

Entrevista a Philip N. Johnson-Laird

JUAN A. GARCÍA MADRUGA
UNED



Palabras clave: *razonamiento, modelos computacionales, conciencia, ciencia cognitiva.*

Dirección: Facultad de Psicología. UNED. Ciudad Universitaria. 28040 Madrid.

INTRODUCCION

El profesor Johnson-Laird nació en el norte de Inglaterra, cerca de Leeds, en 1936. En sus años jóvenes, antes de dedicarse a la psicología, realizó diferentes trabajos entre los que cabría destacar los de músico y crítico musical. En 1961 comenzó sus estudios de psicología en el University College de Londres, donde obtuvo la licenciatura en 1964 y el doctorado en 1967. De aquel tiempo proviene su amistad con Peter Wason quien le dirigió la tesis y con el que comenzó una colaboración que culminó con la publicación en 1972 del libro *Psicología del razonamiento*; un libro en que se recogen las investigaciones realizadas en el campo del razonamiento, utilizando tareas como la de las cuatro tarjetas o tareas silogísticas, y que por su claridad y su riqueza de ideas se ha convertido en un clásico en su campo.

En 1971 estuvo como profesor visitante en el Instituto de Estudios Avanzados de Princeton donde comenzó una colaboración con George Miller sobre teoría del significado, que condujo a la publicación conjunta de *Language and Perception* (1976) y que, indudablemente, ha tenido una notable influencia en su trabajo psicolingüístico posterior.

En 1973 se trasladó a la Universidad de Sussex, en la que fue nombrado posteriormente profesor de Psicología Experimental. En 1974 recibió la medalla Spearman concedida por la Sociedad de Psicología Británica. Desde 1982 desarrolla sus investigaciones sobre lenguaje e inferencia en la Unidad de Psicología Aplicada del «Medical Research Council» en Cambridge, de la que es subdirector. En 1983 publicó *Mental Models*, obra en la que propone una muy atractiva e influyente teoría semántica sobre representación e inferencia y defiende la necesidad de los modelos computacionales en psicología. En el mismo año es nombrado doctor *honoris causa* por la Universidad de Goteborg. Las importantes contribuciones teóricas y experimentales del profesor Johnson-Laird le han convertido en uno de los psicólogos europeos de más prestigio e influencia.

La entrevista fue realizada el 31 de marzo de 1987, en la Universidad de Psicología Aplicada de Cambridge, en donde el autor de la misma realizó una estancia de cuatro meses, disfrutando de una bolsa de estudios concedida por el MEC.

1. La Psicología del Razonamiento, fruto de la colaboración con Peter Wason, que fue publicada en 1972, es un espléndido ejemplo de análisis cognitivo de un campo concreto. ¿Cuáles han sido en tu opinión los cambios más importantes en la Psicología Cognitiva en estos quince últimos años?

R. Yo diría que el cambio más notable en este período ha sido la creciente comprensión de la importancia de construir modelos computacionales. En aquel libro no había modelos computacionales para ninguno de los procesos que postulábamos y aunque comenzábamos a darnos cuenta de que sería una buena idea, sólo utilizábamos diