principio de completud es particularmente útil cuando interactúa con el razonamiento basado en la similaridad para generar nuevas reglas, como se muestra en la siguiente derivación.

- Hipótesis:
  1. $(\forall x)P(x) \Rightarrow q(x)$, que puede ser completado.
  2. $(\forall x)P'(x) \Rightarrow q'(x)$.
  3. $P \cap P' \neq \varnothing$ (i.e. *P* y *P'* tienen algunos atributos en común).
  4. *q(x)* es verdadera.
- Derivación:
  5. *q(x)* de la hipótesis 4.
  6. *P(x)* de la completud aplicada a la hipótesis 1.
  7. *q'(x)* del razonamiento basado en similaridad usando las hipótesis 2 y 3.

Una nueva implicación entre conceptos, llamada $q(x) \Rightarrow q'(x)$, es entonces generada. Aunque el sistema es capaz de derivar q'(x) a partir de q(x), actualmente no añade la implicación nueva a su base de conocimiento.

El siguiente ejemplo, propuesto por (Collins y Michalski, 1989) ilustra el uso de la derivación anterior. Supongamos que el sistema ha aprendido una descripción del área de Chaco en términos de un conjunto G de condiciones geográficas (i.e. $G(x) \Rightarrow$ área(x, Chaco)). Además, supongamos que el sistema conoce una regla que codifica un conjunto de condiciones C suficientes para el establecimiento de ganado (i.e. $C(x) \Rightarrow$ establecer(x,ganado)) y C es tal que C y G comparten cierto número de condiciones. Si al sistema se le dice que el área de interés es Chaco, primero va a deducir por completud que las condiciones en G se cumplen y luego, aprovechando la similitud entre G y C, concluirá que el ganado puede establecerse en Chaco. Hay que notar que el nivel de confianza en la conclusión depende del grado de similaridad.

En un sistema como FLARE, la deducción se aplica hacia adelante. Es decir, los hechos son el punto de partida para el razonamiento. Tales hechos

son codificados en un vector que contiene los atributos de valores conocidos, así como los desconocidos; éstos últimos se expresan como “?”. Uno de los atributos es designado como el atributo-objetivo y, si se conoce, se le asigna su valor. Entonces, el sistema utiliza las reglas de su base de conocimiento y los hechos para derivar un valor para dicho atributo-objetivo. Se asume que la base de conocimiento no está vacía nunca; de lo contrario, el sistema no deduciría nada sino “?”.

El atributo-objetivo siempre tiene una asignación, ya sea mediante la aplicación de una regla o mediante una aserción basada en la similaridad. De este modo, el sistema siempre llega a una conclusión. En el peor de los casos, cuando no hay información acerca del atributo-objetivo en la base de conocimiento actual, el valor derivado en la conclusión debe ser evidentemente “?”. En cualquier otro caso, la validez y precisión de la conclusión derivada depende de la información disponible.

Para efectos de una análisis humano, el sistema aporta toda la información acerca del modo en el que logra sus metas. Actualmente FLARE no es interactivo, esto es, no puede recurrir a un usuario para obtener valores de atributos que podrían ayudar a una mayor precisión en sus resultados.

Aprendizaje.

El sistema FLARE aprende en tanto se adapta al medio de acuerdo a la información que va recibiendo. Expande su aprendizaje inductivo a partir de ejemplos y del conocimiento previo en la forma de preceptos.

A lo largo del tiempo, el sistema va teniendo ante sí una serie de ejemplos y preceptos que han de ser usados para actualizar su base de conocimiento. El conjunto de todos los ejemplos, reglas y preceptos que comparten el mismo atributo-objetivo puede entenderse como una función parcial del sistema que le permite establecer una correspondencia entre distintas instancias y metas. En este sentido, un ejemplo relaciona una instancia particular con un cierto valor dentro del espacio de metas, mientras que los preceptos y reglas son hiperplanos que direccionan todos sus puntos e instancias correspondientes hacia el mismo valor en el espacio de metas.

El aprendizaje consiste, entonces, en la aprehensión mediante el hiperplano más próximo. Es decir, en una primera fase, se aplica el esquema

de razonamiento y, posteriormente, se hacen ajustes a la base actual de conocimiento de forma que refleje la nueva información adquirida. La razón por la que se dice que el algoritmo de aprendizaje es el del hiperplano más próximo es que la fase de razonamiento básicamente pretende identificar la más próxima equivalencia para el vector de entrada. Esta equivalencia puede ser una regla (*i.e.*, hiperplano verdadero) o un ejemplo previamente almacenado (*i.e.*, un punto o hiperplano degenerado).

La aplicación previa del razonamiento permite al sistema predecir el valor del atributo-objetivo en base a la información en la base actual del conocimiento. Si hay atributos con valor desconocido en el vector de entrada y la base de conocimiento contiene reglas que pueden aplicarse para asignar tales valores, las reglas son aplicadas de tal forma que el mayor número posible de atributos sean asignados antes de que la meta final sea anticipada. Así, la exactitud de la predicción se incrementa y la generalización se mejora potencialmente, lo cual le permite al sistema adaptar más efectivamente su base de conocimiento.

Extensionalidad e Intensionalidad.

Ya que es posible usar el conocimiento previo en la forma de preceptos junto con ejemplos (hechos), un sistema como FLARE combina efectivamente el enfoque intensional (basado en características, expresadas como preceptos) y el extensional (basado en instancias, expresado en ejemplos) para aprender y razonar.

La mayoría de los sistemas que aprenden son puramente extensionales, mientras que la mayoría de los sistemas que razonan son puramente intensionales. Aquí surgen las diferencias entre algunos autores acerca de si la inducción y la deducción han de integrarse, ya que una combinación de los dos enfoques sería deseable. Es claro que la combinación incrementa la flexibilidad. Por una parte, la extensionalidad es relevante en cuanto a la habilidad del sistema para adaptarse a su ambiente actual en cada momento, es decir, para ser más autónomo. Por otra parte, la intensionalidad provee un mecanismo por el que el sistema puede adquirir enseñanzas.

Dentro del contexto del razonamiento, los preceptos proveen un medio útil para codificar ciertos enunciados de primer orden (e.g., la regla base de un

sistema experto) que pueden, en su momento, ser apendidos por el sistema y ser usados posteriormente para efectos de razonamiento.

Conclusión.

Un sistema como FLARE combina el aprendizaje inductivo usando conocimiento previo junto con el razonamiento dentro del marco de la lógica proposicional no recursiva. Podemos apuntar algunas conclusiones importantes. En particular,

- El desempeño en cuanto a la inducción puede optimizarse en términos tanto de requerimientos de memoria como de generalización, cuando se utiliza conocimiento previo.
- La inducción a partir de ejemplos puede usarse para solucionar en forma efectiva ciertos conflictos referentes a la extensionalidad.
- La combinación del razonamiento basado en reglas con el basado en similaridad provee un medio útil para ejecutar una función que se aproxima al razonamiento y, al mismo tiempo, reduce la fragilidad.
- La inducción, en la forma presentada, se convierte en un complemento valioso para las técnicas clásicas de adquisición de conocimiento a partir de expertos.

Aunque el sistema ha mostrado buenas perspectivas en aplicaciones experimentales, queda mucho trabajo por hacer para lograr una más completa y significativa integración entre aprendizaje y razonamiento. Las áreas relevantes para futuras investigaciones son:

- Diseño de mecanismos dirigidos al aprendizaje mediante el razonamiento.
- Modelos que ilustren el uso adecuado de la dependencia respecto al orden.
- Fundamentos de la disyunción interna.
- Optimización del uso de reglas aprendidas inductivamente en el razonamiento (aunque existe el fundamento, la inducción puede no producir reglas útiles).

- Estudios sobre la posibilidad de incorporar encadenamiento hacia atrás.
- Traducción de la base de conocimiento del sistema a la inversa, es decir, de AVL a FOL.
- Experimentación con aplicaciones más grandes.
- Expansión del lenguaje al primer orden.

La combinatoria INRC de Piaget como base de la operacionalidad.

La epistemología genética de Jean Piaget estudia el origen y desarrollo de las capacidades cognitivas desde su base orgánica, biológica y genética, encontrando que cada individuo se desarrolla a su propio ritmo. Describe el curso del desarrollo intelectual desde la fase del recién nacido hasta la etapa adulta.

Piaget indica que el aprendizaje es una reorganización de estructuras cognitivas que son consecuencia de procesos adaptativos al medio, la asimilación de la experiencia y la acomodación de tales estructuras.

Períodos del desarrollo genético según Piaget

Sensorio Motriz (nacimiento a los 2 años).

El lactante aprende a diferenciarse asimismo del ambiente que lo rodea. Busca estimulación y presta atención a sucesos interesantes que se repiten.

Operaciones Concretas (2 a 11 años).

Preoperatorio**:** evidencia el uso de símbolos y la adquisición de la lengua. Destaca el egocentrismo, la irreversibilidad de pensamiento y sujeción a la percepción.

Operaciones Concretas: Los niños dominan en situaciones concretas, las operaciones lógicas como la reversibilidad, la clasificación y la creación de ordenaciones jerárquicas.

Operaciones Formales (12 años en adelante).

Transiciones al pensamiento abstracto.

Capacidad para comprobar hipótesis mentalmente.
Se caracteriza por unas destrezas que tienen especial relación con procesos de pensamiento frecuentes en la ciencia.
- Características funcionales: son los enfoques y estrategias para abordar los problemas y tareas.
- Características estructurales: son estructuras lógicas, sirven para formalizar el pensamiento de los sujetos.

Piaget considera el pensamiento y la inteligencia como procesos cognitivos que tienen su base en un substrato orgánico-biológico que va desarrollándose de forma paralela con la maduración, y por otro lado el crecimiento biológico, que va desarrollándose de forma paralela con la maduración y el crecimiento biológico.

Funciones del Proceso Cognitivo

Asimilación.
El organismo incorpora información al interior de las estructuras cognitivas a fin de ajustar mejor el conocimiento que posee. Adapta el ambiente asimismo y lo utiliza según lo concibe.

Acomodación.
Comportamiento inteligente que necesita incorporar la experiencia de las acciones para lograr su cabal desarrollo.

*Esquemas***.**
Un plan cognoscitivo que establece la secuencia de pasos que conducen a la solución de un problema.

Características Funcionales.

 *Se concibe lo real como un subconjunto de lo posible***:** a diferencia de los sujetos que están todavía en el estadio de las operaciones concretas, los que han alcanzado el estadio formal pueden concebir otras situaciones distintas de

las reales cuando abordan las tareas a que son sometidos y son capaces de obtener todas las relaciones posibles entre un conjunto de elementos.

 Carácter hipotético deductivo: la hipótesis es el instrumento intelectual que se utiliza para entender las relaciones entre elementos, porque muchas de las relaciones que el sujeto concibe no han sido comprobadas. Los sujetos estarían capacitados para comprobar estas hipótesis mediante las deducciones correspondientes y ello podría hacerse con varias hipótesis a la vez, de manera simultánea o sucesiva.

 Carácter proposicional: las hipótesis se expresan mediante afirmaciones y lo sujetos pueden razonar sobre estas afirmaciones mediante el uso de la disyunción, la implicación, la exclusión y otras operaciones lógicas. Mientras los sujetos en el estadio de las operaciones concretas realizarían estas operaciones directamente a partir de los datos de la realidad, los sujetos formales convierten los datos en proposiciones y actúan sobre ellas.

Características Estructurales.

 La combinatoria: las posibles combinaciones de unos elementos determinados constituyen una estructura que representa la capacidad de los sujetos para concebir todas las relaciones posibles entre los elementos de un problema.

 El grupo de las cuatro transformaciones: esta estructura representa la capacidad de los sujetos formales para operar simultáneamente con la identidad, la negación, la reciprocidad y la correlación. Estas operaciones formarían una estructura de conjunto, ya que cualquiera de ellas puede expresarse como una combinación de las restantes.

Piaget define una estructura como un sistema que presenta leyes o propiedades de totalidad, en tanto que sistema. Estos sistemas que constituyen estructuras son sistemas parciales en comparación con el organismo o el espíritu. Se trata de un sistema parcial, pero que, en tanto que sistema, presenta leyes de totalidad, distintas de las propiedades de los elementos. Pero el término sigue siendo vago, mientras no se precisa cuáles son estas leyes de totalidad. En ciertos campos privilegiados es relativamente

fácil hacerlo, por ejemplo en las estructuras matemáticas, las estructuras de los Bourbaki, las cuales se refieren a las estructuras algebraicas, a las estructuras de orden y a las estructuras topológicas. Estas estucturas son reversibles.

Por otra parte, la génesis es una cierta forma de transformación que parte de un estado A y desemboca en un estado B, siendo B más estable que A. La génesis es un sistema relativamente determinado de transformaciones que comportan una historia y conducen por tanto de manera continuada de un estado A a un estado B, siendo el estado B más estable que el estado inicial sin dejar por ello de constituir su prolongación. Ejemplo: la ontogénesis, en biología, que desemboca en ese estado relativamente estable que es la edad adulta.

En filosofía, la fenomenología de Husserl, presentada como un antipsicologismo, conduce a una intuición de las estructuras o de las esencias, independientemente de toda génesis.

Toda génesis parte de una estructura y desemboca en una estructura.

A manera de ejemplo se puede considerar el grupo de las cuatro transformaciones, que es un modelo muy significativo de estructura en el campo de la inteligencia, y cuya influencia es muy notable en todos los dominios de la inteligencia formal a este nivel: la estructura de un grupo que presenta cuatro tipos de transformaciones, idéntica I, inversa N, recíproca R y correlativa C. Tomando como ejemplo la implicación p implica q, cuya inversa es p y no q, y la recíproca, q implica p. Ahora bien, la operación p y no q, reciprocada, nos dará: no p y q, que constituye la inversa de q implica p, lo cual resulta ser por otra parte la correlativa de p implica q, puesto que la correlativa se define por la permutación de las “o” y las “y” (de las disyunciones y las conjunciones). Estamos pues ante un grupo de transformaciones, ya que por composición de dos en dos, cada una de estas transformaciones N, R o C dan como resultado la tercera y que las tres a la vez nos remiten a la transformación idéntica I, a saber NR. Además, NC=R, CR=N y NRC=I.

Esta estructura tiene un gran interés en psicología de la inteligencia, ya que explica un problema que sin ella sería inexplicable: la aparición entre 12 y 15 años de una serie de esquemas operatorios nuevos de los que no es fácil

entender de dónde vienen y que, por otra parte, son contemporáneos, sin que pueda verse de inmediato su parentesco. Por ejemplo, la noción de proporción en matemáticas, que no se enseña hasta los 11-12 años. Segundo, la posibilidad de razonar sobre dos sistemas de referencias a la vez: el caso de un caracol que avanza sobre un listón que a su vez es desplazado en otra dirección, o también la comprensión de los sistemas de equilibrio físico (acción y reacción, etc.). Esta estructura tiene una génesis. Se reconocen, en la estructura, las formas de reversibilidad distintas: por una parte, la inversión que es la negación, y por otra parte la reciprocidad. En un doble sistema de referencias, por ejemplo, la operación inversa marcará la vuelta al punto de partida en el listón, mientras que la reciprocidad se traducirá por una compensación debida al movimiento del listón con relación a las referencias exteriores a él. Ahora bien, esta reversibilidad por inversión y esta reversibilidad por reciprocidad están unidas en un solo sistema total, mientras que, para el niño de menos de 12 años, si bien es cierto que ambas formas de reversibilidad existen, cada una de ellas está aislada. Un niño de siete años es capaz ya de operaciones lógicas; pero son operaciones que Piaget llama "concretas", que se refieren a objetos y no a proposiciones. Estas operaciones concretas son operaciones de clases y de relaciones, pero no agotan toda la lógica de clases y de relaciones. Al analizarlas, se descubre que las operaciones de clases suponen la reversibilidad por inversión, + a - a = O, y que las operaciones de relaciones suponen la reversibilidad por reciprocidad. Dos sistemas paralelos pero sin relaciones entre sí, mientras que con el grupo INRC acaban fusionándose en un todo.

Esta estructura, que aparece hacia los 12 años, viene así preparada por estructuras más elementales, que no presentan el mismo carácter de estructura total, sino caracteres parciales que habrán de sintetizarse más tarde en una estructura final. Estos agrupamientos de clases o de relaciones, cuya utilización por parte del niño entre los 7 y los 12 años puede analizarse, vienen a su vez preparados por estructuras aún más elementales y todavía no lógicas, sino prelógicas, bajo forma de intuiciones articuladas, de regulaciones representativas, que no presentan sino una semireversibilidad. La génesis de estas estructuras nos remite al nivel sensorio-motor que es anterior al lenguaje y en el que se encuentra ya una estructuración bajo forma de constitución del

espacio, de grupos de desplazamiento, de objetos permanentes, etc. (estructuración que puede considerarse como el punto de partida de toda la lógica ulterior). Dicho de otro modo, cada vez que nos ocupamos de una estructura en psicología de la inteligencia, podemos volver a trazar su génesis a partir de otras estructuras más elementales, que no constituyen en sí mismas comienzos absolutos, sino que derivan, por una génesis anterior, de estructuras aún más elementales, y así sucesivamente hasta el nacimiento, en lo sensorio-motor, y a ese nivel se plantea todo el problema biológico. Porque las estructuras nerviosas tienen, también ellas, su génesis, y así sucesivamente.

Toda estructura tiene una génesis.

Uno de los resultados más claros de las investigaciones de Piaget, en el campo de la psicología de la inteligencia, es que las estructuras, incluso las más necesarias, en el espíritu adulto, tales como las estructuras lógico-matemáticas, no son innatas en el niño: se van construyendo poco a poco. En una palabra, génesis y estructura son indisociables. Son indisociables temporalmente, es decir, que si estamos en presencia de una estructura en el punto de partida, y de otra estructura más compleja, en el punto de llegada, entre ambas se sitúa necesariamente un proceso de construcción, que es la génesis.

Equilibrio.

Piaget introduce la noción de equilibrio para poder concebir la intima relación entre estructura y génesis. Primeramente, el equilibrio se caracteriza por su estabilidad. Pero observa en seguida que estabilidad no significa inmovilidad. Como es sabido, hay en química y en física equilibrios móviles caracterizados por transformaciones en sentido contrario, pero que se compensan de forma estable. En el campo de la inteligencia tenemos una gran necesidad de esa noción de equilibrio móvil. Un sistema operatorio será, por ejemplo, un sistema de acciones, una serie de operaciones esencialmente móviles, pero que pueden ser estables en el sentido de que la estructura que las determina no se modificará ya más una vez constituida.

Segundo carácter: todo sistema puede sufrir perturbaciones exteriores que tienden a modificarlo. Diremos que existe equilibrio cuando estas perturbaciones exteriores están compensadas por acciones del sujeto, orientadas en el sentido de la compensación.

El equilibrio no es algo pasivo sino, por el contrario, una cosa esencialmente activa. Es precisa una actividad tanto mayor cuanto mayor sea el equilibrio. Gracias al juego de las operaciones, puede siempre a la vez anticiparse las perturbaciones posibles y compensarlas mediante las operaciones inversas o las operaciones recíprocas.

Así definida, la noción de equilibrio parece tener un valor particular suficiente como para permitir la síntesis entre génesis y estructura, y ello justamente en cuanto la noción de equilibrio engloba a las de compensación y actividad. Ahora bien, si consideramos una estructura de la inteligencia, una estructura lógico-matemática cualquiera (una estructura de lógica pura, de clase, de clasificación, de relación, etc., o una operación proposicional), hallaremos en ella la actividad, ya que se trata de operaciones, porque encontramos en ellas sobre todo el carácter fundamental de las estructuras lógico-matemáticas que es el de ser reversibles. Una transformación lógica, en efecto, puede siempre ser invertida por una transformación en sentido contrario, o bien reciprocada por una transformación recíproca. Pero esta reversibilidad está muy cerca de lo que se ha referido aquí como compensación en el terreno del equilibrio. Sin embargo, se trata de dos realidades distintas. Cuando nos ocupamos de un análisis psicológico, se trata siempre de conciliar dos sistemas, el de la consciencia y el del comportamiento o de la psicofisiologia. En el plano de la consciencia, estamos ante unas implicaciones mientras que, en el plano del comportamiento o psicofisiología, estamos ante unas series casuales. La reversibilidad de las operaciones, de las estructuras lógico-matemáticas, constituye lo propio de las estructuras en el plano de la implicación, pero, para comprender cómo la génesis desemboca en esas estructuras, tenemos que recurrir al lenguaje causal. Entonces es cuando aparece la noción de equilibrio en el sentido en que se ha definido, como un sistema de compensaciones progresivas; cuando estas compensaciones son alcanzadas, es decir, cuando el equilibrio es obtenido, la estructura está constituida en su misma reversibilidad.

La Psicología cognitiva y la metáfora computacional.

La revolución cognitiva coincidió, desde sus inicios, en aspectos importantes con la posición piagetiana. Ambas consideran que el objetivo fundamental de estudio es la «capacidad» del individuo. Ambas postulaban que esta capacidad era el reflejo del tipo de representaciones y procesos instanciados en el cerebro.

Estos dos son los principales supuestos de la psicología cognitiva: no se puede explicar la conducta humana sin apelar a un nivel de representaciones y, en segundo lugar, sean cuáles sean las representaciones, éstas han de estar instanciadas en el cerebro materialmente.

La primera idea proporciona una visión activa del individuo; la mente representacional no es un mero dispositivo reactivo, sino que selecciona aspectos relevantes y abstractos de los estímulos ambientales, los coordina en estructuras cerebrales complejas y actúa de acuerdo a su estado interno y los datos externos. La segunda idea da respuesta a una preocupación metodológica; postulando la naturaleza física de los procesos mentales, estos pueden ser concebidos como causas potenciales de la conducta, es decir, como estados objetivos de la materia, susceptibles, por tanto, de investigación científica.

Podemos afirmar que la máquina de Turing opera con símbolos que son análogos o equivalentes a las operaciones básicas que realizan las neuronas. Es por tanto razonable defender que la mente podría ser un dispositivo computacional y que lo que computan los procesos mentales son símbolos. La forma concreta en que la actividad de las neuronas se relaciona con nuestro pensamiento es un asunto discutible. De hecho, podemos considerar que, determinar esta relación con exactitud, es el objetivo nuclear de la psicología cognitiva.

La teoría representacional de la mente humana.

Los postulados fundamentales de esta teoría son los siguientes:

1. Las entidades que conocen actúan sobre la base de representaciones.
2. Sean cuales sean las representaciones, éstas están instanciadas en el cerebro materialmente.

3. El cerebro es un órgano natural. Todo lo que haga el cerebro lo puede hacer en virtud de su estructura lógica (arquitectura funcional).
4. Las representaciones están instanciadas en el cerebro mediante códigos simbólicos de naturaleza física.
5. Sean cuales sean los contenidos que están representados, los símbolos y las reglas que sirven para combinarlos no pueden cambiar.

Los ordenadores no trabajan en modo alguno con números, trabajan con cifras. Los números son entidades abstractas; las cifras son símbolos que pueden interpretarse como sustitutos de números (o de otras muchas cosas). Entre esas otras muchas cosas, las cifras con las que opera un ordenador pueden utilizarse para simbolizar imágenes visuales, palabras o, en general, según los psicólogos cognitivos, estados mentales, tales como intenciones o metas.

Existen numerosos tipos de sistemas simbólicos pero todos ellos tienen tres componentes básicos: un conjunto de símbolos primitivos, unas reglas de combinación para construir símbolos más complejos, y un método para relacionar los símbolos con la realidad que representan. En general, cualquier sistema de símbolos que contenga las propiedades de la máquina de Turing puede representar cualquier cosa.

La psicología cognitiva ha seguido la conjetura según la cual, la mente opera con símbolos y, al igual que una máquina de Turing, ha asumido que el formato y la sintaxis de estos símbolos es independiente del contenido que procesan.

La tarea de los psicólogos evolutivos sería descubrir cuáles son exactamente los símbolos primitivos y cuáles son sus principios combinatorios. Turing fue consciente del problema y lo manifestó haciendo patente lo paradójica que puede resultar, para algunos lectores, la idea de una máquina aprendiz. ¿Cómo pueden variar las reglas de operación de la máquina? Ellas deben describir por completo como reaccionará la máquina, cualquiera que fuese su historia, cualquiera que fuese el cambio sufrido. Esto es completamente cierto. La explicación de la paradoja es, según Turing, que las

reglas que se alteran en el proceso de aprendizaje son de una clase menos pretenciosa, que sólo pretende una validez efímera.[46]

Pero si los símbolos y sus principios combinatorios son independientes de la realidad que representan, entonces ¿qué les permite mantener una relación adaptativa con esa realidad?

Si concebimos el cambio evolutivo a la piagetiana, entonces, deben producirse cambios representacionales y operacionales producidos por la interacción con el ambiente que afecten a la arquitectura funcional, es decir, que afectan a «las reglas operativas de la máquina». El cambio, según Piaget, es un cambio en la estructura lógica del sistema. Para defender esta posición, hay que postular un determinado nivel para el formato de representaciones iniciales y unos mecanismos de transformación capaces de convertir la mente sensoriomotora, en mente representacional, y finalmente, ésta última, en un dispositivo con operatividad formal. Son dos, por tanto, las variables críticas: el formato de las representaciones iniciales y el mecanismo de transformación. Si cada una por separado, o ambas conjuntamente no pueden explicar la transformación, entonces cabría concluir que los cambios evolutivos no son posibles.

Por otra parte, el mayor peligro en las teorías de la representación radica en que cometen la llamada "falacia simbólica", consistente en identificar a la significación únicamente como una cuestión de relacionar un conjunto de símbolos con otro. La ciencia cognitiva necesita explicar cómo es que los símbolos se refieren al mundo. Para entender la mente no basta con determinar la relación de los símbolos entre ellos, sino que resulta indispensable especificar cómo estos sistemas se ajustan adaptativamente a la realidad. "Las redes semánticas, tanto si están basadas en la descomposición como en postulados de significado, no pueden explicar su actuación. Pueden decirnos que dos palabras están relacionadas, o que una oración es una paráfrasis de otra, pero son tan circulares como los diccionarios. Cometen la «falacia simbólica» relativa a que el significado es simplemente una cuestión de relacionar un conjunto de símbolos verbales con otro".[47]

---

[46] A. M. Turing (1950) Computing Machinery and Intelligence. Revista *Mind 49:* 458.

[47] Johnson-Laird, P. N., Herrmann, D.J. and Chaffin, R. (1984). Only connections: a critique of semantic networks. Psychological Bulletin 96, 2: 292-315.

Digamos que tanto la descomposición en primitivos semánticos como los postulados de significado (reglas de producción), son dos propuestas que pretenden explicar la construcción del significado mediante combinaciones sintácticas de elementos primitivos, en un caso, o mediante reglas formales de inferencia, en el segundo. El problema es: ¿cuántas reglas formales de inferencia habría que poseer para dar cuenta, no ya de aspectos conocidos del mundo, sino de relaciones totalmente novedosas? Aun teniendo un conjunto infinito de reglas, ¿cómo decidir cuáles son aplicables en una determinada situación?

De la representación a la computación.

Los principios de la teoría representacional tras la revolución cognitiva pueden resumirse así:

1. Las representaciones están instanciadas en el cerebro mediante códigos simbólicos de naturaleza física.

2. Sean cuáles sean los contenidos que estén representado, los códigos y las reglas que sirven para combinarlos no pueden cambiar.

En base a estos principios habría dos tipos posibles de representación:

a) Representaciones conceptuales de alto nivel, o

b) Representaciones subconceptuales de muy bajo nivel.

La relación de los símbolos con la realidad se daría según dos enfoques posibles:

1. Enfoque sintáctico.

* Las causas de la conducta se deben, en exclusiva, a estas representaciones y a las reglas de combinación de carácter sintáctico prefijadas e inalterables;

* Los conceptos son primitivos de alto nivel y no pueden aprenderse.

La interacción se limitaría a introducir nuevos contenidos, sin que esto produzca cambios en la complejidad o abstracción del formato de las representaciones.

2. Enfoque semántico.

* Las causas de la conducta se deben a computaciones que operan sobre los significados, es decir, sobre los contenidos representacionales.

* Los conceptos son primitivos de bajo nivel que no pueden aprenderse. Sin embargo, permiten recombinarse de formas novedosas construyendo significados. Los significados sí son alterables ya que surgen como «modelos mentales» que captan la estructura de relaciones presente en el ambiente. Son modelos isomórficos de la realidad.

El trasfondo de la teoría computacional de la mente es que todo lo que hace el cerebro, lo hace en virtud de su estructura como sistema lógico, y no porque esté dentro de la cabeza de una persona, o porque sea cierto tejido esponjoso compuesto por un determinado tipo de formación biológica de células. En tanto que los psicólogos cognitivos están empeñados en comprender cómo es posible que las cosas que sólo están en la cabeza, las representaciones, afecten de forma tan determinante el comportamiento.

Piaget era un psicólogo computacional: mantenía ni más ni menos que la mente del niño puede o no realizar determinadas actividades durante las diferentes etapas del desarrollo en virtud de la estructura lógica subyacente en cada momento.

En una máquina como un reloj, podemos describir los movimientos internos de la máquina y estos corresponderán uno por uno, a la ejecución que observamos desde fuera. Tal y como señala Pylyshyn (1984), este es un sistema cerrado y determinístico. Es un sistema operativamente cerrado en el sentido de que las conexiones causales que explican su conducta se limitan al conjunto de variables internas que definen el sistema. Es un sistema determinístico en la medida en que la secuencia de estados por la que pasa el sistema depende, en exclusiva, del estado anterior y del diseño estructural del reloj, en otras palabras, la información interna sería suficiente para predecir el siguiente estado del reloj o para deducir cuál ha sido el paso inmediatamente anterior. Pero además, podemos añadir que es un sistema agenético; su nivel de competencia no cambia nunca. En este tipo de sistemas resulta fundamental el hecho de que no se necesita la noción de representación para

explicar cómo funciona un aparato, sino sólo para explicar cómo se ejecuta la función para la cual fue diseñado. No es necesario que digamos que el sistema hace tal o cual cosa "porque" tiene una representación determinada, aunque frecuentemente utilizamos un lenguaje de este tipo para expresar la razón por la cual el sistema fue diseñado de una forma determinada. Esto supone un fuerte contraste con la situación en la que se tratan de explicar procesos representacionales o intencionales".

La máquina representacional humana no es, desde luego, un sistema cerrado. No es un sistema operativamente cerrado porque la descripción interna de sus estados podría resultar insuficiente para explicar la conducta. Si para explicar cómo un estado interno del sistema produce una conducta tengo que apelar al contenido del mismo, es decir, al estado de cosas externas con el que se corresponde, entonces, los sistemas representacionales no son sistemas cerrados.

El enfoque conexionista en el marco de la psicología cognitiva.

En la década de los noventa, muchos investigadores comenzaron a aplicar las redes neuronales a la psicología cognitiva. Este tipo de trabajos en redes neuronales configuran un punto de vista que se ha dado en llamar *enfoque conexionista* (McLelland y Rumelhart, 1989; McLelland y Rumelhart, 1986; Hanson y Burr, 1990; Hertz, Krogh y Palmer, 1991; Rumelhart, McClelland y el grupo PDP, 1992).

El enfoque conexionista en psicología cognitiva está basado en el cálculo mediante redes neuronales de determinadas propiedades que tienen su contrapartida psicológica. Una red neuronal está compuesta por un conjunto conectado de *neuronas artificiales*. Estas neuronas artificiales, en general, no existen físicamente y se implementan mediante programas de ordenador que almacenan en matrices de datos los parámetros que caracterizan la neurona, de la misma manera que almacenan y se tratan, por ejemplo, los datos correspondientes a las distintas posiciones físicas de un modelo climático. El cálculo matricial permite gestionar un conjunto más o menos grande de estas neuronas artificiales.

Los modelos conexionistas se inspiran en la forma de trabajar en paralelo que tienen las neuronas del cerebro humano. Para ello se utilizan modelos basados en el paralelismo masivo, mediante los que se trata de modelar procesos cognitivos humanos con distinto nivel de complejidad (p.e., desde la percepción al pensamiento consciente). La interpretación sobre el sentido que debe asignarse a dichas unidades básicas originó ciertos debates en los primeros tiempos de la aplicación del enfoque conexionista a la psicología cognitiva (Smolensky, 1988; Fodor y Pylyshyn, 1988). Además, algunos autores han señalado limitaciones importantes del conexionismo y de los modelos basados en redes neuronales.

A continuación se describen brevemente los elementos básicos que intervienen en una red neuronal:

a) Un conjunto de **n** unidades de procesamiento (neuronas artificiales) que reciben impulsos de entrada de otras unidades y envían impulsos de salida a las restantes unidades o nodos. Cada nodo puede representar una variable, un rasgo, un concepto, etc. Estas neuronas artificiales no tienen existencia física real y se implementan mediante programas de ordenador.

b) Un estado de activación de cada unidad (neurona), **a(n)**. Este estado de activación es función de las entradas que recibe la unidad y determina la salida que se envía a las restantes unidades. El estado de activación global viene dado por un vector de orden **n** y es la magnitud más importante de la red neuronal; de hecho, el vector final de activaciones es la magnitud que se suele comparar con los datos experimentales que se desea interpretar.

c) Un vector de orden **n** que representa los impulsos de salida de cada unidad. Estos impulsos de salida son función de la activación de cada una de las unidades.

d) Un patrón de conexiones entre unidades. Este patrón de conexiones puede representarse mediante una matriz de pesos o conexiones **W(n,n)** que se multiplica por el vector de salida para obtener el vector de impulsos de entrada de cada unidad. Las conexiones pueden ser fijas o variables. En este último caso, las conexiones suelen depender del vector de activación.

e) Una regla de activaciones que combine, en un instante determinado, los impulsos de entrada en cada unidad junto con la activación actual de la unidad para obtener la activación en el instante siguiente.

De la descripción anterior se desprende una característica importante de los modelos conexionistas: todas las unidades están realizando cálculos, recibiendo entradas y enviando salidas simultáneamente (en paralelo). Las unidades son “tontas”, en el sentido de que realizan cálculos sin “saber” lo que están haciendo. Simplemente reciben impulsos, realizan unas operaciones matemáticas con dichos impulsos y generan otros impulsos que van a parar a otras neuronas con las que están conectadas. Como consecuencia de ello y añadiendo, cuando resulte necesario, una regla de aprendizaje, es posible simular algunas de las características típicas del sistema cognitivo humano, tales como la degradación progresiva con el daño, el aprendizaje y generalización a partir de ejemplos, el procesamiento de información parcial, contradictoria o confusa y la actuación del cerebro en la resolución de ambigüedades (Rumelhart, McClelland y el grupo PDP, 1992).

Una de las aplicaciones más interesantes del conexionismo es simular el funcionamiento de los esquemas mediante redes neuronales (McLelland y Rumelhart, 1989), de ahí la enorme utilidad de este enfoque. Así, por ejemplo, diferentes elementos del vector de activaciones **a(n)** pueden representar diferentes características o variables de un esquema y estas variables se pueden activar o no dependiendo de la interrelación o interconexión entre ellas. La activación final de cada elemento reflejaría la importancia o peso final de un nodo en una representación cognitiva. Si un conjunto de nodos está activado, y dado que cada nodo representa un concepto o una variable, se entiende que esos conceptos y variables son importantes en la representación final que tiene el sistema acerca de un dominio o una situación determinada.

El paradigma conexionista ha dado lugar a una nueva disciplina basada en el uso de redes neuronales.

Aunque el enfoque conexionista y los tratamientos basados en el uso de redes neuronales han alcanzado un desarrollo notable en los últimos años, los modelos conexionistas adolecen de algunas limitaciones que es preciso conocer para valorar adecuadamente las posibilidades de esta metodología.

En primer lugar, existen fenómenos relacionados con el aprendizaje que no resultan fáciles de implementar mediante el uso de redes conexionistas; por ejemplo, la generación de nuevas construcciones cognitivas que aportan significados más ricos y extensos a los conceptos ya conocidos. Otra situación

similar se produce cuando se intenta simular procesos de carácter implícito o procedimientos complejos. Por otra parte, muchos de los logros de los modelos conexionistas actuales tienen que ver con el reconocimiento de patrones, pero estos procesos son mucho más limitados que los procesos cognitivos implicados en la comprensión y aprendizaje complejo de conceptos, principios y teorías. Además, un concepto clave de los modelos conexionistas, el de activación, no tiene una traducción inmediata y unívoca en términos relacionados con procesos cognitivos. Por otra parte, los mecanismos y aplicaciones matemáticas que utilizan los modelos conexionistas están tan alejados de los enfoques clásicos en psicología que Fodor y Pylyshin, dos críticos bien conocidos, señalan con cierta ironía que «todas las razones que hay para pensar que los modelos conexionistas podrían ser verdaderos, son razones para pensar que pudieran no ser psicología».[48]

[48] García-Madruga, J.A. (1992). Introducción a la edición española, en Rumelhart, D.E., Mclelland, J.L. y el grupo PDP (1992). *Introducción al procesamiento distribuido en paralelo.* Madrid: Alianza Editorial.

VI. DISCUSIÓN, APLICACIONES Y CUESTIONAMIENTOS

Modelos de toma de decisiones.

La corriente clásica de la Inteligencia Artificial presupone que en el mundo no existe la incertidumbre; es posible, entonces, planificar la toma de decisiones ya que se tiene un conocimiento total de las condiciones sobre las que se ejecutaría un determinado plan, así como de los resultados posibles de acuerdo a toda acción que se realizara. En otras palabras, desde el punto de vista clásico de la IA no hay lugar para las contingencias ya que las circunstancias que rodean a las diferentes acciones posibles permanecen sin cambio.

Por el contrario, la elaboración de un plan de contingencias supone que acciones diferentes tienen lugar en diferentes circunstancias. Un sistema que elabore un plan de este tipo establece determinados pasos de decisión, de tal forma que el agente ejecutante tiene la facultad de decidir qué rama del plan seguir en cada momento. Además, en algunos sistemas tales como el llamado "Cassandra", los pasos de decisión significan medios de adquirir conocimiento, es decir, son metas intermedias que, al ser logradas, permiten al agente sobrellevar las circunstancias que pudieran surgir. Este sistema distingue entre los procesos para obtener información, de aquéllos para tomar decisiones.

Muchos planes que usamos en forma cotidiana especifican ciertas maneras de afrontar diversos problemas que pueden surgir en algún momento. Estos planes se conocen como planes de contingencias. Éstas generalmente se hacen explíctas bajo la forma de instrucciones a seguir: "trata de tomar la Av. Western, pero si está bloqueada toma Ashland". Los llamados *sistemas de planeación clásicos* no pueden elaborar planes de este tipo, debido fundamentalmente a que se sustentan en tres supuestos del *conocimiento perfecto*:

1. El sistema tiene un conocimiento completo de la condiciones iniciales a partir de las cuales el plan será ejecutado;
2. Los resultados de todas las acciones son totalmente predecibles;
3. Todo cambio en el mundo ocurre debido a las acciones realizadas por el sistema.

Sobre esos supuestos el mundo es totalmente predecible, de modo que no hay necesidad de ningún plan de contingencias.

Los supuestos del conocimiento perfecto son una idealización del contexto de planeación que se utilizan para simplificar el proceso. Ellos permiten el desarrollo de algoritmos de planeación que tienen propiedades verificables, tales como la completud y la consistencia. Desafortunadamente, hay muy pocos dominios en los que tales algoritmos son realistas; por lo general, el mundo resulta ser más bien impredecible. Desarrollado sobre los supuestos del conocimiento perfecto en un mundo impredecible, un sistema tal puede probar ser redituable en términos de efectividad y costo, si su grado de incertidumbre sobre un determinado dominio es bajo, o si el costo de recuperación por alguna falla es poco. Pero en general, dichos supuestos podrían guiar al sistema a descartar opciones que podrían haber estado disponibles si los problemas potenciales hubiesen sido determinados con anticipación. Por ejemplo, sobre el supuesto de que el tiempo estará soleado, como pronóstico, uno no tendría por qué llevar un paraguas consigo; si, posteriormente, el pronóstico resulta ser erróneo, entonces es imposible usar la sombrilla para no mojarse. Cuando el costo de recuperación por falla es alto, el fallar en prepararse por posibles problemas con antelación puede resultar ser un error caro. Para evitar errores de este tipo, un agente autónomo en un dominio complejo debe ser capaz de elaborar y ejecutar planes de contingencia.

Recientemente, algunos investigadores han comenzado a trabajar sobre la posibilidad de relajar los supuestos del conocimiento perfecto sin separarse del esquema de planeación clásico. Un sistema de planeación como Cassandra[49] elabora planes con las siguientes características:

- Incluyen pasos de decisión específicos para determinar cuál de los posibles cursos de acción seguir;
- Los pasos de obtención de información son distintos de los pasos de decisión;

[49] Cassandra fue una profetiza troyana en la que la gente no creyó cuando predijo atinadamente desastres futuros.

- Las circunstancias en las que es posible llevar a cabo una acción se distinguen de aquéllas en las que es necesario realizarlas.

Aspectos de la planeación de contingencias.

Un sistema de planeación debe ser capaz de elaborar planes que puedan ser exitosos independientemente de que las condiciones iniciales sean desconocidas y de los posibles efectos inciertos de acciones no determinables. Por tanto, un sistema efectivo debe poseer las siguientes capacidades:

- Debe poder anticipar efectos de acciones no determinables;
- Debe poder reconocer cuando un resultado incierto afecta el logro de una meta;
- Debe poder elaborar planes de contingencia para todos los resultados posibles de las diversas fuentes de incertidumbre que afectan a un determinado plan;
- Debe poder programar acciones lo suficientemente sensibles como para detectar la ocurrencia de una contingencia particular;
- Debe generar planes que puedan ejecutarse correctamente independientemente de la contingencia que surja.

El diseño de Cassandra apunta a estas características. Sin embargo, hay otras que no han sido consideradas:

- No se ha considerado el problema de determinar si vale la pena planear para un resultado particular;
- Cassandra no es un sistema de planeación probabilística; no puede hacer uso de ninguna información sobre la similaridad o sucesos de ese tipo;
- se ha ignorado la posibilidad de intercalar planeación y ejecución;
- Cassandra no maneja sucesos exógenos;
- este sistema únicamente puede encontrar planes que involucren decisiones entre cursos de acción que resultan exitosos en contingencias distintas.

En los modelos de planeación clásicos se asume que todas las fuentes de incertidumbre y todos sus resultados posibles son conocidos, y elabora planes para todos aquéllos que afecten el logro de sus metas. La tarea de este tipo de

sistemas es elaborar planes que garanticen el logro de sus metas. No decide cuándo planear y para qué planear. Además, aunque el modelo de Cassandra es consistente y completo, no es sistemático, como se verá más adelante.

Un efecto incierto es un efecto dependiente de un contexto con una precondición desconocida, es decir, una precondición que el sistema no puede ni percibir ni alterar deliberadamente. Consideremos lo que sucede cuando lanzamos una moneda al aire: en principio, dado un conocimiento perfecto de todas las fuerzas y distancias involucradas, sería posible predecir el resultado. En la práctica, tal conocimiento no está disponible y el efecto del lanzamiento es incierto. En principio, podría ser posible especificar las condiciones que dirigirían a la moneda a caer en "cruz"; en la práctica, tales condiciones son desconocidas.

Hay que notar que en algunas circunstancias podría ser posible que un sistema aprendiera a predecir resultados que, hasta un momento dado, hubiese considerado desconocidos: por ejemplo, si aprendiera la dinámica durante el lanzamiento y el trayecto de la moneda. "Desconocida" se refiere únicamente a la situación actual en un momento dado. Un cierto modelo podría facilitar tal aprendizaje, si implicara simplemente el aprendizaje de algunas precondiciones secundarias, más que del algoritmo completo de una cierta acción.

Por otra parte, en este tipo de sistemas se asume que las diferentes fuentes de incertidumbre son independientes entre sí. Cada fuente de incertidumbre tiene un conjunto exhaustivo de sucesos mutuamente excluyentes.

Representación de un Plan Básico de Contingencias.

Un plan se representa como un esquema con los siguientes componentes:

- un conjunto de *pasos*;
- un conjunto de *efectos* debidos a tales pasos;
- un conjunto de *enlaces* entre pasos y efectos producidos y consumidos (un paso consume un efecto cuando se requiere tal efecto para lograr alguna de sus precondiciones). Nótese que los enlaces denotan *intervalos de protección*, esto es, intervalos sobre

los que las condiciones particulares tienen que permanecer ciertas para que el plan funcione adecuadamente;

- un conjunto de *marcos variables* de instancias de operatividad;
- un *ordenamiento parcial* de pasos;
- un conjunto de *condiciones abiertas*, esto es, de objetivos no establecidos;
- un *conjunto de enlaces inseguros*, es decir, de enlaces a partir de las condiciones que pudieran ser falseadas por otros efectos en el plan.

Un plan es *completo* cuando no contiene condiciones abiertas ni enlaces inseguros.

Representación de Contingencias.

Un plan de contingencias se implementa para lograr su objetivo, independientemente de cuáles contingencias conocidas de antemano puedan tener lugar durante la ejecución. Para construir un plan válido de contingencias, el planeador debe ser capaz de enumerar tales contingencias. El conjunto de contingencias conocidas de antemano puede ser determinado a partir de las fuentes de incertidumbre asociadas con el plan. En efecto, una contingencia es un conjunto de resultados posibles para todas las fuentes de incertidumbre relevantes.

Identificación de contingencias.

El seguimento del logro de objetivos de un plan en cada contingencia es un proceso algo complejo. Cada objetivo, paso y efecto es identificado para indicar las contingencias en las que esos elementos participan:

- Las metas son identificadas de acuerdo a las contingencias en las que deben lograrse;
- los efectos son identificados de acuerdo a las contingencias en las que ellos pueden ocurrir, es decir, las contingencias en las que tienen lugar las metas que dichos efectos satisfacen;
- los pasos son identificados de acuerdo a las contingencias en las que ellos deben ser accionados, es decir, la unión de las contingencias en las que cualquiera de sus efectos pueden ocurrir.

Las precondiciones de cada efecto se convierten en nuevas metas, cuyos indicadores corresponden a los efectos por los que ellas se originan.

En general, se asume que un paso particular puede ser ejecutado en cualquier contingencia, aunque posiblemente sin objetivo alguno. Sin embargo, algunas veces es necesario dirigir un paso particular fuera de cierta contingencia para evitar interferencias con el plan. Por ejemplo, consideremos un plan que pretenda lograr que una moneda lanzada al aire caiga en "cara". Cuando se obtiene "cara", la meta se logra y no se requiere tomar ninguna otra acción. En otra contingencia, si la moneda da "cruz", ésta debe ser volteada para lograr la meta propuesta. Es claro, entonces, que la acción de voltear la moneda no debe ser ejecutada en la primer contingencia, y éste es el propósito del direccionamiento de ciertos pasos fuera de ciertas contingencias. Esto se organiza identificando las contingencias con indicadores negativos asociados a aquellos pasos que no han de ser ejecutados. Peot y Smith llaman a dichos pasos *condicionamiento*.

Además, cada paso que dependa, directa o indirectamente, de un resultado particular proveniente de una determinada fuente de incertidumbre, es direccionado fuera de toda contingencia que considere un resultado alternativo de tal fuente de incertidumbre. La razón de esta restricción se discute detalladamente más adelante.

El sistema de indicadores de Cassandra provee al agente con una guía clara para ejecutar el plan, según la cual, simplemente se implementan aquellos pasos cuyos indicadores positivos identifican las condiciones actuales en cada momento de la ejecución. Los pasos sin identificación ni positiva ni negativa implicados en una determinada contingencia no afectan el logro de las metas y su ejecución no se garantiza. En contraste, la ejecución de un plan producido por otros sistemas (como *CNLP*) es guiada por los indicadores de razón asociados a los pasos; en este tipo de planes, una acción requiere ser ejecutada cuando al menos una de las metas representadas en sus indicadores de razón es factible. Por lo tanto, el agente necesita tener algún método para decidir cuál de las metas principales es factible. Se supone que esto puede hacerse comparando los indicadores de contexto de cada meta principal (identificadas como "acciones tontas") con las circunstancias dadas en un momento determinado. El método de Cassandra es más simple: el agente

simplemente acciona los pasos identificados positivamente, en lugar de utilizar los indicadores asociados a un paso para indicar aquellas metas cuyos indicadores de contexto deben ser analizados.

Los principios generales de la propagación de indicadores en Cassandra son:

- Indicadores positivos, que denotan que un determinado elemento del plan contribuye al logro de la meta en cierta contingencia, y se propagan a lo largo de enlaces causales que van desde submetas a los elementos del plan que los establecen;
- indicadores negativos, los cuales denotan que un determinado elemento del plan podría prevenir el logro de la meta en cierta contingencia, y se propagan a lo largo de enlaces causales que van desde los efectos a los elementos del plan que ellos establecen.

Representación de Decisiones.

La planeación puede verse como el proceso de decisión sobre lo que hay que hacer, antes del momento de actuar. La necesidad de planes de contingencia surge cuando las decisiones necesarias no pueden tomarse con anticipación debido a la falta de información. En este caso, las decisiones deben ser tomadas cuando el plan es ejecutado. El agente que implementa un plan de contingencia debe decidir, en un punto determinado, qué curso de acción tomar dentro de las posibilidades, o sea, qué rama del proceso tomar.

En lo dicho hasta aquí se ha asumido que el agente ejecutará aquellos pasos que sean consistentes con cierta contingencia dada en un momento determinado. Sin embargo, tales pasos consistentes no pueden determinarse con antelación; para saber qué contingencia considerar durante la ejecución, el agente que ejecuta el plan debe obtener información que le permita sustentar su decisión. Un plan seguro y viable debe garantizar que los pasos requeridos para la obtención de dicha información no interfieren con los requeridos para llevar a cabo el resto del plan. La planificación ha de considerar tanto los pasos para obtener información, como los otros pasos para soportar la toma de decisiones. Cassandra logra esto considerando explícitamente las decisiones como pasos dentro del plan. Las precondiciones de tales pasos de decisión incluyen metas para obtener la información relevante para la toma de

decisiones; la programación de tales acciones está incluída entonces en el proceso normal de planeación.

Consideremos, por ejemplo, el plan de contingencias aludido arriba: “tratar de tomar la Avenida Western, pero si está bloqueda, tomar Ashland.” Durante la ejecución de este plan, el agente debe decidir en algún punto qué rama del plan ejecutar. El paso de decisión en este caso podría considerar la precondición de conocer si la Avenida Western está bloqueda o no, para lo cual el planeador podría programar una acción de obtención de información para checar el nivel del tráfico en la Av. Western. Esta operación podría, a su vez, considerar la precondición de encontrarse en Western, lo cual podría lograrse viajando hacia la intersección de Western y Belmont. Después de que la decisión sea tomada, el agente puede tomar Western hacia Evanston, o bien continuar sobre Belmont hacia Ashland.

Asumiendo que la meta del plan sea estar en Evanston, el plan final podría ser como se muestra en la fig. 5. Nótese que el flujo de control después de una decisión se representa por líneas gruesas. Las líneas sólidas en el diagrama representan enlaces, con la acción al final del enlace que, a su vez, satisface una precondición de la acción, la cual se indica en el otro extremo del enlace. En este plan, el agente tomará Western a Evanston en una de las contingencias, y tomará de Belmont a Ashland y luego a Evanston en la otra.

Nótese que para determinar la precondición apropiada para un paso de decisión dado, el planeador debe contar con algún medio para determinar exactamente lo que necesitará saber para tomar la decisión durante el tiempo de ejecución. Esta determinación un tanto compleja depende en parte de la manera en que se lleve a cabo el proceso de toma de decisiones. En Cassandra, las decisiones son modeladas como la evaluación de un conjunto de reglas de condición-acción de la forma:

si condición1 entonces contingencia1
si condición2 entonces contingencia2
…
si condición_ entonces contingencia_

Cada resultado posible de una incertidumbre dada origina una regla de decisión; la condición de esta regla de decisión especifica un conjunto de

efectos que el agente debería probar para determinar si se ha de ejecutar el plan de contingencia para tal resultado. Por ejemplo, las reglas de decisión para el plan del ejemplo podría ser como esto:

si Av. Western está bloqueda entonces ejecutar contingencia por Ashland

si Av. Western no está bloqueda entonces ejecutar contingencia por Western.

Las precondiciones para un paso de decisión son metas para conocer los valores verdaderos de las condiciones en las reglas de decisión: se trata de la *metas de conocimiento*. Tales metas se manejan tal como las precondiciones de cualquier paso. En Cassandra no se requieren más elementos para construir planes para obtener información.

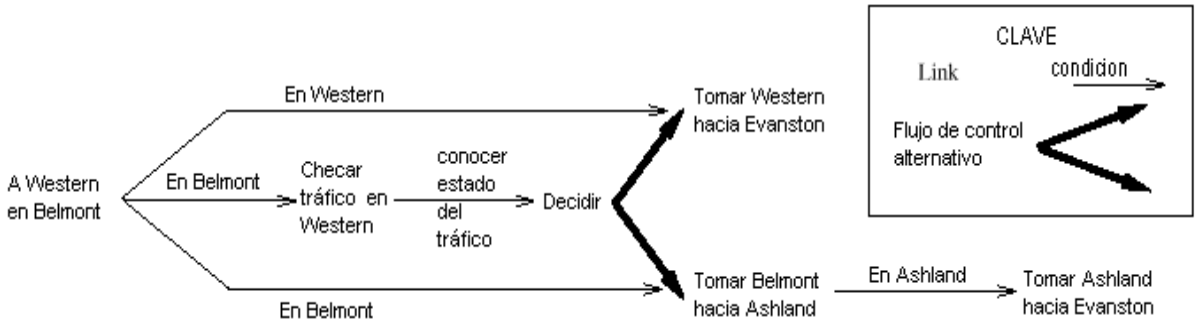

Fig. 5. Un plan incluyendo un paso de decisión

La representación explícita de pasos de decisión provee una base para soportar procedimientos alternativos de decisión. Mientras que el proceso de decisión en el modelo básico de Cassandra es bastante simple, los procedimientos más complejos de decisión pueden ser soportados con el mismo esquema. Por ejemplo, el modelo podría ser cambiado a un procedimiento de diagnosis diferencial. La representación de procedimientos de decisión como acciones permitiría al planeador escoger entre métodos alternativos de toma de decisiones del mismo modo en que puede escoger métodos alternativos para lograr submetas. Una mejor aproximación podría obtenerse formulando una meta explícita para hacer una decisión correcta,

permitiendo que el sistema construyera un plan para lograr dicha meta utilizando operadores inferenciales. Sin embargo, se requeriría que las metas para esos operadores fueran establecidas en un meta-lenguaje que describiera las precondiciones y resultados de los operadores. No hemos apuntado a esa posibilidad en modo alguno.

En Cassandra, la separación entre la obtención de información y la toma de decisiones permite que un mismo paso para obtener información sirva para varias decisiones. Esto proporciona flexibilidad en el uso de las acciones de obtención de información: prácticamente no hay diferencia entre tales acciones y cualquier otra contenida en el plan.

Planeación de Contingencias.

Esta sección describe cómo son introducidas las incertidumbres y cómo son manejadas.

Como ejemplo de un plan involucrando una incertidumbre, consideremos una versión del problema clásico de Moore "bomba en el baño", en el que la meta es *bomba desarmada,* y las condiciones iniciales son *bomba en el paquete1* o *bomba en el paquete2*. La incertidumbre en este caso descansa en las condiciones iniciales: dependiendo del resultado de la incertidumbre, el operador *inicio* puede tener ya sea el efecto de que la bomba esté en el *paquete1* o el efecto de que la bomba esté en el *paquete2*.

Contingencias.

La incertidumbre es introducida dentro de un plan cuando una condición abierta se activa debido a un efecto incierto, es decir, un efecto con una condición desconocida. En el ejemplo de la bomba-en-el-baño, Cassandra puede activar la condición *bomba desarmada* seleccionando el operador *dejar*, el cual tiene las precondiciones *el paquete está en el baño*, y *la bomba está en el paquete*. La condición *la bomba está en el paquete* puede establecerse identificándola con *la bomba está en el paquete1*, la cual es un efecto del operador *inicio*. Sin embargo, esta condición es incierta, como puede determinarse notando que se trata de una precondición desconocida. Cassandra trata con esta incertidumbre introduciendo una o varias contingencias nuevas en el plan. El

estado del plan inmediatamente después de que la incertidumbre es introducida se ilustra en la fig. 6.

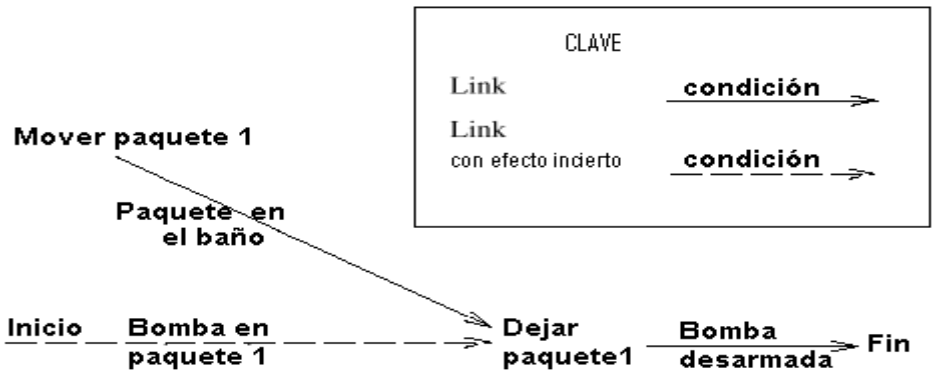

Fig. 6. Introducción de incertidumbre en un plan.

Introducción de contingencias.

Cassandra identifica una incertidumbre cuando el plan se hace dependiente de un resultado particular de tal incertidumbre, es decir, de un efecto incierto cuya precondición desconocida se manifiesta como un resultado de tal incertidumbre. El plan que Cassandra ha desarrollado hasta ese punto es una rama del plan para ese resultado. Dado que las ramas también deben ser construídas para todos los resultados posibles de la incertidumbre, Cassandra genera una copia de su meta global para cada resultado posible. Cada copia conlleva un indicador que la identifica de acuerdo al resultado de incertidumbre del cual se origina. Entonces el plan se divide en un conjunto de ramas, cada una de las cuales representa un posible resultado de la incertidumbre.[50]

En la planeación de metas idénticas para diferentes condicionamientos, Cassandra debe cerciorarse de que ningún elemento de una rama direccionada a cierta meta en virtud de un resultado, sea también direccionada por algún otro resultado de la misma incertidumbre. En otras palabras, ninguna meta, ni cualquiera de sus submetas puede ser direccionada por algún efecto que dependa, directa o indirectamente, de algún resultado de incertidumbre distinto

[50] Un método alterno podría ser dividir el plan en dos ramas, independientemente del número de resultados. Entonces, una rama estaría asociada con un resultado de incertidumbre dado, mientras que la otra lo estaría con todos los demás posibles resultados. De este modo trabaja SENSp.

al que señala el indicador de meta. Como se describe arriba, Cassandra logra esto mediante un sistema de *indicadores negativos*, los cuales señalan las contingencias de las que deben ser excluídos ciertos elementos particulares del plan.

En el ejemplo de la bomba-en-el-baño, cuando del plan se hace dependiente del resultado de incertidumbre *bomba en paquete1*, se añade una copia de la meta principal *bomba desarmada* al conjunto de condiciones abiertas. La copia es identificada con un indicador que señala su pertenencia a una contingencia en la que la bomba está en el *paquete2*. La meta principal existente y todas sus submetas están identificadas para indicar que pertenecen a la contingencia en la que la bomba está en el *paquete1*. El efecto bomba en *paquete1* y la acción *dejar paquete1* junto con todos sus efectos, están identificados para indicar que no pueden jugar un rol en la contingencia en la que la bomba está en el *paquete2*.

Nótese que la acción mover *paquete1*, aunque juega un rol dentro de la contingencia en la que la bomba está en el *paquete1*, no depende del hecho de que la bomba esté en el *paquete1*. Tal acción podría, en principio, ser parte de la contingencia de que la bomba está en el *paquete2*, y probaría ser igualmente útil. Esto se indica por el hecho de que no se tiene un indicador negativo para la contingencia relativa al paquete2.

Cuando Cassandra intenta activar la nueva condición abierta *bomba desarmada*, puede escoger el operador *dejar* otra vez (nótese que está prohibido utilizar cualquier efecto del operador *dejar* existente). Esta nueva instancia del operador *dejar*, a su vez, origina la submeta de que la bomba esté en el paquete supuesto.

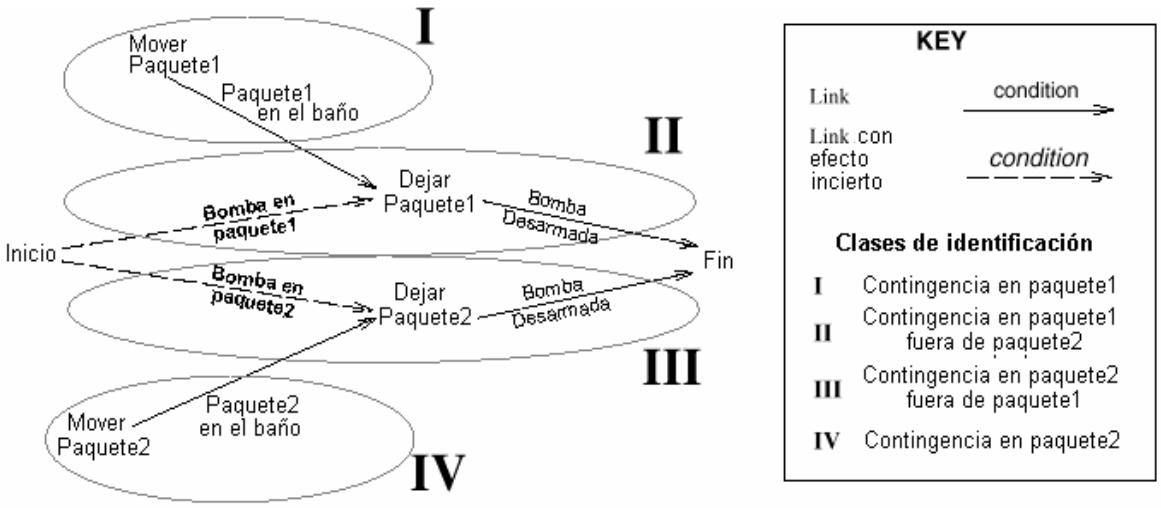

Fig.7. Un plan de contingencias para desarmar una bomba

Incertidumbres con múltiples resultados.

Aunque el algoritmo descrito puede tratar con incertidumbres que tienen cualquier número de resultados posibles, se han discutido ejemplos con únicamente dos posibles resultados. En efecto, estos ejemplos son suficientes para describir la mayoría de los problemas que hemos considerado. Sin embargo, no es difícil pensar en situaciones que pudieran respresentarse en términos de una fuente de incertidumbre con más de dos resultados posibles. Por ejemplo, supongamos que el planeador estuviera interesado en hacerse de un objeto, del cual se supiera que podría estar en alguno de tres lugares. En este caso, el pseudo-operador *inicio* podría representarse con tres efectos inciertos (uno por cada posible localización del objeto) asociados con resultados alternativos a partir de una misma fuente de incertidumbre. El plan de Cassandra para asir el objeto involucraría entonces tres contingencias, una por cada localización posible.

Múltiples Fuentes de Incertidumbre.

Un plan puede involucrar dos o más fuentes de incertidumbre, en cuyo caso se tendría más de un conjunto de ramas. Por ejemplo, supongamos que a Cassandra se le propone la meta de tomar un paquete que puede encontrarse en dos lugares posibles, y que puede usar uno de dos carros disponibles para

ello. Si la incertidumbre relativa a la localización del paquete fuera encontrada primero durante la construcción del plan, Cassandra desarrollaría un plan que involucrara dos contingencias, una para cada localización. Sean tales contingencias *A* y *B*.

En algún punto durante la construcción del plan para la contingencia *A*, Cassandra tendrá que encontrar la incertidumbre relativa al carro disponible, lo cual hará que el plan sea dependiente de un resultado particular de tal incertidumbre. Ya que esta nueva fuente de incertidumbre surge al considerar la contingencia *A* dentro del plan, dicha contingencia se subdividirá en dos contingencias: A_1, en la que el paquete se encuentra en el sitio 1 y el carro 1 está disponible; y A_2, en la que el paquete está en el sitio 1 y el carro 2 está disponible. Ahora Cassandra debe reemplazar todos los indicadores para la contingencia *A* por indicadores de A_1. Por lo tanto, se introducirá una nueva copia de la meta principal identificada con la contingencia A_2.

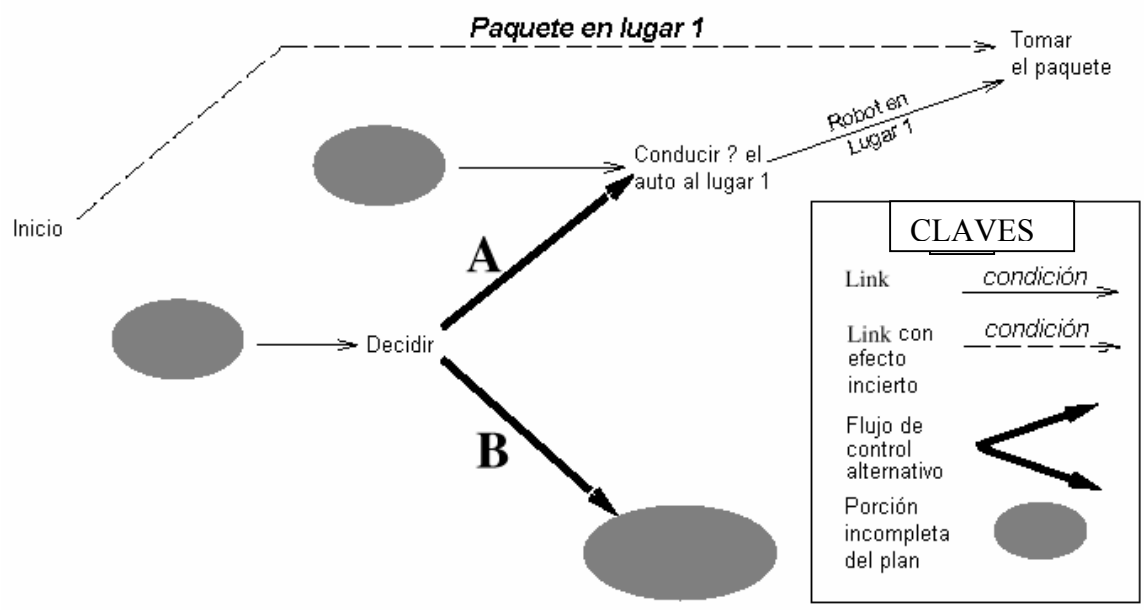

Fig. 8. Un plan parcial para tomar un paquete.

Nótese que Cassandra debe planear desde cero con miras a la meta principal en la contingencia A_2, no obstante el hecho de que ya tenga un plan viable para dicha meta en la contingencia A_1. Esto es necesario porque pueden darse situaciones en las que los únicos planes exitosos impliquen usar

diferentes métodos para lograr una misma meta en dos contingencias. Por ejemplo, si los dos carros fueran en extremo diferentes entre sí, se podrían requerir planes diferentes para manejarlos (tales diferencias podrían incluso afectar las rutas en la que pueden conducirse los carros o los lugares en los que pueden estacionarse). Para que el plan esté completo, Cassandra tiene que considerar todos los modos posibles para lograr la meta en la contingencia *A*2. Si el usar uno u otro carro no afecta el plan de manejo, entonces una ruta a través del espacio de búsqueda producirá planes de contingencia isomórficos para A_1 y A_2 (fig. 9).

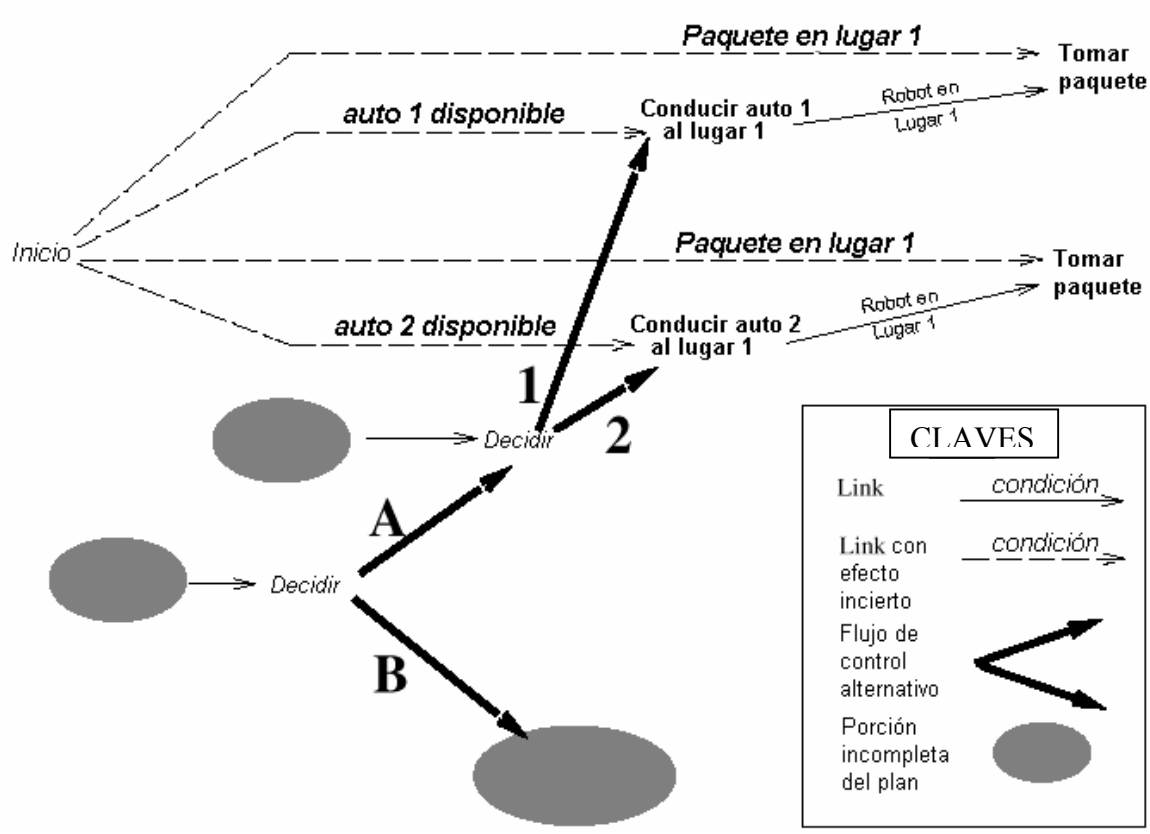

Fig. 9. Un plan con dos fuentes de incertidumbre.

El mismo razonamiento se aplica a la extensión del plan relativo a la contingencia *B*. No se puede asumir a priori que el plan para la contingencia *B* será en modo alguno igual al relativo a la contingencia *A*. Una consecuencia interesante de esto es que la incertidumbre concerniente a la disponibilidad de los carros no necesariamente desemboca en un determinado plan para la contingencia *B*. Por ejemplo, si la localización del paquete en la contingencia *B*

estuviera tan cercana que el agente pudiera llegar ahí sin tener que usar un carro, el plan final podría tener únicamente tres contingencias: A_1 (localización 1 con carro 1), A_2 (localización 1 con carro 2), y *B* (localización 2, a pie).

Desde luego que Cassandra puede producir una ampliación del plan en el que el carro vaya a ser usado en la contingencia *B* también, en cuyo caso encontraría la incertidumbre asociada con la localización del carro, y procedería a bifurcar la contingencia *B* como fue hecho previamente para la contingencia *A*. En última instancia, el plan involucraría una contingencia por cada miembro del producto cruz de los resultados posibles de las incertidumbres relevantes. Sin embargo, es importante notar que no todos los miembros del conjunto del producto cruz deben aparecer como contingencias ya que, como hemos visto, algunas incertidumbres pueden generar únicamente resultados particulares de otras incertidumbres.

Pasos de Decisión.

Cada vez que Cassandra encuentra una nueva fuente de incertidumbre, añade un paso de decisión al plan para representar al acto de determinar la ruta que se ha de seguir durante la ejecución. Las siguientes restricciones de orden son añadidas al plan al mismo tiempo:

- El paso de decisión debe ocurrir después del paso con el que está asociada la incertidumbre;
- El paso de decisión debe ocurrir antes de cualquier paso precondicionado cuyo logro dependa de un resultado particular de la incertidumbre.

Formulación de Reglas de Decisión.

Para que un paso de decisión sea operacional, debe haber un procedimiento efectivo mediante el cual, el agente que ejecuta el plan pueda determinar qué decisión tomar. En Cassandra, la acción de decidir qué contingencia ejecutar se realiza por la evaluación de un conjunto de reglas de condicionamiento de la forma:

Si condición1 entonces contingencia1
Si condición2 entonces contingencia2
Si condición3 entonces contingencia3

. . .

Cassandra anota cada paso de decisión en un plan de acuerdo al conjunto de reglas que serán usadas para tomar tal decisión. El agente entonces toma la decisión evaluando esas reglas al llegar al paso de decisión durante la ejecución del plan. Para evaluar una regla de decisión, el agente debe ser capaz de determinar si se cumple el antecedente de la regla. Las precondiciones del paso de decisión deben pues incluir metas para conocer el estado actual de cada condición que aparezca como un antecedente de una regla en esa condición. Las precondiciones de un paso de decisión se convierten en condiciones abiertas en el plan, al igual que lo hacen las precondiciones de cualquier otro paso.

Como el objetivo de evaluar las reglas de decisión es seleccionar la contingencia apropiada dado el resultado de una incertidumbre particular, las condiciones deberían ser diagnóstico de resultados particulares de la incertidumbre. El agente no puede, desde luego, determinar directamente el resultado de una incertidumbre, ya que ésta debe inferirse de la presencia o ausencia de efectos que dependan de tal resultado.

El camino más directo para construir las condiciones de antecedente de una regla de decisión sería analizar los operadores del plan para identificar todos los efectos que pudieran esperarse, a partir de un resultado de incertidumbre dado, y hacer que la condición sea la conjunción de tales efectos. Pero esto resulta sumamente impráctico. De hecho, únicamente es necesario checar que los efectos de un resultado de incertidumbre dado *sean actualmente usados para establecer precondiciones en la contingencia asociada a tal resultado*. En otras palabras, únicamente es necesario verificar que el plan de contingencias pueda efectivamente tener éxito. Consecuentemente, es interesante observar que el agente podría, en principio, llegar a seleccionar un plan de contingencia aún cuando el resultado de la incertidumbre no fuera aquél con el que el plan estuviera asociado. Hay que notar que esto no causaría un problema en la ejecución del plan, ya que esto ocurriría sólo si se reunieran todas las condiciones para el éxito del plan. De hecho, como veremos, Cassandra depende de esta situación en ciertas circunstancias.

La condición de antecedente de una regla de decisión es pues una conjunción de todos los efectos directos de un resultado particular que son usados para establecer precondiciones en el plan de contingencias para tal resultado. Las reglas de decisión se construyen a medida que el plan se va elaborando. El camino usado en la formulación de las reglas de decisión de Cassandra está basado en la hipótesis de que un agente puede ejecutar un plan si puede "estar seguro" de que todos los eventos del plan son ejecutables.

Adición de una regla de decisión.

En el ejemplo de la bomba-en-el-baño, Cassandra introduce un paso de decisión para determinar si la bomba está o no en el *paquete1*. Como la incertidumbre está en las condiciones iniciales, la decisión será restringida a ocurrir después del paso *inicio*. A la vez, ésta debe ocurrir siempre antes de las acciones de *dejar*, ya que éstas dependen de resultados particulares de la incertidumbre. El paso de decisión tendrá una precondición para saber si la bomba está en el *paquete2*. Si hay acciones disponibles que podrían permitir determinar esto –rayos x, por ejemplo- Cassandra satisfará esta precondición con alguna de esas acciones y, sobre esa base, decidirá qué rama del plan ejecutará.

La construcción de reglas de decisión en Cassandra.

En el punto del proceso de planeación en el que Cassandra construye una regla de decisión, solamente se conoce una sola precondición que dependa de un resultado particular de la incertidumbre que motivó la decisión, a saber, aquélla que guió a Cassandra a descubrir la incertidumbre en primera instancia. El conjunto de reglas de decisión que Cassandra inicialmente construye es algo como lo siguiente:

Si efecto1 entonces contingencia1
Si T entonces contingencia2
Si T entonces contingencia3

Durante la elaboración del plan, Cassandra debe modificar el conjunto de reglas iniciales cada vez que un efecto dependiente directamente de la fuente de incertidumbre es usado para establecer una condición abierta en el plan. Específicamente, Cassandra debe determinar la contingencia en la que

reside dicha condición abierta, y conjuntar el efecto con el antecedente, que ya existe, de la regla de decisión para esa contingencia.

Consideremos, por ejemplo, lo que sucede cuando es lanzada un moneda al aire. Podríamos decir que, teóricamente, existen tres resultados posibles de esta acción: la moneda puede caer con la cara hacia arriba; con la cruz hacia arriba; o de canto (fig. 10). Supongamos que la meta de Cassandra sea que la moneda caiga sobre algún lado plano. Esto puede establecerse usando el *efecto-cara* plana al lanzar la moneda. Pero debido a que éste es un efecto incierto, Cassandra introduce dos nuevas contingencias en el plan, una para el resultado en que la moneda cae en cruz, y otro para el caso en que caiga de canto.

```
Acción:          (lanzar-moneda  ?moneda)
Precondiciones:  (tener  ?agente  ?moneda)
Efectos:         (:cuando  (:V  H)
                  :efecto (:y  (plana  ?moneda)          efecto incierto
                               (cara  ?moneda)))
                 (:cuando  (:V  T)
                  :efecto (:y  (plana  ?moneda)          efecto incierto
                               (cruz  ?moneda)))
                 (:cuando  (:V  E)
                  :efecto  (de canto  ?moneda)))         efecto incierto
```

Fig. 10. Representación del lanzamiento de una moneda

La introducción de esas contingencias exige la introducción de un paso de decisión cuyo conjunto de reglas iniciales es algo como esto:

```
Si  (moneda plana)  entonces  [V1 : H]  regla para “cara”
Si  T               entonces  [V1 : T]  regla para “cruz”
Si  T               entonces  [V1 :E]   regla para “canto”
```

Al mismo tiempo, una nueva condición abierta (saber-si (cara plana)) es introducida como una precondición del paso de decisión, y nuevas condiciones de meta son introducidas para ser satisfechas en las contingencias [V1 : T] y [V1 : E]. Después, Cassandra establece la condición de meta en la contingencia [V1 : T] usando el *efecto-cruz* en el paso de lanzamiento. Las

reglas de decisión asociadas con la contingencia *cruz* son entonces modificados como sigue:

Si (moneda plana) entonces [V1 : H] regla para “cara”
Si (moneda plana) entonces [V1 : T] regla para “cruz”
Si T entonces [V1 :E] regla para “canto”

Finalmente, la condición de meta es establecida en la contingencia [V1 :E] mediante la introducción de un nuevo paso, *poner-de-lado* , en el plan. Una precondición del paso tip es que la moneda esté de canto, lo que se establece por el *efecto-canto* en la acción de lanzamiento. Como este efecto depende directamente de la incertidumbre V1, la regla de decisión para la contingencia *canto* es modificada para incluir esta condición:

Si (moneda plana) entonces [V1 : H] regla para “cara”
Si (moneda plana) entonces [V1 : T] regla para “cruz”
Si (moneda de canto) entonces [V1 : E] regla para “canto”

Ya que el plan está completo, éste es el conjunto de reglas de decisión final. Nótese que esas reglas no distinguen entre los resultados “cara” o “cruz”. De hecho, cualquier resultado tendrá lugar, de modo que no es necesario hacer esta distinción. Cuál plan será ejecutado en cualquiera de esas condiciones depende solamente del orden en que el agente elija evaluar las reglas de decisión.[51]

Un problema algo más complejo se presenta si proponemos a Cassandra la meta de que la moneda caiga sobre un lado plano y específicamente en “cara”. En este caso, ambos efectos pueden establecerse usando la acción de lanzamiento. Esto nos llevará, otra vez, a la introducción de dos nuevas contingencias en el plan, una para el caso de que la moneda caiga en “cruz”, y otra para cuando caiga de canto. Aunque Cassandra podría establecer (cara plana) en el caso de “cruz”, fallaría para completar el plan, ya que la moneda podría no estar en “cara”. Sin embargo, podría ejecutarse la acción de *voltear*, para dejar la moneda en “cara”, dado que se tuviese que comenzar con la moneda en “cruz”. En este punto, las reglas de decisión son como sigue:

[51] Naturalmente, una ampliación de Cassandra podría ser la construcción de un post-procesador que revisara las reglas de decisión que no discriminan entre conjuntos particulares de resultados, y ajustara el plan removiendo las contingencias superfluas. Hay que notar que la viabilidad de esto sólo puede determinarse hasta que el plan esté completo.

Si (y (moneda plana) (cara)) entonces [V1 : H] regla para “cara”
Si (y (moneda plana) (cruz)) entonces [V1 : T] regla para “cruz”
Si T entonces [V1 : E] regla para “canto”

Cassandra debe ahora planear una meta conforme a la situación de que la moneda esté de canto. Ambos efectos pueden establecerse como resultado de la acción poner-de-lado. Sin embargo, el resultado “cara” es un efecto incierto de esa acción, ya que la moneda podría fácilmente estar en “cruz”. Por lo tanto, Cassandra debe añadir una nueva contingencia para cuando la moneda caiga en “cruz” después del *poner-de-lado*. En esta instancia, la meta puede establecerse mediante la acción *voltear*, y la condición “cruz” de esta acción mediante el resultado incierto de la acción *poner-de-lado*. La regla de decisión final para la primera decisión quedaría así:

Si (y (moneda plana) (cara)) entonces [V1 : H] regla para “cara”
Si (y (moneda plana) (cruz)) entonces [V1 : T] regla para “cruz”
Si (moneda de canto) entonces [V1 : E] regla para “canto”

Si la contingencia “de canto” es requerida, otra decisión, proveniente de un resultado incierto de *poner-de-lado*, tiene que ser añadida al plan. Si llamamos V2 a esta segunda fuente de incertidumbre, las reglas de decisión son:

Si (cara) entonces [V2 : H]
Si (cruz) entonces [V2 : T]

El plan se ilustra en la fig. 11.

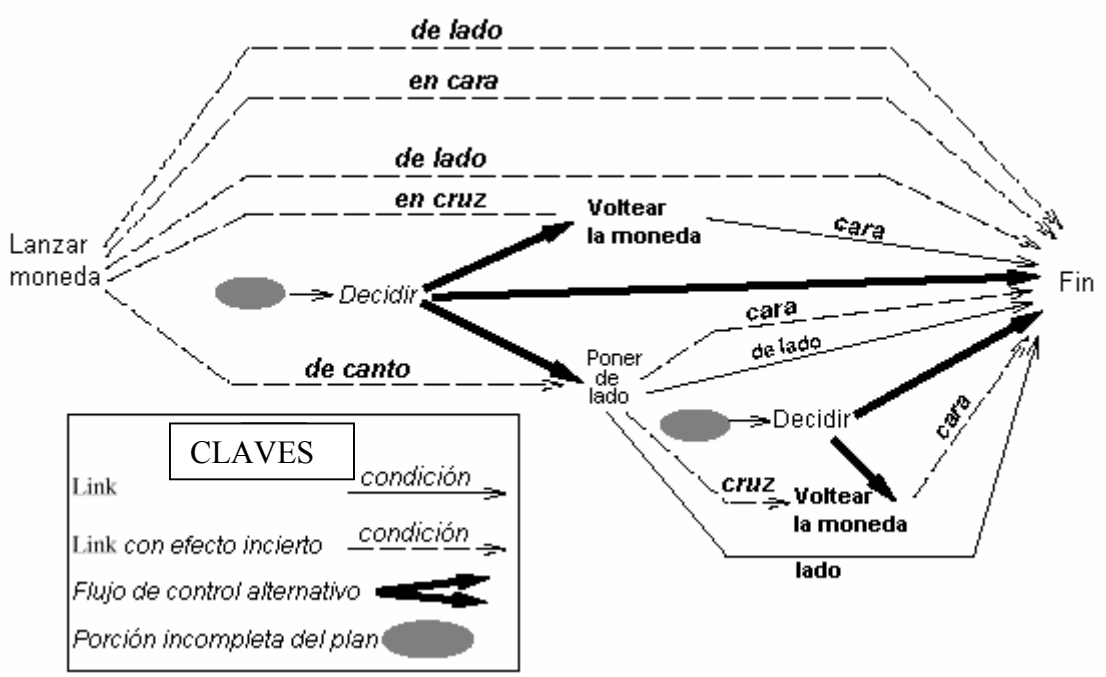

Fig. 11. Un plan con dos decisiones

Reglas de decisión y vínculos (Links) inciertos.

El hecho de que Cassandra permita reglas de decisión que no distinguen completamente entre resultados de incertidumbre, provoca una situación un tanto complicada. Consideremos el plan parcial para abrir una puerta asegurada como se muestra en la fig. 12. La acción de patear la puerta tiene, digamos, dos posibles resultados, una en la que el cerrojo es abierto y otra en la que el pie del agente se rompe. Un plan para la contingencia en la que el cerrojo se rompe, simplemente consiste en abrir la puerta. Para la contingencia alternativa, el plan sería quitar el cerrojo y luego abrir la puerta.

Ya que el segundo plan no depende causalmente de ningún resultado de la incertidumbre (el pie del agente no tiene que romperse para poder quitar el cerrojo y abrir la puerta), la reglas de decisión correspondientes serían:

Si (cerrojo abierto) entonces [D : L] *regla para cerrojo abierto*

Si T entonces [D : F] *regla para pie roto*

Hay que notar que, en este caso, la acción de *abrir del cerrojo* depende de que la cerradura esté intacta, mientras que la acción de *patear la puerta* puede tener el efecto de que la cerradura ya no quede intacta. En otras palabras, la acción de patear obstruye potencialmente la precondición de

patear. Sin embargo, el planeador puede con razón ignorar esta obstrucción, puesto que las dos acciones pertenecen a distintas contingencias. No obstante, esto es válido solamente si la estructura de las reglas de decisión garantiza que el agente no escogerá ejecutar la contingencia de *abrir* el cerrojo cuando el resultado de patear sea que el cerrojo se rompa. Las reglas de decisión indicadas más arriba no fuerzan esto. En tal caso, la solución sería ampliarlas para una contingencia en la que la cerradura no se rompa y probar si efectivamente está intacta. Se tendrían entonces las siguientes reglas de decisión:

Si (cerrojo roto) entonces [D : L] *regla para cerrojo roto*

Si (no (cerrojo roto)) entonces [D : F] *regla para pie roto*

Cassandra amplía de este modo las reglas de decisión cada vez que un efecto directo de una incertidumbre podría obstaculizar algún *link* en una contingencia diferente.

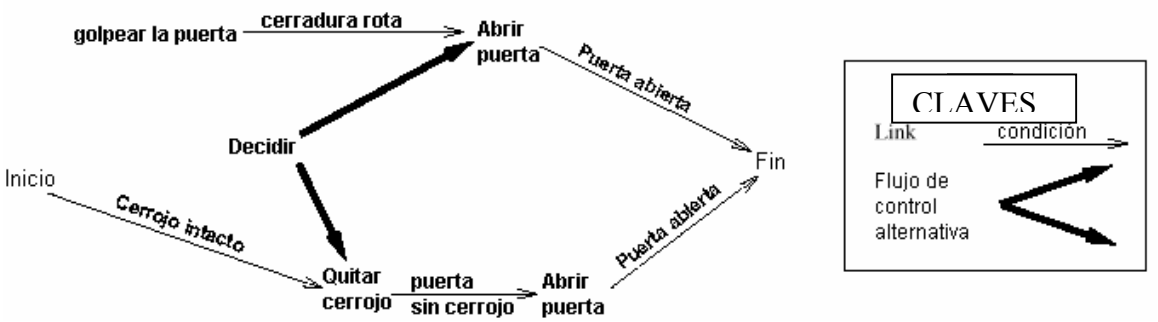

Fig. 12. Plan parcial para abrir una puerta.

Puntos Importantes en los Planes de Contingencias.

Cassandra es un planeador de orden parcial que desciende directamente de UCPOP, del cual hereda las propiedades de consistencia, completud y sistematicidad. Todos los planes producidos por UCPOP están garantizados para lograr sus metas, encontrar el plan adecuado y nunca duplicar la ejecución de un plan parcial. En esta sección se discuten estas propiedades y todos los aspectos relacionados dentro de la planeación de contingencias.

Consistencia.

La consistencia de UCPOP depende de los supuestos del conocimiento perfecto. Un plan de UCPOP es consistente si las condiciones iniciales están totalmente especificadas, y si todos los posibles efectos de las acciones están especificados en los operadores que los respresentan. Si no hay incertidumbres involucradas en el plan, entonces Cassandra es equivalente a UCPOP y, por tanto, produce planes consistentes.

Si hay incertidumbres implicadas en el plan, entonces ya no puede asumirse que las condiciones iniciales y todos los posibles efectos están completamente especificados. De modo que las incertidumbres surgen cuando tales supuestos son violados. Sin embargo, los supuestos pueden ser adaptados para tratar con la inciertidumbre: sería posible, por ejemplo, lograr que todas las posibles condiciones iniciales, así como sus efectos, fueran especificados. En la representación de Cassandra, esto significa que cada fuente de incertidumbre debe especificarse a través del uso de precondiciones secundarias desconocidas, y que cada posible resultado de tales fuentes debe también especificarse.

Se podría afirmar que Cassandra, bajo esas condiciones, es válida. La prueba de ello se seguiría del procedimiento de adición de nuevas metas cada vez que surge una nueva fuente de incertidumbre, ya que esto asegura el logro de cada meta en cualquier resultado posible de incertidumbre.

Completud.

Se afirma que Cassandra es completa en el sentido limitado de que, si existe un plan consistente de la forma que se puede construir, entonces Cassandra lo encontrará. Esto es simplemente una extensión de la completud de UCPOP. Si no existen incertidumbres involucradas, Cassandra siempre encontrará un plan del mismo modo de UCPOP. La introducción de una fuente de incertidumbre dentro de un plan provoca la adición de nuevas metas contingentes. Cassandra encontrará un plan para cada una de tales metas dentro de la contingencia apropiada. Por tanto, si la meta puede lograrse en cada contingencia, Cassandra encontrará un plan para lograr tal meta, así como determinará cuál de esas contingencias acatar.

La completud de UCPOP, como su consistencia, depende de los supuestos del conocimiento perfecto. En el caso particular de Cassandra, dicha completud se asienta sobre tres extensiones a esos supuestos:

- Todas las fuentes de incertidumbre están especificadas;
- los resultados especificados son exhaustivos;
- existen acciones disponibles que permiten la determinación del resultado de cualquier incertidumbre, aún cuando pueda ser sólo en forma indirecta.

Desafortunadamente, esas condiciones son necesarias pero no suficientes. Cassandra puede encontrar planes únicamente si las acciones que utiliza para determinar la contingencia no interfieren con el logro de la meta. Podría, por ejemplo, estar disponible una acción de abandono que podría detonar cualquier bomba dentro de determinado paquete. Ciertamente, ésta es una acción que permitiría determinar el resultado de la incertidumbre, pero no existe un plan sólido que haga uso de ella.

Para tener una noción útil de la completud de Cassandra, debemos especificar entonces la forma de los planes que pueden construirse. Éste es un problema común para probar la completud de cualquier planeador: por ejemplo, no podemos afirmar que, digamos, SNLP es incompleto porque no puede encontrar un plan para el problema de la bomba-en-el-baño. Más bien decimos que no existe un plan consistente de la forma en que puede ser construído. Es fácil especificar la forma de los planes que SNLP puede construir: consisten en secuencias de pasos parcialmente ordenados y que existen para ser todos ejecutados. La introducción de contingencias hace de la descripción de los planes de Cassandra algo más complejo; se necesita aún formalizar una descripción, pero se está ya trabajando en esa dirección. De manera informal, Cassandra puede construir únicamente planes que, para cada fuente de incertidumbre, incluyen un paso para decidir sobre una de las ramas relevantes del plan. La extensión de Cassandra para resolver el problema de la bomba-en-el-baño puede hacer eso porque ella puede construir planes para los que no aplica ese criterio.

Sistematicidad.

UCPOP es sistemático: nunca ejecuta un plan parcial más de una vez mientras busca soluciones. Cassandra, como se ha descrito, no es sistemático; puede ejecutar algunos planes parciales en el espacio de búsqueda más de una vez. Consideremos nuevamente el plan para desarmar la bomba. En este plan hay dos maneras diferentes de establecer la meta: optando por el paquete1 y, por el paquete2. Inicialmente, Casandra puede escoger cualquiera de las dos maneras y, en ambas, provoca la introducción de una contingencia y la necesidad de replantear la situación para lograr la meta en la otra contingencia. Ambas rutas de búsqueda conducen al mismo final del plan, de modo que la búsqueda no es sistemática.

Cassandra podría sistematizarse restringiendo el manejo de las contingencias únicamente de acuerdo a cierto orden, de tal manera que una ruta bajo otro orden de búsqueda sea tratada como un fin completo. Sin embargo, esta extensión no ha sido añadida dado el debate que, actualmente, existe en torno a lo deseable que pudiera ser la sistematicidad. Por ejemplo, Langley (1995) arguye que un método de búsqueda no sistemático, mediante iteraciones, resulta frecuentemente mejor que uno sistemático, de primera búsqueda en profundidad, para problemas que tienen múltiples soluciones así como rutas de solución profundas. El desempeño de una versión no sistemática de SNLP resulta ser mejor que el de la versión sistemática original. Este comportamiento se atribuye al hecho de que la duplicidad en la exploración de planes resulta más económico que el que se hace en el aseguramiento de la sistematicidad.

Metas de Conocimiento.

Un agente que lleva a cabo planes de contingencia debe ser capaz de obtener información acerca del estado actual del mundo, de tal manera que pueda determinar cuál de los posibles cursos de acción seguir. Un sistema que construye planes de contingencia debe ser capaz de planear la obtención de esta información; en general, el proceso de obtención puede ser arbitrariamente complejo.

Las discusiones sobre metas para obtener conocimiento acerca del mundo han tenido como hilo conductor la concepción de que las metas de conocimiento surgen de la necesidad de especificar las acciones que van a ser

desempeñadas; en otras palabras, de la necesidad de hacer de las acciones algo operacional. Los trabajos en esta área se han concentrado totalmente en la capacidad de describir y representar metas de conocimiento y, en cambio, han ignorado en gran medida los aspectos relacionados con el desarrollo de planeadores que construyan planes que los contengan a sí mismos.

La estructura de Cassandra está basada en la noción de que las metas de conocimiento surgen fuera de la necesidad de tomar decisiones al modo de acciones que han de ejecutarse. En un mundo conformado a las suposiciones del conocimiento perfecto, como en la planeación clásica, esto siempre es posible puesto que el mundo es totalmente predictible, de modo que los planes no necesitan contener metas de conocimiento. Sin embargo, cuando esas suposiciones se tambalean, puede no ser posible tomar todas las decisiones con anticipación si la información necesaria para hacerlo no está disponible para el planeador. La información puede no estar disponible ya sea debido al limitado conocimiento del planeador acerca del mundo, o porque los eventos que causan, de modo no determinístico, las condiciones que afectan las decisiones, no han ocurrido aún. En ambos casos puede ser posible que el planeador determine que una decisión pueda ser tomada aún cuando no pueda, en ese momento actual, tomarlas. En este caso, el planeador puede diferir la decisión: puede planear hacerla en el futuro, cuando la información necesaria esté disponible. Parte del plan es, entonces, obtener información y, por tanto, contiene metas de conocimiento.

El uso que hace Cassandra de las precondiciones desconocidas para indicar no-determinismo es, pues, una parte crucial de su mecanismo. En Cassandra, las metas de conocimiento surgen como resultado de decisiones diferidas. Tales deciciones están representadas explícitamente en sus planes, y emergen directamente de la incompletud del conocimiento del mundo por parte de Cassandra, ya sea a través del los efectos de acciones no determinísticas o a través de condiciones iniciales no especificadas completamente. Ambas formas de incertidumbre son manejadas del mismo modo: una vez que Cassandra ha reconocido la necesidad de diferir una decisión, la *razón* de ello no es importante excepto por el hecho de que resulta de un conocimiento incompleto del mundo.

La opinión de las metas de conocimiento como surgiendo de decisiones diferidas es básicamente consistente con la opinión de que son necesarias para realizar acciones operacionales, pero difiere del punto de vista tradicional de que las metas de conocimiento no son directamente precondiciones de acciones físicas, sino que son, en cambio, precondiciones de las acciones que toman decisiones. Por ejemplo, McCarthy y Hayes consideran el problema de una caja de seguridad con combinación: comunmente se sostiene que la acción de abrir la caja tiene como precondición el conocer la combinación. En Cassandra, sin embargo, la meta de conocer la combinación puede surgir como una submeta para decidir cuál rama del plan seguir, donde podría haber una rama para cada cambinación posible.[52] Las ramas podrían surgir debido al conocimiento incompleto del mundo por parte de Cassandra: las condiciones iniciales en las que el plan va a ser ejecutado no están completamente especificadas.

Cassandra se limita a utilizar la forma *saber-si(hecho)* para representar metas de conocimento, lo cual parece ser adecuado si, como se ha supuesto, todos los resultados posibles de una incertidumbre dada son conocidos.

Aspectos varios de la planeación de contingencias.

Los lineamientos de Cassandra arrojan varias cuestiones en torno al comportamiento deseable de un planeador de contingencias, muchas de las cuales de tienen una respuesta fácil. En esta sección examinamos brevemente algunos de estos puntos.

Dependencia de resultados y de contingencias superfluas.

El hecho de que un plan de contingencias asuma un resultado particular de una incertidumbre, significa únicamente que éste no depende de otro resultado diferente de dicha incertidumbre. Cassandra no restringe el plan a que deba ser causalmente dependiente del resultado que éste asuma. Así, el plan para tomar Ashland, descrito anteriormente, no depende actualmente de que Western esté bloqueda; el plan podría ejecutarse exitosamente independientemente del nivel de tráfico en Western.

[52] De aquí surge la pregunta de si el establecimiento de ramas por cada combinación posible es algo razonablemente práctico.

Esta observación arroja una cuestión interesante: ¿Si un plan para una contingencia prueba que no depende de ningún resultado de la incertidumbre que lo provocó, no obviaría ello la necesidad de tener planes para contingencias alternativas? Así, en nuestro ejemplo, podría parecer sensato ejecutar el plan para tomar Ashland independientemente de si Western está bloqueda. Entonces, podría parecer que el planeador debería editar el plan de algún modo tal que eliminara las contingencias aparentemente superfluas. Sin embargo, fácilmente puede mostrarse que la versión del plan que no involucra dependencia de ningún resultado de la incertidumbre será generado en cualquier lugar del espacio de búsqueda. En el ejemplo, esto significaría que el planeador podría de hecho considerar un plan que simplemente implicara tomar Ashland. Si la heurística de búsqueda provoca desventajas en los planes que involucran contingencias, entonces sería apropiado preferir este otro plan, en lugar del plan de contingencias.

Contingencias laterales.

A pesar de la discusión precedente, un plan que no involucra contingencias no siempre es superior a uno que implica alguna. Y así es como un planeador podría de hecho construir un plan como el de Western/Ashland. Para tener una idea más precisa, supongamos que Pat necesita $50 para apostar a un caballo. Ella podría tratar de pedir prestado a Chris, pero el resultado de esta acción es incierto –Chris podría rehusarse. Como alternativa, ella podría robar una tienda. Mientras que el plan de robo podría no involucrar contingencias, sería un mal plan por otras razones. Sería mejor tratar primero de obtener el préstamo de Chris y, entonces, si no funciona, robar la tienda. Cassandra podría generar este plan. Para hacerlo preferible a la alternativa libre de contingencias, sin embargo, su métrica de búsqueda tendría que tomar en cuenta los costos estimados de varias acciones, y desempeñar algo similar a un cálculo probabilístico. Para ejecutar el plan apropiado, podría ser necesario también tener algún modo de conocer que el plan del préstamo debería preferirse al del robo, si fuera posible ejecutar cualquiera de los dos.

Ramas Idénticas.

Es posible que un solo plan pueda funcionar igualmente bien para diferentes resultados de una incertidumbre. Por ejemplo, supongamos que la acción de pedir $50 a Chris tiene tres posibles consecuencias: ya sea que Pat consigue el dinero y Chris se siente satisfecha (por haber tenido la oportunidad de hacer un favor); o bien Pat obtiene su préstamo y Chris no está satisfecha (por haberse sentido obligada a hacer un favor); o simplemente Pat no consigue el dinero. Si Pat desarrolla un plan en la que ella trata de obtener un préstamo de $50 para apostarle a un caballo, entonces, asumiendo que este plan no depende de la felicidad de Chris (podría depender, por ejemplo, si Pat necesitara de Chris un pase para el hipódromo), el plan funcionará ya sea para "conseguir dinero + Chris contenta", o para "conseguir dinero + Chris molesta".

Cassandra podría encontrar tal plan, pero podría tener que encontarlo dos veces –una para cada resultado de la incertidumbre- y requeriría aún un paso de decisión para discriminar entre esos resultados. Esto muestra ser ineficiente de dos modos: el tiempo extra de búsqueda requerido para encontrar dos veces el mismo plan y, el esfuerzo inútil para tomar una decisión innecesaria. Se está trabajando para evitar el primer problema. El segundo podría resolverse mediante un post-procesador que pudiera "fusionar" planes de contingencia idénticos, pero aún no se ha desarrollado esa técnica.

Ramas Combinadas.

Es posible diseñar un plan en el que las ramas pudieran separarse y luego unirse. Por ejemplo, consideremos el plan Western/Ashland otra vez. El contexto en el que se da la meta para tomar Evanston podría ser la obligación de enviar un brindis para la cena a un cierto restaurant de Evanston. La contingencia debida a la incertidumbre acerca del tráfico en la Av. Western parecería afectar, en este caso, únicamente la parte del plan concerniente a tomar Evanston; quizás tenga cierta relación con la manera de expresar el tipo de brindis, la selección del vino, etc. La forma más natural de diseñar este plan podría ser, entonces, asumir que, independientemente de cuál contingencia tenga lugar, el planeador tendrá eventualmente que llegar a determinado lugar en Evanston y, desde ese punto, habrá que desarrollar un plan para lograr la meta final.

El diseño de un plan en esta forma tendria una descripción más concreta, y reduciría el esfuerzo en desarrollarlo evitando, por ejemplo, la construcción de múltiples copias del mismo subplan. Actualmente se está trabajando en métodos de fusión o combinación de ramas. Sin embargo, los que hemos analizado hasta ahora, parecen complicar considerablemente el proceso de planeación.

Planeación a prueba de fallas.

Como se ha visto, la operación de Cassandra se sustenta en la capacidad de determinar, aunque sea indirectamente, la consecuencia de alguna incertidumbre. Sin embargo, esto no siempre es posible y tampoco significa una precondición necesaria para la existencia de un plan viable. En el problema de la bomba-en-el-baño, por ejemplo, hay un plan consistente que Cassandra no puede encontrar: atender ambos paquetes a la vez.

Esto nos lleva a tratar de idear un método para diseñar planes que puedan hacer frente a la incertidumbre cuando la consecuencia no puede ser determinada –lo que se podría llamar *plan a prueba de fallas*. Cada vez que tiene lugar una incertidumbre es posible, en principio, que pueda existir un plan sin contingencias que pueda lograr la meta sea cual fuere la consecuencia de la incertidumbre. Para desarrollar tal tipo de plan, el planeador debe diseñar una versión del plan de contingencia en la que todas las acciones de las ramas de contingencia que provengan de la incertidumbre hayan de ser ejecutadas en forma incondicional. Cassandra se ha ampliado en esa línea, añadiendo un nuevo tipo de decisión, de modo que se ejecuten todas las ramas en paralelo. (Collins & Pryor, 1996). Un plan que contenga tal tipo de decisión puede ser *íntegro* si: a) ninguna de las acciones para lograr la meta en alguna contingencia interfiere con ninguna de cualquier otra contingencia y, b) la habilidad para realizar las acciones es independiente del resultado de la contingencia. Estas condiciones se cumplen claramente en el problema de la bomba-en-el-baño.

Cassandra puede razonar sobre esta posibilidad porque su esquema de identificadores distingue aquellas acciones que no deben realizarse en cierta contingencia, de aquéllas que no necesitan serlo. Es posible ejecutar todas las

ramas solamente si las acciones en cada rama pueden ser realizadas (pero no necesitan serlo) en todas las demás ramas.

Cuando se añade una decisión paralela al plan en la versión ampliada de Cassandra, se añaden nuevas metas en la forma usual, pero la identificación se maneja en forma diferente. Las ramas no están separadas, así que Cassandra nunca va a inferir que las ligas causales en una rama no serán afectadas por las acciones en otra rama.

Falla contingente.

Cassandra puede construir un plan siempre y cuando sea posible lograr la meta del plan en todas las contingencias posibles. No obstante, con frecuencia la meta no puede realizarse sobre algunos resultados de la incertidumbre subyacente. Consideremos el caso en el que se trata de llegar a un centro de ski en coche, cuando el único camino que conduce ahí puede estar libre o bloqueado por nieve. Si el camino está libre, entonces la meta puede lograrse, pero si está bloqueado, todos los planes están destinados a fallar.

No puede esperarse que un planeador reconozca la imposibilidad de lograr una meta en el caso general. Sin embargo, se podría introducir un método alternativo de resolución de condiciones de meta abiertas: simplemente se asume que la meta en cuestión falle.

Este es un método de resolución de condiciones de meta abierta si la meta de hecho puede lograrse, de modo que en teoría los planes que implican una falla contingente debería considerarse únicamente después de que el planeador haya fallado en encontrar un plan en el que todas las metas sean logradas. Algunas veces esto es posible pero, en general, el problema de determinar si hay un plan exitoso es indecidible. Siempre puede haber planes parciales que no impliquen fallas de meta pero que no pueden llevarse a cabo completamente. Por ejemplo, a medida que un plan parcial se modifica va resultando más y más complejo, y la resolución de cada condición abierta va exigiendo la introducción de más submetas no logradas. En este caso, los planes que implican falla contingente nunca serán considerados, a menos que estén clasificados por encima de algunos planes que no implican falla contingente. En general, para lograr alguna utilidad, la dirección tiene que suavizarse: en lugar de considerar una falla de meta únicamente después de

que todas las alternativas han fallado, aplicar restricciones más severas a los planes que involucren metas fallidas. La intención debiera ser: ajustar las restricciones de tal manera que la falla contingente pueda sólo ser aplicable en los casos en que las metas realmente no sean factibles. Sin embargo, esto necesariamente sería un planteamiento heurístico y la completud podría perderse.

Planeación reactiva.

Un planteamiento diferente al problema de la planeación bajo incertidumbre es el que representa el paradigma de la planeación reactiva. En este lineamiento, no se planea ninguna secuencia específica de acciones. Así como en la planeación contingente, se da un conjunto de condiciones iniciales y una meta. Sin embargo, en lugar de generar un plan con ramificaciones, se produce un conjunto de reglas de condición-acción: por ejemplo, planes universales o Reglas de Control Situado (*SCRs*).

En teoría, un sistema de planeación reactivo puede manejar eventos exógenos así como efectos de incertidumbre y condiciones iniciales desconocidas: es posible dar una regla de reacción para cada situación posible que pueda ser encontrada, independientemente de que las circunstancias que llevarían a ellas puedan ser percibidas. En contraste, un planeador de contingencias como Cassandra no puede manejar eventos exógenos dado que no puede predecirlos. Cassandra y otros planeadores concentran sus esfuerzos de planeación en las circunstancias que pueden predecirse como posibles.

Podríamos representar los planes de contingencia de Cassandra como conjuntos de reglas de condición-acción, usando las relaciones causales y las precondiciones para especificar las condiciones en las que cada acción debería ser realizada. Sin embargo, se requiere más razonamiento en el tiempo de ejecución para usar reglas de reacción, del requerido para ejecutar planes de contingencia. En lugar de ejecutar simplemente cada paso en un plan, razonando solamente en los nodos de ramificación, el uso de reglas de reacción requiere la evaluación de condiciones en cada ciclo para poder seleccionar la regla relevante.

Discusión.

Se ha descrito a Cassandra como un planeador de contingencias de orden parcial que puede representar efectos inciertos y construir planes de contingencia para tales resultados. El diseño de Cassandra está basado en una visión coherente de los aspectos que surgen de la planeación bajo incertidumbre. Se observa que, en un mundo incierto, debe hacerse una distinción entre el estado actual del mundo y el modelo que de él posee el planeador; éste ejemplifica intuitivamente un panorama natural de por qué existen las metas de conocimiento y cómo surgen; y basa su tratamiento del plan ramificándolo de acuerdo a los requerimientos del agente. Como resultado, Cassandra planea explícitamente la obtención de información conforme a ciertas acciones en un marco de total generalidad. La coherencia de su diseño provee una base sólida para la implementación de capacidades más avanzadas tales como el uso de procedimientos de toma de decisiones.

Contribuciones.

La principal contribución de un sistema como Cassandra descansa en la representación explícita de pasos de decisión y sus implicaciones en el manejo de metas de conocimiento. Se trata del primer planeador en el que las decisiones son representadas como acciones explícitas dentro de sus planes. Sus metas de conocimiento surgen específicamente de la necesidad de decidir entre cursos de acción alternativos, como precondiciones de acciones de decisión. Cassandra es consistente con la idea de que la planeación es el proceso de tomar decisiones con antelación. En esta postura, los planes de contingencia difieren algunas decisiones hasta que la información en la que éstas se basan está disponible. (Pryor, 1996). Distintas ramas del plan corresponden a distintos resultados de decisión.

A través del uso de pasos de decisión explícitos, Cassandra distingue entre acciones de conocimiento o recopilación de información, por una parte, y de toma de decisiones por otra. Una razón importante para hacer esta distinción es que ésta puede depender de más de un bloque de información, cada uno de los cuales está disponible a través de la realización de diferentes acciones. Además, separar la obtención de información de la toma de decisiones provee una base para introducir métodos alternativos en la toma de

decisiones. Por ejemplo, una ampliación de Cassandra introduciría un tipo de decisión que dirigiría al agente ejecutante a realizar todas las ramas resultantes de una fuente dada de incertidumbre, lo cual permitiría la construcción de planes que podrían tener éxito en situaciones en las cuales no hay modo de determinar cuál es el evento resultante actual (e.g., el problema de la bomba-en-el-baño). La representación explícita de diferentes métodos de toma de decisiones es una línea importante para futuras investigaciones.

Debido a que las metas de conocimiento surgen como precondiciones para las decisiones de Cassandra, la necesidad de conocer si una rama particular del plan funcionará se distingue de la necesidad de conocer el resultado actual de una incertidumbre. Cassandra no planea determinar resultados a menos que sean relevantes para el logro de cualquiera de sus metas. Además, Cassandra no trata las metas de conocimiento como casos especiales: los planes para lograrlas pueden ser tan complejos como los planes para lograr cualquier otra meta. Así como planea para lograr metas de conocimiento que surgen como precondiciones de decisiones, Cassandra también puede generar planes para metas de conocimiento de primer nivel.

Vale la pena notar otras dos características de Cassandra: la flexibilidad generada de su sistema de identificación; y el potencial para aprender y la adaptabilidad debida a su representación de la incertidumbre.

El esquema de identificación de Cassandra, aunque complejo, permite al agente distinguir entre tres clases de acciones: aquéllas que deben ser ejecutadas en una contingencia dada; aquéllas que no deben serlo; y aquéllas cuya ejecución no afectará el logro de la meta en esa contingencia. Esta característica allana el camino para la ampliación descrita arriba, y que permite a Cassandra construir planes que requieran la ejecución de todas las ramas resultantes de una fuente de incertidumbre.

Cassandra no asume nada respecto a la naturaleza instrínseca de la incertidumbre. Una precondición desconocida simplemente denota que la información respecto a qué contexto producirá un efecto particular de una acción, no está disponible al planeador. Puede ser que esta información sea en principio incognoscible; es mucho más factible que la incertidumbre resulte de las limitaciones del planeador o de la información que tenga disponible. En general, un agente operando en el dominio del mundo real será mucho más

efectivo si puede aprender a mejorar su desempeño y a adaptarse a condiciones cambiantes. El uso de precondiciones desconocidas para representar incertidumbre significa que en ciertas circunstancias podría ser relativamente simple incorporar los resultados de tales aprendizaje y adaptación al conocimiento del dominio del planeador. Por ejemplo, el planeador podría descubrir cómo predecir ciertos resultados; podría entonces cambiar las precondiciones desconocidas en otras que reflejaran el conocimiento nuevo. Si, por otra parte, se descubriera que los efectos predichos fueran consistentemente fallidos, se podrían cambiar las precondiciones relevantes por unas desconocidas.

Limitaciones.

Cassandra es uno dentro de un creciente número de planeadores que buscan ampliar las técnicas de la planeación clásica hacia dominios más realistas. Cassandra está diseñada para operar en dominios en los que dos de las tres principales restricciones observadas por los planeadores clásicos son menos estrictas; es decir, se permiten las acciones no determinísticas y un conocimiento incompleto de las condiciones iniciales. Cassandra es, sin embargo, sujeto de la tercera restricción: que los cambios no tienen lugar excepto como un resultado de las acciones especificadas en el plan. Esto limita claramente su efectividad en muchos dominios del mundo real. Además, existen limitaciones en la ampliación del no-determinismo e incompletud del conocimiento que se manejan. Los planes de Cassandra no necesariamente lograrán sus metas si las fuentes de incertidumbre son ignoradas, o si todos los resultados posibles no son especificados.

Cassandra no puede hacer uso de la información acerca de la factibilidad de los resultados particulares, a diferencia de los planeadores probabilísticos; no puede planear para alternar entre planeación y ejecución; y no provee reglas de acción para todas las circunstancias posibles. Únicamente puede resolver problemas para los que hay planes válidos que implican modos de discriminación entre resultados posibles.

El algoritmo descrito tiene dos limitaciones prácticas fundamentales: primera, los planes que genera frecuentemente son más complejos de lo

necesario; y segunda, el tiempo requerido para generar planes imposibilita su uso en todos los problemas excepto los más simples.

La complejidad de los planes de Cassandra proviene de la necesidad de planear para cada contingencia y de la ausencia de combinación de ramas. Por ejemplo, supongamos que se tuviera que abrir una caja de seguridad con combinación para poder obtener dinero para gastar en la tarde. El plan de Cassandra para la meta de disfrutar una tarde podría tener una rama para cada combinación posible. Cada rama podría iniciar con las acciones para abrir la caja, las cuales son distintas para cada combinación, y podría continuar con las acciones para ir a un restaurant y luego al cine, es decir, acciones que podrían ser idénticas en cada rama. Un plan mas simple podría conjuntar las ramas separadas después de haber sido abierta la caja. La consideración de métodos de combinación de ramas es un área para futuro trabajo.

En ciertas circunstancias, como en este ejemplo, la complejidad de un plan podría reducirse mediante el uso de variables de tiempo de ejecución (*run-time variables*). Cuando la única incertidumbre está en el valor que un parámetro de acción toma (que es el caso de la apertura de la caja de combinación), sería posible usar una variable de tiempo de ejecución para representar tal parámetro, obviando la necesidad de separar las ramas del plan. La implementación de esta estrategia podría requerir métodos para determinar cuándo los efectos de incertidumbre están limitados a valores de parámetros. En general, esta noción significa un posible acercamiento al problema de la combinación de ramas: el de tomar un mínimo compromiso de acercamiento a la conjunción de variables, en el mismo sentido que se toma hacia el ordenamiento de pasos en un planeador de orden parcial. Esto daría lugar al concepto de conjunción “condicional” de variables: una conjunción de variables podría ser identificada como requerida o no utilizable en una contingencia dada.

Aunque no se ha analizado totalmente la complejidad del algoritmo de Cassandra, se considera que sería exponencial, debido a las múltiples ramas del plan, cuya presencia no sólo incrementa el número de pasos en un plan sino que incrementa también el número de interacciones potenciales y el número de alternativas para resolverlas. Es difícil encontrar heurísticas de control de búsqueda de dominios independientes efectivas y, en muchos de los

dominios –tan simples- en los que se ha usado a Cassandra, aún heurísticas de problemas específicos son difíciles de abstraer.

Conclusión.

Cassandra es un sistema de planeación basado firmemente en el paradigma de la planeación clásica. Muchas de de sus virtudes y debilidades son como las de otros sistemas de planeación clásicos. Por ejemplo, que bajo ciertas circunstancias sus planes serán válidos y que se garantiza encontrar un plan válido si existe alguno. Sin embargo, las técnicas que utiliza son válidas solamente en ciertas circunstancias, y su complejidad computacional es tal que, pretender un crecimiento directo, no parece ser factible.

Aparentemente, las principales virtudes de Cassandra surgen de la representación explícita de las decisiones en sus planes. Se ha mostrado cómo este uso de las decisiones proporciona una descripción natural de cómo surgen las metas de conocimiento durante el proceso de planeación. También se ha bosquejado cómo las decisiones pueden usarse como base de ampliaciones que proporcionan funcionalidad adicional. Un nuevo tipo de decisión permite elaborar planes a prueba de fallas, los cuales pueden proveernos de un método de resolución de problemas tales como el de la bomba-en-el-baño; y otro tipo de decisión puede proporcionar un método efectivo de planeación y ejecución alternadas.

Por otra parte, puede considerarse que el uso de procedimientos de decisión explícitos permitirá la ampliación del rango de aplicabilidad de las técnicas de planeación clásica. En general, la idea de construir un único plan que tenga éxito en todas las circunstancias es aparentemente improductiva: el mundo real es tan complejo e incierto que, tratar de predecir su comportamiento en detalle, es simplemente imposible. No obstante, el uso de procedimientos de decisión que, por ejemplo, impliquen técnicas probabilísticas o de planeación y ejecución alternas, parece proporcionar un marco flexible que, aunque inevitablemente sacrifique completud y consistencia, proveerá una base para la planeación efectiva y práctica en el mundo real.

Computación y comprensión.

Gödel probó que si se toma un conjunto de axiomas lo suficientemente amplio - que contenga los axiomas de la aritmética como mínimo - no es posible probar, con las armas de deducción del sistema, que tal conjunto sea a la vez consistente y completo.

El artículo, publicado en 1931, en el que Gödel trata sobre este tema se titula “Sobre proposiciones formalmente indecidibles en *Principia Mathematica* y sistemas relacionados”. Trata sobre la imposibilidad de demostrar la verdad o falsedad de ciertas proposiciones dentro de sistemas axiomáticos como el de *Principia Mathematica* o sistemas similares, basados en un conjunto finito de axiomas. *Principia Mathematica* presenta una formulación, por parte de Russell y Whitehead, escrita entre 1910 y 1913, aparentemente completa y consistente del razonamiento matemático. Entendiendo como un sistema lógico completo aquél cuyos teoremas - proposiciones verdaderas - siempre pueden ser demostrables, y consistente, como un sistema lógico que no da lugar a contradicciones, como la de que una proposición resulte ser verdadera y falsa a la vez, lo que Gödel, en la proposición VI de su citado artículo demuestra es que en cualquier sistema axiomático formal existen aseveraciones cuya verdad o falsedad es imposible de decidir desde dentro del propio sistema. Esta afirmación se conoce como el Teorema de Indecibilidad de Gödel. De aquí se derivan dos debilidades de los sistemas axiomáticos: su incompletud, en tanto que siempre será posible encontrar proposiciones indecidibles, así como la inconsistencia, ya que, si el sistema es suficientemente completo - pudiendo demostrar la verdad o falsedad de cualquier proposición - entonces presentará contradicciones, al dar cabida a proposiciones que afirman su propia indemostrabilidad. En resumen, si un sistema axiomático es consistente, entonces es incompleto; y si un sistema axiomático es completo, entonces es inconsistente. De hecho, Gödel demostró que el propio enunciado de la consistencia dentro de algún sistema axiomático - al ser codificado en la forma de una proposición numérica adecuada - es una proposición indecidible, es decir, ni demostrable ni indemostrable con los medios permitidos dentro del sistema.

Lógicos y matemáticos como Cantor, Frege, Hilbert, Russell y Whitehead habían hecho serios intentos por construir un conjunto finito de axiomas y reglas de inferencia que permitiera derivar la totalidad de las matemáticas presentes y futuras. Los resultados de Gödel pusieron fin a esta perspectiva, mostrando que las matemáticas siempre quedarán incompletas y que siempre habrá problemas que no tendrán solución dentro de un mismo sistema axiomático.

Roger Penrose sostiene que una consecuencia evidente del argumento de Gödel es que el concepto de verdad matemática no puede estar inscrito en sistema formalista alguno. La verdad matemática va más allá del formalismo. La decisión que se tome para elegir las reglas adecuadas de inferencia parte de una comprensión "intuitiva" de lo que es "evidentemente verdadero", en virtud de los significados de los símbolos del sistema. La noción de consistencia formal no parece ser la guía para decidir qué sistemas formales son adecuados - conforme a nuestras ideas intuitivas de evidencia y significado - y cuáles no. Se podrían tener muchos sistemas consistentes que no serían razonables en este sentido. Penrose afirma que tales nociones intuitivas seguirían siendo necesarias - aún sin el teorema de Gödel. La noción de verdad matemática estaría más allá del concepto global de formalismo y habría algo absoluto e "infuso" en la verdad matemática. Aunque los sistemas formales son instrumentos muy valiosos dentro de las discusiones matemáticas, sólo actúan como una guía aproximada hacia la verdad pero, "la verdad matemática real va más allá de las simples construcciones humanas".

Tanto Penrose como John Lucas, filósofo de Oxford, coinciden en que algunas facultades mentales deben estar realmente más allá de lo que puede lograrse computacionalmente. Penrose usa el argumento de Gödel para demostrar que la comprensión humana no puede ser una actividad algorítmica. El argumento de Gödel no es un argumento a favor de que haya verdades matemáticas inaccesibles. Lo que afirma es que las intuiciones humanas están más allá del argumento formal y más allá de los procedimientos computables. He aquí la reformulación penrosiana del teorema de 1930:

Los matemáticos humanos no están utilizando un algoritmo cognosciblemente válido para asegurar la verdad matemática.

O bien,

Ningún matemático concreto asegura la verdad matemática solamente por medio de un algoritmo que él sabe que es correcto.

Más filosóficamente, el argumento de Gödel demuestra que cualquiera que sea el punto de vista adoptado, dicho punto de vista no puede ser (saberse) encerrado en las reglas de cualquier sistema formal concebible. Por eso, el teorema de Gödel supuso también un paso capital en la filosofía de la mente, pues demostró que la intuición y la comprensión humanas no pueden reducirse a ningún conjunto de reglas computacionales. Ningún sistema de reglas podrá ser nunca suficiente para demostrar siquiera aquellas proposiciones de la aritmética cuya verdad es accesible, en principio, a la intuición común, de modo que la intuición humana no puede reducirse a ningún conjunto de reglas. Esto sirve de base a Penrose para concluir que debe haber más en el pensamiento humano (físicamente) de lo que puede alcanzar nunca un ordenador, al menos en el sentido de lo que entendemos hoy por "ordenador".

Dentro de esta misma línea y sobre la experiencia de Penrose acerca de las teorías modernas de la física, incluyendo la Relatividad General y la Mecánica Cuántica, este físico-matemático ha propuesto su propio “reduccionismo objetivo” en términos de una mecánica cuántica revisada en virtud de que el ámbito de la mecánica clásica no es suficiente para garantizar algunas actividades mentales, como la compresión de las matemáticas, dado que esta comprensión supone la posibilidad real de entender números no computables. Es decir, debido a que la mente es capaz de entender cosas que no pueden probarse en el ámbito matemático (como lo demuestra el teorema de Gödel) y, dado que las físicas clásica y cuántica responden a procedimientos deterministas y computables - aún dentro del ámbito de la probabilidad estadística de la mecánica cuántica - es posible concluir que la mecánica cuántica ordinaria es inadecuada para explicar la mente.

A diferencia del misticismo, que niega la pertinencia del criterio científico para la búsqueda del conocimiento, Penrose sostiene un criterio científico, solo que tendría que ser revisado y requeriría una ampliación que diera cuenta de la

complejidad suficiente para explicar el enigma de la mente: conciencia, intuición, libertad e inteligencia.

De acuerdo con Penrose, la computación significa un conjunto de operaciones lógicas bien definidas que pueden ser descritas algorítmicamente. Tanto los procedimientos llamados, en el argot computacional, de-arriba-abajo (axiomático-deductivos) como los de-abajo-arriba (procedimientos de aprendizaje). Estos son los llamados "sistemas computacionales", e incluyen los sistemas "caóticos" ya que la aleatoriedad que los caracteriza no significa no-computacionalidad. No obstante, existen ciertos tipos de actividad matemáticamente exacta que están más allá de la computacionalidad. Podría construirse un modelo de un universo físico cuya acción fuera completamente determinista y que estuviera más allá de la simulación computacional.

De acuerdo a lo anterior, existiría algún aspecto de la comprensión que no puede simularse adecuadamente por ningún medio computacional. En consecuencia, no sería idéntico hablar de inteligencia humana e inteligencia artificial. Aunque Penrose apela al libre albedrío como factor esencial de la conciencia en el aspecto activo al buscar el conocimiento, la argumentación en contra de la posibilidad de la existencia de una inteligencia auténtica en algún ordenador estaría suficientemente sustentada en la imposibilidad - en principio - de la no-computabilidad en el ámbito de la llamada "inteligencia artificial". Y, a propósito de esto, Turing, al igual que Yuri Matiyasevich, mostraron que existen ciertos problemas que no tienen solución algorítmica.

En favor de su argumentación podría citarse el hecho de que los mayores fracasos de la Inteligencia Artificial no se ha dado en un ámbito de competencia alcanzable sólo por mentes extraordinarias, sino en el ámbito del sentido común, esto es, en problemas referentes a la determinación de la mejor y más eficaz acción en vistas a un cierto propósito final, donde mientras el sentido común acierta, el cerebro electrónico falla.

Un ejemplo de este tipo de problemas de obviedad o de sentido común lo muestra Daniel C. Denett:[53]

[53] Boden, Margaret (compiladora), Filosofía de la Inteligencia Artificial, F.C.E. México 1994, pág. 167

Había una vez un robot, al que sus creadores llamaron R1, cuya única tarea era valerse por sí mismo. Un día, sus diseñadores hicieron arreglos para que R1 aprendiera que su batería de repuesto, su preciada reserva de energía, se encontraba bajo llave en una habitación, junto con una bomba de tiempo que iba a detonar pronto. R1 localizó la habitación y la llave de la puerta y formuló un plan para rescatar su batería. Dentro de la habitación había una carreta y allí estaba la batería; R1 supuso que con cierta acción que denominó JALAR (CARRETA, HABITACIÓN), podría sacar la batería de la habitación. Actuó de inmediato y logró sacar la batería de la habitación antes de que estallara la bomba. Por desgracia, la bomba también se encontraba en la carreta. R1 *sabía* que la bomba estaba en la carreta dentro de la habitación, pero no se dio cuenta de que al jalar la carreta hacia afuera, la bomba saldría junto con la batería. ¡Pobre R1! Pasó por alto esa implicación obvia de su plan.

La comprensión subyacente a las reglas computacionales parece estar más allá de la simple computación.

Dentro de sus *Collected Works*, el mismo Gödel parece haber considerado que aunque el cerebro debía funcionar computacionalmente, la mente, trascendiendo al cerebro, no estaría limitada a las leyes computacionales. Y, aunque estaba de acuerdo con Turing respecto a que el cerebro se comporta básicamente como un ordenador digital, discrepaba de él en cuanto a que la mente no estaría separada de la materia, considerando esta afirmación como un prejuicio de la época.[54] Por el contrario, Turing sostenía que no era necesaria la existencia de ninguna entidad no física para explicar ciertos tipos de actividad como la intuición matemática, apoyándose en el hecho de que si aun los matemáticos más capaces pueden cometer equivocaciones, ¡habría que esperar que los ordenadores con inteligencia genuina también cometieran errores! Dentro de esta temática, Turing señaló, en su conferencia ante la *London Mathematical Society*, en 1947 que, aunque varios teoremas como el de Gödel expresan que no se puede esperar

[54] Hao Wang, From mathematics to philosophy, Londres, 1974; y Gödel, Collected Works, vol. II, Oxford, 1990, p. 297, citado por M. Boden, *op. cit*., pp. 168 ss.

inteligencia en una máquina que a la vez sea infalible, tales teoremas no dicen nada acerca de cuánta inteligencia podría mostrarse en una máquina que no pretendiera ser infalible.

Penrose considera que la acción física propia del cerebro resulta ser, algunas veces, no computable. Gödel, por su parte, consideraba que no todo pensamiento es computación (precisa o imprecisa) y que habría que buscar puntos débiles en las propias leyes para encontrar lugar a la no computabilidad que está presente en la actividad mental. Y, mientras que Turing sostenía que el pensamiento es una actividad puramente computacional, podríamos aseverar que, según la opinión de Gödel y Penrose, la falibilidad del pensamiento humano no es meramente imprecisión computacional - como diría Turing - sino una suerte de imprecisión no computacional que es la que, precisamente, proporciona al pensamiento humano una potencia mayor que la que pueda derivarse de la algoritmicidad formalmente válida. Y desde el ámbito de la comprensión matemática, a partir de la postura de Penrose, esta imprecisión del pensamiento no es cuestión de validez formal, aunque sí de realidad matemática.

Lenguaje y comprensión.

Desde el punto de vista de la psicología computacional, las lenguas naturales pueden caracterizarse como procedimientos, de tal forma que las oraciones, las frases y las palabras son un tipo de programas computacionales. Para el manejo del lenguaje, estos programas aplicarían un cálculo lógico no interpretado, en base a las conexiones que pueden establecerse, como en el uso del lenguaje natural, no sólo entre las palabras (símbolos) y el mundo, sino entre las palabras y una multiplicidad de procedimientos causales. Los programas de Inteligencia Artificial han demostrado que es posible derivar una fórmula particular bien planteada a partir de las estructuras de datos y las reglas de inferencia, independientemente de la relación existente entre los símbolos y el mundo real. Así, representando la oración en castellano "Leonardo es el padre de María", un programa podría dar como respuesta 'LEONARDO' al dato de entrada 'PADRE[MARIA]' basándose en criterios meramente formales y - aunque la relación de parentesco entre Leonardo y María fuese cierta - sin tener forma de interpretar si este modelo deductivo se

refiriese a gente real. Así, la existencia de una correspondencia entre un cierto formalismo y un dominio de la realidad, no proporciona a determinado ente artificialmente inteligente ninguna comprensión de ese dominio.

El debate acerca de si un sistema (hardware y software) de inteligencia artificial puede ser inteligente o no parece tener su punto álgido cuando se considera si la inteligencia implica comprensión o no. En cuanto no se acepta la identificación entre pensamiento - esto es, pensamiento inteligente - y comprensión o por lo menos la implicación de ésta, la argumentación se desliza sobre el tratamiento de la inteligencia en términos únicamente de manipulación formal de símbolos abstractos (Newell y Simon). El argumento de Searle, llamado el argumento de "Searle en la habitación" arroja luz sobre este punto y

> ... encerrado en una habitación donde hay una serie de fajos de papeles que contienen garabatos; una ventana a través de la cual la gente puede pasarle más papeles con garabatos y por la que él puede entregar papeles; y un libro de reglas (en inglés) que le indican cómo aparejar los garabatos, que siempre pueden identificarse por su forma y figura. Searle pasa el tiempo dentro de la habitación manipulando los garabatos de acuerdo con las reglas.
>
> Una de estas reglas le indica, por ejemplo, que cuando le pasen *güiri güiri* entregue *guara guara*. El libro de reglas también contiene secuencias más complejas de pares de garabatos, donde sólo el primero y el último pasos mencionan la transferencia de papel hacia adentro o hacia afuera de la habitación. Antes de encontrar alguna regla que le ordene directamente entregar una tira de papel, deberá localizar el garabato *plonque* y compararlo con un garabato *plonque*, en cuyo caso el resultado de la comparación es lo que determina la naturaleza del garabato que entregue. Algunas veces tendrá que hacer muchas de estas comparaciones de un garabato con otro y las consiguientes selecciones de garabatos, antes de encontrar la regla que le permita transferir algo hacia afuera.
>
> En lo que concierne a "Searle en la habitación", estos *güiris* y *guaras* no son más que garabatos sin sentido. Sin embargo, aunque él no lo sabe, se trata de caracteres chinos. Las personas que se encuentran fuera de la habitación, que son chinas, las interpretan como tales. Más aún, los modelos que entran y salen por la ventana son interpretados por ellos como *preguntas* y *respuestas* respectivamente: las reglas son tales que la mayoría de las preguntas tienen su contraparte, ya sea directa o indirectamente, en lo que ellos reconocen como una respuesta sensata. No obstante, el propio Searle (dentro de la habitación) ignora todo esto. (Tomado de M. Boden, *op. cit.*, pp. 105-106).

tiene mucha relación con la cuestión del manejo del lenguaje. Se trata de un experimento mental en el que Searle se imagina a sí mismo...

Según Searle, este experimento es una representación concreta de un programa de computadora que, como lo muestra, realiza una manipulación formal de modelos o símbolos no interpretados: utiliza reglas de sintaxis y está fuera de toda semántica. La argumentación de Searle, sobre este experimento, en contra de la posibilidad de alguna actividad de inteligencia, pensamiento o comprensión por parte de un sistema computacional, radica en el hecho de que el proceso de preguntas y respuestas no consiste en responder las preguntas. "Searle en la habitación" no está realmente respondiendo, dado que no *comprende* las preguntas.

Como se puede apreciar en el experimento mental anterior, las argumentaciones a favor y en contra de la posibilidad de la inteligencia artificial se sustentan en las nociones que puedan tenerse acerca de "el entendimiento" o "la comprensión". Los procesos que subyacen a la actividad mental, tales como los procesos neuronales son producto de interacciones bioquímicas y resultan tan poco inteligentes como los procesos físicos que realiza un simple termostato para regular la temperatura. La contrargumentación de Margaret Boden sobre el ejemplo de "Searle en la habitación" se apoya en este tenor y aporta algunas precisiones muy importantes en torno al problema del lenguaje y la comprensión en Inteligencia Artificial. La autora[55] hace énfasis en el hecho de que "un programa de computadora es un programa para una computadora: cuando el programa corre en el hardware adecuado, la máquina hace algo en consecuencia (de aquí que en las ciencias de la computación se utilicen palabras como "instrucción" y "obedecer")." En el nivel del código de máquina, nivel al que deben ser traducidas las instrucciones establecidas en un nivel de más alto nivel, a través de un cierto programa de computación, es el nivel en el que una instrucción dada produce una operación única. Las instrucciones programadas no son meros modelos formales. El lenguaje de programación es el medio tanto para expresar representaciones como para causar la *actividad representativa* de ciertas máquinas. Para una computadora, la representación es una actividad antes que una estructura. En contra de

[55] M. Boden, *op. cit.* pp. 116 - 120

Newell, para quien los símbolos son signos formales manipulables, Hofstadter considera que una máquina como el cerebro no manipula símbolos, sino que es el medio en el que flotan los símbolos y se activan entre sí. El enfoque conexionista visualiza las representaciones y las relaciones entre símbolos como aspectos dinámicos. En este tono, haría falta una teoría adecuada de la semántica de los lenguajes de programación, según lo ha señalado el científico de la computación B. C. Smith[56]. Según este autor, las confusiones existentes en las ciencias cognitivas y la Inteligencia Artificial sobre fenómenos tales como intencionalidad, manipulación de símbolos y representación del conocimiento, se han dado fundamentalmente debido a una distinción teórica demasiado radical entre las funciones de control de un programa y su naturaleza como un sistema sintáctico formal, es decir, entre una estructura representativa declarativa y el código de procedimiento que es ejecutado directamente por una máquina computadora. Indudablemente, no es lo mismo visualizar la expresión "PADRE[MARIA]" como representación denotativa de una relación de parentesco, que visualizarla como una expresión que active a la computadora a localizar el dato "LEONARDO". Sin embargo, Smith propone que se adopte una "teoría unificada" de los lenguajes de programación que dé cuenta tanto de los aspectos denotativos como de los procedimentales. De aquí se derivaría una puntual precisión en cuanto a la semántica de los lenguajes de programación, de tal forma que se considere que éstos tienen una semántica causal, más que denotativa. Es decir, los programas de computación poseen consecuencias inherentes de procedimiento. En la semántica causal, el significado de los símbolos se sustentaría en los vínculos causales, tanto con fenómenos externos como con otros símbolos o estructuras simbólicas. Estos vínculos llevarían a la construcción de otros símbolos y a la activación de otras instrucciones. Desde el punto de vista de la computadora, la interrelación existente entre sus estructuras simbólicas, en base a los programas con los cuales funciona, constituiría lo que, desde el punto de vista humano, se denominaría "entendimiento". Y aunque pudiera pensarse que tal entendimiento fuera tan mínimo que no debería utilizarse tal vocablo en lo que a las computadoras se refiere, no se justifica, según Boden, que "Searle en la

[56] citado por Boden, *op.cit.*., p. 117

habitación" sea considerado como una representación concreta de un programa de computación careciendo de toda comprensión. La pregunta "¿cuándo entiende algo una computadora?" sería tan paradójica como preguntar "¿cuántos granos de arena son un montón de arena?"

Teoremas limitativos e inteligencia artificial.

Los teoremas de Gödel han puesto la piedra angular respecto a otros teoremas que representan limitaciones a los sistemas formales. Si bien los trabajos de Gödel apuntaron a un sistema formal específico - el de Russell y Whitehead - que pretendía ser la formalización axiomática de la aritmética, las consecuencias para las investigaciones en Inteligencia Artificial y, en general, para las ciencias cognitivas, se han hecho patentes como limitaciones cruciales. Esa piedra angular descubierta por Gödel es la noción de autoreferencialidad, y es alrededor de la cual giran otros teoremas tales como el de Church y el de Tarski, pasando por el de Turing.

Como ya se ha vislumbrado en el apartado anterior, lo que llamamos autoconocimiento constituye la premisa fundamental, de principio, para las posibilidades de lo que se podría denominar, propiamente, inteligencia artificial. Y es que la noción de autoreferencialidad funcionaría como la conceptualización, técnicamente hablando, de las nociones un tanto vagas de autoconocimiento o autocomprensión y nos conduce por un derrotero en el que lo que está en juego no es la inteligencia, la cual, en principio, aparecería sin limitaciones para ser artificialmente reproducida, y aun en mayor medida, que la humana, sino en el problema de la representación de los procesos involucrados en el conocimiento. Esta problemática parece condicionar la eficiencia de un sistema computacional al problema de la representación integral de sí mismo. "Todos los teoremas limitativos de la metamatemática y de la teoría de la computación insinúan que, una vez alcanzado determinado punto crítico en la capacidad de representar nuestra propia estructura, [...] se cierra toda posibilidad de que podamos representarnos alguna vez a nosotros mismos en forma integral. El Teorema de Incompletud, de Gödel; el Teorema de la Indecibilidad, de Church; el Teorema de la Detención, de Turing; el Teorema de la Verdad, de Tarski: todos ellos tienen las resonancias de

antiguos cuentos de hadas, advirtiéndonos que 'perseguir el autoconocimiento es iniciar un viaje que. . . nunca estará terminado, no puede ser trazado en un mapa, nunca se detendrá, no puede ser descripto'.[57]

El teorema de Church parte de la tesis de que los procesos mentales pueden ser simulados por un programa de computadora siempre y cuando todas las funciones recursivas finitas puedan ser programadas. Este teorema, formulado en 1936 por el lógico estadounidense Alonzo Church, establece que no hay método infalible que discrimine entre teoremas y no teoremas en un sistema axiomático que contenga, por lo menos, la formalización de la teoría de los números. Este teorema resulta de la mayor importancia en la filosofía de la matemática, del cerebro y del pensamiento porque se apoya en el factor común a estas tres áreas, y que consiste en la suposición crucial de que la teoremidad, la propiedad de una proposición de ser un teorema es, no solamente expresable, sino representable mediante alguna fórmula de teoría de los números.

Tres son, entonces, las nociones básicas que enmarcan el teorema de Church: una, la recursividad, como condición de la tesis de Church; la segunda, la representabilidad como el supuesto sobre el que descansa la afirmación del teorema de Church. Ambas nociones, a su vez, se apoyan en una tercera, en el proceso por el cual una proposición de teoría de los números se convierte en una afirmación (o negación) acerca de sí misma aunque en distintas versiones pero conservando su identidad genuina: autoreferencialidad. Este proceso, conocido como "gödelización", implica la aplicación de las reglas de formalización, un número indefinido - aunque finito - de veces (recursividad) sobre una proposición candidata a ser teorema hasta llegar al momento en que tal proposición, ya gödelizada, no es sólo una proposición que habla acerca de sí misma (autoreferencialidad), sino una que, afirmándose o negándose a sí misma, es ella misma (representabilidad) una proposición de teoría de los números. Es en ese momento cuando es, o no es, un teorema.

Pues bien, como se verá en seguida, el problema surge por la suposición de que le teoremidad puede ser representada por alguna fórmula de teoría de los números:

[57] *Idem*, p. 827

Supongamos que una proposición G dice: “G no es un teorema”. (“G” es la inicial de Gödel para hacer referencia al autor de este proceso de Gödelización). Supongamos que G es un teorema, es decir, es una fórmula bien formada dentro del sistema formal. Luego, en virtud de que la teoremidad es supuestamente representable, al fórmula que afirma “G es un teorema” sería, ella misma, un teorema. Pero esta fórmula resulta ¬G, la negación de G, de modo que el sistema formal resulta incoherente. Por el contrario, si suponemos que G no es un teorema; entonces, nuevamente debido a la supuesta representabilidad de la teoremidad, la fórmula que afirma “G no es un teorema” sería un teorema del sistema formal. Pero esta fórmula es G, lo que nos lleva de nueva a la paradoja. Como se ve, el problema se genera por la suposición de que la teoremidad es representable. Entonces, si suprimimos tal suposición, tendríamos que concluir que ningún proceso recursivo puede distinguir entre números Gödel de teoremas y de no teoremas, lo cual es la negación de la tesis de Church y, con ello, la confirmación del Teorema del mismo autor. En referencia al ámbito de la Inteligencia Artificial, este resultado conduciría a la conclusión de que no existe método que pueda habilitar a los seres humanos a distinguir con seguridad entre teoremas y no teoremas.

Por lo que se refiere al Teorema de Tarski (publicado en 1933), es importante señalar, como lo muestra el Teorema de Gödel, que la capacidad para distinguir la teoremidad no se identifica, necesariamente, con la capacidad para distinguir entre enunciados verdaderos y falsos. Pero más allá de eso, ambas alternativas son imposibles. En efecto, en forma análoga al interés de Church, el de Tarski se refiere a la posibilidad de existencia de una proposición tal como “la fórmula cuyo número Gödel es *a* expresa una verdad”. Este teorema afirma que es imposible contar con un procedimiento de decisión para aplicar a las verdades teórico-numéricas - en caso de que existieran. La autoreferencialidad resulta nuevamente el punto clave para esta afirmación aun cuando la finalidad de Tarski consistía en proponer una definición satisfactoria de la “verdad”. Como se muestra en su trabajo, publicado en 1933, *The concept of truth in formalized languages*, Tarski distingue primero entre los nombres de los enunciados y los enunciados en sí, para evitar precisamente la autoreferencialidad. La premisa fundamental de este lógico y matemático polaco, nacido en 1902, es la noción semántica de la verdad, caracterizando a

la semántica como algún tipo de relación entre las expresiones de un lenguaje - un lenguaje exacto, formalizado - y los objetos a los cuales se refieren esas expresiones. Pero el término "verdad" no consiste en una relación entre expresiones y objetos, sino en una propiedad de algunos enunciados. La definición de "verdad" en Tarski pretende tener el calificativo de satisfactoria para caracterizarla como una propiedad de los enunciados. Un enunciado será verdadero si es satisfecho por todos los objetos, y falso en caso contrario. Para su tratamiento de la verdad fue necesario realizar su fundamentación de la metalógica y de la metamatemática, con el fin de evitar paradojas y antinomias como la paradoja del mentiroso, en la que se genera la afirmación de una oración falsa, análogamente a la proposición G, de la que se trató más arriba.

De acuerdo al análisis que hace Tarski para descubrir las causas de esas paradojas, se muestra que: (1) el lenguaje ordinario contiene tanto las expresiones como los nombres de esas expresiones, así como términos de naturaleza semántica tales como "verdadero"; y (2) en ese lenguaje valen las leyes ordinarias de la lógica. Este análisis lo lleva a proponer la elaboración de un meta-lenguaje en el que se pudiera formular la definición de la "verdad", y que se distinguiera del lenguaje-objeto, el cual estaría incluido en el primero. Además, el meta-lenguaje habría de contener variables de tipo lógico superior al de las del lenguaje objeto, esto es, el meta-lenguaje pretendería ser un lenguaje exacto, formal. Y aunque Tarski proseguiría con su propósito de evadir las paradojas y poder establecer una definición de la "verdad" dentro de este lenguaje formal, cualquier formalización conllevaría - como lo ha mostrado Gödel - a la inconsistencia o a la incompletud enraizada en la autoreferencialidad.

La máquina de Turing.

Otra de las dificultades que presenta el modelo de la mente/cerebro como una estructura algorítmica interactuante consigo misma a través de varios niveles formales, es el relativo a la garantía que puede tenerse respecto a la finitud de los procesos algorítmicos. Esta es una manera de referirnos al llamado "problema de la detención", formulado por el matemático inglés Alan

Turing, uno de los pioneros de las computadoras, entre 1935 y 1936. Esta formulación se refiere a la caracterización misma de lo que se entiende por algoritmo. Básicamente, el concepto de "algoritmo" puede ser entendido como una "máquina de Turing". Esta máquina ha de entenderse como un elemento de la matemática abstracta y no como un objeto físico. Este concepto fue introducido por Turing para afrontar el llamado *Entscheidungsproblem*, planteado por el matemático alemán David Hilbert, en 1900, y que se conoce como el "décimo problema de Hilbert". Este problema consistía en tratar de encontrar un procedimiento algorítmico, mecánico, general para resolver cuestiones matemáticas. La máquina de Turing pretendía precisar lo que se podía entender como "procedimiento mecánico".

Una máquina de Turing es una máquina matemáticamente idealizada consistente en un conjunto finito de posibles estados diferentes, es decir, estados internos. Además, este dispositivo requiere de la capacidad de manejar un *input* de cualquier magnitud. Esta máquina ideal dispondría también de un espacio ilimitado de almacenamiento externo, esto es, "papel" para sus cálculos. Por último, este dispositivo sería capaz de producir un *output* de tamaño también ilimitado. La máquina operaría los datos - inputs y outputs - examinando sólo los datos pertinentes para un momento determinado. En el espacio de almacenamiento externo, el papel, podría ir anotando los resultados importantes de las operaciones realizadas, e ir pasando a las siguientes etapas predeterminadas en función al resultado final buscado. (Una división un tanto simple entre parte interna y parte externa de la máquina de Turing sería análoga a la que existe, en las computadoras actuales, entre hardware y software, respectivamente).

El espacio de almacenamiento externo fue visualizado por Turing como una cinta de longitud ilimitada y con cierta anchura suficiente para que el dispositivo hiciera marcas en ella. Esta cinta marcada sería leída por el dispositivo, cuando fuera necesario, y movida - como parte de la operación - hacia adelante o hacia atrás, haciendo nuevas marcas o borrando las anteriores en función de cada paso en la operación, de tal manera que la cinta funcionaría como espacio externo - impresión - y como medio de *input* (de resultados intermedios del proceso de cálculo). Por supuesto, la cinta funcionaría también como el medio para expresar el *output* final. Mientras éste

no fuera alcanzado, la máquina continuaría con el proceso de cálculo haciendo pasar la cinta indefinidamente hacia adelante o hacia atrás. Por el contrario, alcanzado el resultado final, esto es, finalizado el cálculo, la cinta se detendría quedando expresado el resultado en la cinta.

La cinta fue concebida por Turing como una secuencia lineal de cuadros, algunos marcados y otros en blanco, es decir, puede ser analizable en términos de elementos discretos. Además, dado que en la realidad los cálculos, el *input* y el *output* son finitos, debe haber en la cinta un número finito de marcas reales.

Por otra parte, por lo que se refiere a la "mecanicidad" del proceso de cálculo, puede observarse que el comportamiento del dispositivo, en cualquier instante, estaría determinado en base a su estado interno, conjuntamente con el *input*.

Este procedimiento mecánico es el que se identifica, según las ideas de Turing, con lo que se conoce como algoritmo, y fue formulado para dar respuesta - como se ha indicado al comienzo de este apartado - al problema planteado por Hilbert, consistente en la determinación de un procedimiento mecánico general para responder a todas las cuestiones matemáticas dentro de un marco suficientemente amplio y bien definido - como la teoría de los números. Turing reformuló este problema en términos de poder garantizar que una determinada máquina de Turing *se detendrá o no* cuando opere sobre un cierto número. Esta formulación se conoce como el *problema de la detención*. En base a la abstracción matemática idealizada de Turing, cabe pensar en la codificación de todos los algoritmos posibles, uno por cada máquina de Turing, en un lenguaje binario - de unos y ceros - que representaran la codificación de las instrucciones o pasos predefinidos (es decir, el algoritmo) para operar sobre un número determinado.

Como ejemplo consideremos el caso de formular un algoritmo - diseñar una máquina de Turing - que tenga por objeto simplemente sumar la unidad a cualquier número natural. Sea el caso del número 167 que, en código binario sería 10100111. Para sumar 1 a un número binario hay que encontrar el último 0, reemplazarlo por 1 y luego cambiar todos los 1's siguientes por 0's. De esta forma, el número 10100111 + 1 resultaría 10101000. Este número corresponde, en notación decimal, al 168. (Nótese que el número resultante es

igual que el original hasta las cuatro primeras cifras - de izquierda a derecha - y que, a partir de la quinta cifra, el 0 cambió por 1 y los tres últimos 1's cambiaron por 0's.)

Ahora bien, dado que el algoritmo debe ser aplicable a *cualquier* número natural, la máquina de Turing que suma la unidad al número dado estaría representada por la siguiente lista de instrucciones:

$$_{0}0 \rightarrow {}_{0}0d,$$
$$_{0}1 \rightarrow {}_{1}1d,$$
$$_{1}0 \rightarrow {}_{0}0d,$$
$$_{1}1 \rightarrow {}_{10}1d,$$
$$_{10}0 \rightarrow {}_{11}0i,$$
$$_{10}1 \rightarrow {}_{10}1d,$$
$$_{11}0 \rightarrow {}_{0}1\text{ALTO},$$
$$_{11}1 \rightarrow {}_{00}0i,$$
$$_{100}0 \rightarrow {}_{101}1i,$$
$$_{100}1 \rightarrow {}_{100}1i,$$
$$_{101}0 \rightarrow {}_{110}0d,$$
$$_{101}1 \rightarrow {}_{10}1d,$$
$$_{110}1 \rightarrow {}_{111}1d,$$
$$_{111}0 \rightarrow {}_{11}1d,$$
$$_{111}1 \rightarrow {}_{111}0d$$

Cada instrucción representa un paso determinado y codifica tres elementos básicos: a) el estado interno de la máquina de Turing, representado por los números subíndice a la izquierda de los dígitos 1 ó 0; b) los dígitos posibles que conformarían, en lenguaje binario, el número input al que se le sumará la unidad y, c) el movimiento de la cinta, a la izquierda o a la derecha (i o d). También, como se puede observar, cada instrucción representa una posibilidad que, lógicamente, corresponde a una proposición de la forma si *p* entonces *q* ($p \rightarrow q$). Esto es, dado un estado interno y un dígito en el *input*, entonces ejecuta el cambio a otro estado (o la permanencia en el mismo), así como el cambio o no de dígito.

Como se aprecia en la lista anterior, la séptima instrucción indica la detención del proceso (ALTO) que, para este algoritmo, se daría en el

momento de encontrar el último 0 y cambiarlo por 1. Hay que hacer notar que, dado que en la formulación de Turing, la cinta tiene movimiento, a la derecha o a la izquierda, la consecuencia para el - digámoslo así - "dispositivo lector" de la cinta es leer el dígito a la izquierda o a la derecha respectivamente a dicho movimiento. También es importante señalar que los estados internos - ocho en este algoritmo - están numerados, en código binario, del 0 al 7, es decir, del 0 al 111.

La lista de instrucciones representada anteriormente puede ser, ella misma, codificada en notación binaria. En otras palabras, la lista de instrucciones puede ser representada - al igual que el número particular (167) sobre el que se opera - como una cadena también de 1's y 0's. Para ello se requeriría codificar los símbolos d, i, ALTO, flecha ($\rightarrow$) y la coma (,) mediante números en código binario. El objetivo de esta codificación de las propias instrucciones es poder construir una matriz en la que cada fila represente cada máquina de Turing particular para un fin determinado y que, por estar codificada binariamente, pueda ser "leída" por una máquina de Turing en un nivel superior. Se trataría de construir, mediante esa matriz, todos los *inputs* posibles para una meta-máquina de Turing que, para el propósito esencial de afrontar el problema de Hilbert, estaría evaluando el comportamiento de todas las máquinas de Turing posibles representadas, cada una, en cada fila de la matriz, como *input* para dicha meta-máquina.

Antes de pasar al problema de la detención, que es la formulación de Turing del problema de Hilbert, se presenta la lista de instrucciones anterior, ya codificada binariamente, como una de las filas de la matriz mencionada:
1010110101001011010100111010010110101111010000111010010101110100010111010100011010010110110101010110101010110101010100.

Esta codificación contiene ya, binariamente, toda la lista de instrucciones incluyendo tanto los dígitos del número decimal 167 como los símbolos de operación (lectura y ejecución).[58]

El problema de la detención.

[58] Roger Penrose, La mente nueva del emperador, FCE, México 2002, p. 74.

Esta codificación de la lista de instrucciones para una máquina de Turing arbitraria, *T*, en una cadena de 0's y 1's en una cinta - la cinta infinita del dispositivo - es la idea básica de una máquina universal de Turing. Esto es, si consideramos que la parte inicial de la cinta, el input, corresponde a la lista codificada de instrucciones de alguna máquina de Turing Universal *U*, que actúa sobre el resto del input de la misma forma que lo haría *T*. La máquina universal de Turing es un imitador universal. La parte inicial de la cinta provee a la máquina *U* de toda la información necesaria para imitar a cualquier máquina *T*.

En vista de lo anterior, es posible construir una lista de todos los algoritmos posibles, es decir, de todas la máquinas de Turing que pueden operar sobre cualquier número natural. De esta forma, en una matriz de n filas por m columnas, las filas representarían los inputs para la máquina U - todas las máquinas T - y las columnas representarían todos los números naturales. El propósito esencial de esta matriz es decidir si T_n - la enésima máquina de Turing - dará o no una respuesta al aplicarse a un número *m*.

Si T_n no da una respuesta entonces T_n nunca se detendrá.

La formulación de Turing del problema de la detención consiste en determinar si existe un procedimiento algorítmico para determinar, automáticamente, si una máquina de Turing se detiene o no.

Turing demostró que no existe tal algoritmo, como se verá a continuación.

Si suponemos que tal algoritmo *sí* existe, entonces hay alguna máquina de Turing que "decide" si la *n*-ésima máquina de Turing, al actuar sobre el número *m*, se detendrá o no. Entonces su *output* es la cinta numerada 0 si no se detiene y 1 si lo hace. Se puede simbolizar esta alternativa de la siguiente forma:

$$H(n;m) = \begin{cases} 0 \text{ si } T_n(m) = \square \\ \\ 1 \text{ si } T_n(m) \text{ se detiene} \end{cases}$$

La tabla correspondiente a todos los *outputs* de todas las posibles máquinas de Turing aplicadas a todos los números *m*, sería como la siguiente tabla infinita.

n ***m***	0	1	2	3	4	5	6	7	8 .....
0	**0**	0	0	0	0	0	0	0	0
1	0	**0**	0	0	0	0	0	0	0
2	1	1	**1**	1	1	1	1	1	1
3	0	2	0	**2**	0	2	0	2	0
4	1	1	1	1	**1**	1	1	1	1
5	0	0	0	0	0	**0**	0	0	0
6	0	0	1	0	2	0	**3**	0	4
7	0	1	2	3	4	5	6	**7**	8
8	0	1	0	0	1	0	0	0	**1**
.									
197	2	3	5	7	11	13	17	19	23.....
.									

En esta tabla los 0's indican que la máquina T_n no se detiene. Esto es posible sobre el supuesto de que la máquina *H* existe, ya que entonces la misma *H* permite reemplazar los □'s por 0's. Este reemplazamiento se consigue precediendo la operación de T_n sobre m por el cálculo *H(n;m)*. Este procedimiento puede expresarse en la forma

$$T_n(m) \times H(n;m).$$

Continuando con la suposición de que *H* existe, las filas constituyen secuencias infinitas computables, es decir, sucesiones infinitas de números que pueden ser generadas por un algoritmo. En este caso, los algoritmos son las máquinas de Turing *T*.

Esta tabla implica que toda secuencia computable de números naturales debería aparecer en alguna de sus filas (o en varias).

Por otra parte, si denominamos Q al procedimiento que genera toda la tabla, podemos escribir

$$Q(n;m) = T_n(m) \times H(n;m)$$

Por último, con ayuda de un artificio matemático ideado por Georg Cantor, llamado “corte diagonal” se puede generar, a partir de la diagonal principal de la tabla (marcada con negritas), una nueva secuencia que, aunque debiera estar en alguna de las filas, no puede realmente existir. En efecto, si a la diagonal principal de la tabla, que es la secuencia 0, 0, 1, 2, 1, 0, 3, 7, 1,..., sumamos 1 en cada uno de sus elementos se obtiene la secuencia:

1, 1, 2, 3, 2, 1, 4, 8, 2,...

Se trata de una secuencia computable, y derivada de una tabla también computable, que es la expresada como $1 + Q(n;n)$, esto es,

$$1 + T_n(n) \times H(n;n)$$

dado que en la diagonal *n* es igual a *m*.

Ahora bien, si la tabla contiene *todas* las secuencias computables, la nueva secuencia generada a partir de la diagonal principal, que debería estar incluida en la tabla, no puede estar debido a que difiere de la primera fila en el primer dígito, de la segunda fila en el segundo dígito, de la tercera en el tercer dígito, y así sucesivamente. Esta contradicción manifiesta que la suposición de que la máquina *H* existe, es falsa.

Otra manera de visualizar esta contradicción es, suponiendo que *H* existe, entonces hay algún número de máquina de Turing, por ejemplo w, tal que se tiene

$$1 + T_n(n) \times H(n;n) = T_w(n).$$

Sustituyendo *n* por *w* se obtiene

$$1 + T_w(w) \times H(w;w) = T_w(w)$$

donde, si $T_w(w)$ se detuviera, se obtendría la relación imposible

$$1 + T_w(w) = T_w(w)$$

(ya que *H(w;w)* sería 1); por el contrario, si $T_w(w)$ no se detuviera (con *H(w;w)* = 0), se obtendría la relación igualmente inconsistente

1 + 0 = □

De aquí el *teorema de la detención*: no existe un algoritmo universal para decidir si una máquina de Turing se detendrá o no.

Los teoremas limitativos de la metamatemática y de la teoría de la computación parecen apuntar, según se podría apreciar a partir de lo expuesto en el capítulo anterior, que existen serias dificultades para lograr la representación integral de la mente/cerebro. Construir una entidad artificial dotada de inteligencia - esto es, inteligencia humana - implica dotarla de comprensión, la cual implica, a su vez, autoconocimiento.

Establecidas estas nociones en el concepto de autoreferencialidad, son portadoras de la problemática derivada de este concepto, en términos de paradojas y conflictos entre niveles de formalización. De forma análoga a la demostración de Gödel respecto a los sistemas formales, en cuanto a que tales cuerpos axiomáticos no pueden ser simultáneamente consistentes y completos, Hofstadter señala las dificultades inherentes a la cuestión de la autorepresentación, es decir, a la autocomprensión. La analogía tendría lugar en tanto que, arribando a ciertos niveles "críticos" de autocomprensión, surgen los cimientos mismos socavados en forma gödeliana, conduciéndonos a la incompletud. En otras palabras, la búsqueda de coherencia hace aparecer cuestiones indecidibles dentro de ese ámbito de coherencia.

Inversamente, si partimos del supuesto de que la coherencia no sea una de nuestras virtudes, entonces pocas esperanzas quedan para la comprensión. Los teoremas limitativos muestran que el punto álgido en el problema de la autocomprensión, identificado como el problema de la autorepresentación, es el de la autoreferencialidad, esto es, la referencia a la mismidad recursivamente. En esto han consistido las demostraciones de los distintos teoremas: referirse a lo que se afirma o se niega considerándolo, desde el nivel de objeto y, pasando a otros niveles, como forma o estructura conteniendo al objeto primario. De aquí la concepción de un modelo artificialmente inteligente operando en base a diferentes niveles de formalización con lenguajes correspondientes de altos y bajos niveles; desde pseudocódigos hasta los

llamados lenguajes de máquina; desde niveles de diseño algorítmico hasta niveles de interpretación y ejecución (que, en el caso de la representación de la estructura de la inteligencia humana, correspondería a los niveles neurofisiológicos o cuánticos).

Desde una perspectiva que pudiera abarcar un ámbito de especulación acerca de la inteligencia, que considere las limitaciones manifestadas por los teoremas expuestos y que, al mismo tiempo, vaya más allá de ellos, es oportuno considerar las puntualizaciones de algunos de los autores más relevantes al respecto. Éstas se refieren a la posibilidad real de que las limitaciones en cuanto a la representación de la inteligencia humana, no tengan su fundamento en cuestiones de autoreferencialidad o - como lo expresa Hofstadter - de "retorcimientos gödelianos", sino a cuestiones de falta de comprensión, propiamente, del funcionamiento real de la inteligencia.

Como apunta Penrose[59], el teorema de la detención, de Turing, no establece que existan algunas cuestiones matemáticas indecidibles. Este teorema no demuestra que exista alguna tabla generada por una máquina de Turing particularmente complicada para que, en principio, sea imposible decidir si tal máquina se detendrá o no en cuanto se aplique a cierto número. Lo que afirma el teorema no versa sobre la insolubilidad de problemas particulares, sino sobre "la insolubilidad algorítmica de familias de problemas". A lo que apunta el teorema es a la imposibilidad de saber de antemano *cuál* es el algoritmo que necesitaríamos utilizar para aplicarlo a un caso concreto; algoritmo que para ese caso sí existe. De aquí que Penrose haga notar que, finalmente, los algoritmos no deciden, por sí mismos, acerca de la verdad matemática; la decisión acerca de ella y la determinación de la validez algorítmica recae, en última instancia, en la inteligencia humana.

Por otra parte, Hofstadter[60] señala, a propósito de las cualidades que deberían constituir la inteligencia, que "la cuestión central es que no pueden existir razones *fundamentales* (esto es, gödelianas) que hagan de tales cualidades algo incomprensible: pueden ser algo perfectamente comprensible a los ojos de seres más inteligentes." La prueba de Gödel sugeriría que,

[59] R. Penrose, *op. cit*., p.89
[60] D. Hofstadter, *op. cit*., p.839

mientras ciertos hechos podrían ser explicados en los niveles altos, podrían no ser explicados, "*de ningún modo*", en los niveles bajos. En esta forma, podría haber alguna panorámica de alto nivel en la mente/cerebro que abarcaría conceptos que no podrían existir en los niveles más bajos: las facultades explicativas del nivel alto no existirían - ni en principio - en los niveles inferiores. Para dar cabida a esta posibilidad es oportuno recordar que la no teoremidad de G tiene una explicación, la cual está basada en la necesidad de comprender no un nivel por vez, sino en la forma como un nivel refleja su metanivel, así como las implicaciones de tal reflejamiento. Desde esta noción del funcionamiento de la mente/cerebro, quedaría por entender cómo es que los niveles superiores abarcan a los inferiores y, al mismo tiempo, están determinados por éstos: una especie de "resonancia" de refuerzo recíproco entre los diferentes niveles.

Estas puntualizaciones sugerirían que la concepción cabal de la inteligencia es una cuestión, por una parte:

a) de la apertura para aceptar que las actividades más sutiles de la inteligencia, entre ellas la cuestión de la verdad matemática, son de naturaleza *no algorítmica*, y
b) del progreso que se pueda lograr en las investigaciones acerca del mecanismo subyacente al reflejamiento existente entre los niveles neuronales con ciertas "señales" o fenómenos de nivel intermedio, y de éstos, a su vez, con el nivel simbólico, incluido el "símbolo del yo". Estas investigaciones - tarea fundamental de la Inteligencia Artificial, de la teoría de la computación y las áreas cognitivas - arrojarían luz acerca de los llamados "fenómenos emergentes", tales como las ideas, esperanzas, analogías, el libre albedrío e incluso, la conciencia.

****

VII. CONCLUSIONES

Husserl nos señala que la conciencia es flujo de totalidad y unitario, percepción interna y representación. Por otra parte – y a la vez - los símbolos constituyen la realidad inmediata a la conciencia humana. Ésta, como dadora de sentido, no afronta al mundo físico sino sobre y desde una perspectiva netamente simbólica, tal como lo afirma Cassirer. Esta urdimbre simbólica, lejos de representar un obstáculo –podría pensarse- para el establecimiento de una relación directa entre el hombre y la naturaleza, resulta ser *el* medio de comprensión del mundo objetivo. La perspectiva simbólica es el modo de acción de la conciencia humana, y es la que fundamenta todo lo que el mundo significa para el hombre, y todo lo que éste significa para sí mismo. Una relación "directa" sólo tiene sentido, precisamente, bajo el ámbito simbólico y, siguiendo a Husserl, aquello que directamente aparece a la conciencia, los fenómenos de la conciencia, son lo más inmediato a ella.

Podría decirse que la realidad "más inmediata" de los fenómenos de la conciencia, que no son sino símbolos, y la realidad "externa" del mundo, son – para el hombre- una y la misma cosa. La ciencia contemporánea, cada vez más, nos presenta un Universo Simbólico cuya realidad, por ende, sólo puede ser entendida desde la inmediatez fenoménico-simbólica de la conciencia. La física no se ocupa ya directamente de lo existente como lo materialmente real, sino de su "estructura", esto es, de su textura formal. La tendencia hacia la unificación ha prevalecido sobre la tendencia hacia la intuitivización; la síntesis, encauzada por los conceptos puros de ley ha resultado ser superior al resumen por medio del concepto de cosa. Con ello, el orden se ha convertido en el verdadero concepto fundamental "absoluto" de la física; el mundo mismo no se presenta ya como un agregado de cosas, sino como un orden de "eventos". Como medio para que la física construya el mundo exterior – así formula Weyl la situación – no deben servir ni el espacio ni el tiempo de la intuición, sino un *continuum* de cuatro dimensiones en sentido aritmético abstracto.(...) Lo que queda es finalmente una construcción simbólica en el sentido en que Hilbert la lleva a cabo en la matemática. El significado objetivo de los símbolos intelectuales de la matemática y de la ciencia natural no proviene de los objetos

trascendentes que se encuentran tras de ellos, sino a través de su rendimiento, de la función de “objetivación” que se efectúa en ellos. (Cassirer)

Por lo que se refiere al enfoque simbólico-computacional, de acuerdo con Newell y Simon, la comprensión de la inteligencia se basa en la descripción de sistemas de símbolos físicos, en términos de designación e interpretación. En otras palabras, la manera de plasmar y hacer comunicable toda relación o proceso inteligente, es a través de estructuras simbólicas, cuyos componentes nos remiten a objetos físicos y cuyas relaciones son interpretables. Es así como el conjunto de símbolos adquiere inteligibilidad, esto es, sentido para la conciencia.

Pero hay que señalar, siguiendo a Cassirer, que tal sentido, tal inteligibilidad del símbolo proviene del trato con un “espacio significativo” que ha dejado atrás tanto al espacio meramente “expresivo” (mítico), como al “espacio representativo” (lingüístico). Nuevamente: las cosas – entidades, procesos, funciones - son lo que significan. El *logos* geométrico de la ciencia traspasa y trasciende lo “dado”. Ese hecho lógico es también el que condiciona y propiamente hace posible el contenido de cuerpo “físico” y de evento “físico”, en la medida en que se toma a estos conceptos no en su sentido substancial sino en su sentido funcional, considerándolos no en primer término como expresión de una simple existencia o acaecer, sino como expresión de un cierto orden, de un cierto modo de consideración. Es aquí donde el enfoque simbólico-computacional puede afinar el trato con los símbolos, para pasar de un trato mecanicista a un trato más holístico. La llamada “programación orientada a objetos”, de alguna manera ya implícita en su práctica, contiene la semilla de este *logos*, de este sentido geométrico-significativo; al considerar la simbolización de entidades tales como funciones, eventos, propiedades como los referentes de la función operatoria de la inteligencia.

Hay, al parecer, una relación indisociable entre símbolo e inteligencia, lo cual nos lleva a recordar la pregunta de McCulloch, “¿qué es un símbolo, en tanto que puede ser usado por la inteligencia, y qué es la inteligencia, en tanto que puede utilizar un símbolo?” Se puede afirmar que el símbolo, para ser tal, requiere *necesariamente* de inteligibilidad ya que, careciendo de ésta, sería

algo sin sentido e irrelevante. Pero se podría cuestionar si la inteligencia, para ser tal, requiere *necesariamente* ser simbólica.

En términos fenomenológicos, la intencionalidad de la conciencia y el horizonte de significatividad, aportan el sentido y el contexto a los fenómenos de la conciencia, de tal manera que pueden ser interpretados. La conciencia inteligente no conoce otro modo de dar forma y estructura al ámbito fenoménico –simbólico- si no es en términos de designación e interpretación. Por otra parte, en estos mismos términos, aplicados a la descripción de los sistemas de símbolos físicos, tiene lugar, según Newell y Simon, la comprensión de la inteligencia. La manipulación de símbolos, en una forma sistemática, es lo que constituye el procesamiento de información.

La capacidad para manipular y almacenar símbolos es, según McCulloch, uno de los requisitos estructurales de la inteligencia. Y es sobre esta aproximación, basada en las manifestaciones de la inteligencia, como este autor trata con ella. No se trata de buscar un "principio de inteligencia", como no se conoce un "principio de vida", sino de concretizar la actividad que puede calificarse como inteligente, a través del manejo sistemático de símbolos, en cualquier sustrato físico. Si bien es cierto que no es posible demostrar lógicamente la *necesidad* de contar con un sistema de símbolos físicos donde se exhiba inteligencia, el análisis fenomenológico de la conciencia humana muestra este hecho. Y en esto consiste el segundo tipo de prueba de la hipótesis de los sistemas de símbolos físicos. El primer tipo de prueba, según se ha visto, consiste en confirmar la *suficiencia* de los sistemas de símbolos físicos para generar inteligencia.

La búsqueda de esta suficiencia representa para la ciencia cognitiva uno de los principales focos de apoyo por parte de la fenomenología y, en general, de la filosofía. Aunque esta perspectiva simbólica de la inteligencia no busca, como se ha dicho, encontrar algún principio del que emane la inteligencia, basado en los sistemas de símbolos físicos, no por ello resulta poco aventurado afirmar la suficiencia de éstos para generar aquélla. Parece indudable que la inteligencia trabaja, se hace patente, mediante sistemas de símbolos; ella almacena y manipula sistemáticamente símbolos sobre la base de la designación y la interpretación (no en términos sustancialistas, como ya se ha indicado, según Cassirer). Esta afirmación no es equivalente a la de

afirmar que los sistemas de símbolos físicos sean suficientes para la inteligencia.

La aproximación fenomenológica nos describe, nos muestra la dinámica de la conciencia. No está dentro de sus objetivos el explicarnos la conciencia ni, menos aún, sus relaciones con la inteligencia. Esto es, la conciencia es, de suyo, inteligente; ambas se funden en una sola entidad. La fenomenología no nos muestra una entidad llamada inteligencia, aparte de la conciencia. El ser conscientes es nuestro modo de ser-en y de ser-con el mundo y, en este modo de ser va implícito el carácter inteligente de relacionarnos con el Universo.

En otros términos, es la conciencia la que tiene un modo de ser que llamamos: inteligente. Conforme a la información recopilada en el presente trabajo, no parece plausible afirmar que algún sistema de símbolos, sobre algún sustrato físico como puede ser una computadora, agote en su dinámica una actividad equivalente a la inteligencia. Es factible decir, en cambio, que tales sistemas de símbolos físicos participan en la exhibición de tal actividad.

Los sistemas que han sido descritos, tanto para aprendizaje-razonamiento, como para toma de decisiones, *FLARE* y *CASSANDRA*, apuntan al hecho general de mostrar, en sus respectivos objetivos, una característica importante de un ser conciente-inteligente: la autonomía. Ésta se basa en la dinámica propia del sistema de símbolos; en la "vida" propia del símbolo. Y este dinamismo está forjado en la combinación que Newell y Simon han propuesto: una teoría sintáctico-formal del simbolismo con una teoría causal de la semántica. De acuerdo a una definición causal de los términos semánticos, el conjunto de cambios que un símbolo permite al sistema realizar, en respuesta a un determinado estadio –interno o externo- constituye la noción de significado. El nivel de acoplamiento dinámico de un símbolo dentro de todo un sistema –su causalidad semántica-, propicia en mayor o menor grado la autonomía, la autosuficiencia del sistema, ya sea para aprender o para decidir, según el caso descrito. Esta característica puede tomarse como una medida de la capacidad de un sistema de símbolos físicos para adaptarse al ámbito prácticamente impredecible del mundo real.

Metafóricamente, podría decirse que la vida de un símbolo radica en su interpretabilidad, en su ámbito semántico. El sistema de símbolos físicos, como una sola entidad, vive en tanto es autónomo, esto es, en la medida en

que su estructura se conforma por relaciones intersimbólicas flexibles, en términos de sintaxis. La perspectiva de Newell y Simon aporta la pauta para las investigaciones en torno a la autonomía de los sistemas de símbolos físicos, en base a la causalidad semántica.

El lograr elaborar una clase de estructura de símbolos de esta naturaleza, parece ser más claramente visualizado desde una perspectiva analógica. En términos cuantitativos-discretos, en términos simbólicos-atómicos un modelo tal insinúa un grado de complejidad no fácilmente descriptible. En cambio, un modelo basado en unidades de estructuras simbólicas, que puedan dar cuenta de relaciones múltiples y simultáneas, así como de grados de interpretación no discretos, sería un modelo basado en representaciones analógicas.

La insuficiencia de los sistemas de símbolos físicos para generar actividad inteligente se ha debido, por una parte, a que los modelos basados en ellos han sido dirigidos básicamente hacia el juego de la imitación de la inteligencia humana y, por otra parte, a que la interpretación, el modelaje de las estructuras de símbolos no ha considerado suficientemente el matiz analógico de la semántica que dé cuenta del espacio de significatividad requerido fenomenológica y simbólicamente.

La aproximación mediante la interacción de niveles a través de relaciones simbólicas y reflejamiento, sustentadas en la noción de autoreferencialidad, es un modelo de la mente/cerebro donde el aspecto semántico se proyecta encontrando un papel fundamental. La significación puede tener lugar en dos o más niveles diferentes de un sistema que opere símbolos si el reflejamiento de la realidad, isomórficamente, ocurre en alguno de esos niveles. Este reflejamiento daría entrada a un matiz metafórico de la significación que, aun estando enmarcado en un sistema con un sustrato rígido -el hardware- y racionalmente descriptible -como el nivel neuronal del cerebro-, arrojaría luz a la sospecha, al parecer no infundada, de que un sistema computacional, infalible, no es apto para generar significatividad.

Y es que, según se ha visto en este trabajo, la búsqueda de racionalidad ha hecho patente la irracionalidad -si ésta puede oponerse a lo metafórico- que ha puesto a la investigación en ciencias cognitivas en una clara alternativa: o se acepta la irracionalidad o se resigna a las limitaciones que ofrece un

tratamiento racional en cuanto a la posibilidad de crear artificialmente un nivel convincente de inteligencia. La investigación actual, a este respecto, no está satisfecha con los juegos de imitación logrados desde los sistemas lógico-deductivos, casi infalibles en el ajedrez, hasta los sistemas de búsqueda heurística para resolución de problemas o los sistemas expertos con capacidad de aprendizaje. Las habilidades humanas más cruciales -aun pareciendo triviales- como el sentido común, no han podido ser abarcadas, y se han mostrado como un factor básico para la toma de decisiones, el aprendizaje y la generación de conocimiento. La irracionalidad aparece como un halo de misterio que las rodea.

No obstante que la cuestión de la irracionalidad no sea sino un asunto de grado de entendimiento, no hay razón para creer que el hardware infalible de una computadora no pueda dar sustento a un comportamiento simbólico de alto nivel, en el que tuvieran lugar estados complejos como equivocaciones, confusiones, olvidos o la apreciación de la belleza. Bajo la premisa de que la irracionalidad se produce en el nivel más alto, y no el infalible nivel más bajo, confiable y lógico, la conducta visible -racional e irracional- se proyecta, para las subsecuentes investigaciones, como un paradigma en la representación de la mente/cerebro. La inteligencia no reside en el nivel inferior sino en el nivel en el que reside también la irracionalidad: en el nivel superior. Y es en este nivel donde tienen efecto los teoremas limitativos que se han mostrado.

No existen, en principio, limitaciones para la imitación de la inteligencia humana. Es la noción misma de inteligencia –y ésta es una de las muchas aportaciones de las ciencias de la computación y cognitivas, en general– la que protagoniza y dirige el rumbo de las investigaciones. Se trata de una noción que, dentro de una de las posturas fundamentales de la IA, conforma un binomio con la de la irracionalidad. Otra postura igualmente importante mantiene una esfera fina, distintiva de la inteligencia humana, más allá del formalismo y de procedimientos algorítmicos. Una ha aceptado la irracionalidad como indisociable compañera; la otra quiere mantenerse en un ámbito de cierto misticismo más parecido al reino de las musas que a la tierra del esfuerzo intelectual.

No es necesario tomar partido por una o por otra: ¿cómo *inteligirse* a sí mismo si uno mismo se manifiesta caprichoso, irracional? O bien, ¿tiene sentido *tratar* de inteligir cuando basta sólo con mirar?

Lo que es indudable es que en el intento mismo de la imitación propia se da un mejor autoconocimiento, se dimensiona lo desconocido, y se valora más lo que se es.

BIBLIOGRAFÍA

Agre, Philip E., the soul gained and lost: artificial intelligence as a philosophical project, *SEHR, volume 4, issue 2: Constructions of the Mind*

Boden, Margaret. Filosofía de la Inteligencia Artificial, F. C. E., México, 1994.

Cassirer, E. Antropología Filosófica. Introducción a una Filosofía de la Cultura, F.C.E, México, 1963.

Cassirer, E. & Morones, A. (1998). Filosofía de las formas simbólicas (2ª ed.). México: FCE.

Collins, A. and Michalski, R. S., *The Logic of Plausible Reasoning: A Core Theory*, Cognitive Science, Vol. 13, 1989.

Corvez, Maurice. La Filosofía de Heidegger, F.C.E., México, 1981.

Crosson, Frederick J. Filosofía y Cibernética, F.C.E., México, 1982.

Dreyfus, Hubert L. What Computers Can't Do, Harper, New York, 1972.

García-Madruga, J.A. (1992). Introducción a la edición española, en Rumelhart, D.E., Mclelland, J.L. y el grupo PDP (1992). *Introducción al procesamiento distribuido en paralelo.* Madrid: Alianza Editorial.

Güven Güzeldere y Stefano Franchi: *SEHR, volume 4, issue 2: Constructions of the Mind.*

Hofstadter, D., Gödel, Escher y Bach: una eterna trenza dorada, Conacyt, México, 1982

Hubert Dreyfus, *What Computers Still Can't Do* (Cambridge, MA: MIT Press, 1992)

Hubert L. y Dreyfus, Stuart E., Mind Over Machine , Free, New York, 1986.

Husserl, Edmund. Investigaciones lógicas 1, trad. Manuel García Morente y José Gaos, 1ªreimpresión, Ed. Alianza editorial, Madrid, 2001.

Investigaciones lógicas 2, trad. Manuel García Morente y José Gaos, 1ª edición, Ed. Alianza, Madrid, 1982.

Ideas relativas a una fenomenología pura y una filosofía fenomenológica, tomo I, trad. José Gaos, 4ª. Reimpresión, Ed. Fondo de Cultura Económica, México, 1997.

Ideas relativas a una fenomenología pura y una filosofía fenomenológica, Tomo II, Investigaciones fenomenológicas sobre la constitución, trad. Antonio Zirión, UNAM, México, 1997.

Johnson-Laird, P. N., Herrmann, D.J. and Chaffin, R. (1984). Only connections: a critique of semantic networks. Psychological Bulletin 96, 2.

Kant, Immanuel. "Prefacio a la Segunda Edición", Crítica de la Razón Pura, Porrúa, México, 1998.

Langley, Pat., *Elements of Machine Learning*, San Francisco, Morgan Kaufmann, 1995.

Minsky, Marvin. "A Framework for Representing Knowledge," The Psychology of Computer Vision, ed. Patrick Henry Winston, McGraw, New York, 1975.

Newell, Allen, Intellectual Issues in the History of Artificial Intelligence, *The Study of Information: Interdisciplinary Messages*, ed. Fritz Machlup y Uma Mansfield (New York: Wiley, 1983)

Newell, Allen and Herbert A. Simon, Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM*. vol. 19, No. 3, 1976.

Pollack, Jordan B. Mindless Intelligence, *IEEE Intelligent Systems*, May/June 2006.

Penrose, Roger, La mente nueva del emperador, FCE, México, 2002.

Piaget, Jean. La formación del símbolo en el niño, FCE, México, 1977
Psicología y epistemología genéticas, Ed. Proteo, B. Aires, 1970.

Pryor, Louise y Collins, Gregg, "*Planning for contingencies: a decision based approach"*, *Journal of Artificial Intelligence Research*, Vol. 4, 1996.

Rich, Elaine y Knight, Kevin. Inteligencia Artificial, McGraw-Hill/Interamericana de España, 1994.

Robberechts, L. El Pensamiento de Husserl, F.C.E., México, 1979.

Schildt, Herbert. Utilización de C en Inteligencia Artificial, McGraw Hill/Interamericana, México, 1990.

Schütz , Alfred y Luckman, Thomas. The Structures of the Life-World, Northwestern UP, Evanston, IL, 1973.

Searle, John R. Mentes, Cerebros y Ciencia, Ed. Cátedra, Madrid, 1995.

Sharoff, Serge, "Constructions of the Mind", *Stanford Humanities Review*, vol. 4, num. 2, 1995.

Winston, Patrick H. Inteligencia Artificial, Addison Wesley Iberoamericana, México, 1994.

Winograd, Terry y Fernando Flores, Understanding Computers and Cognition: A New Foundation for Design (Reading, MA: Addison, 1986)

Wittgenstein, Ludwig. *Tractatus Logico Philosophicus*, Alianza Editorial, Madrid, 1997.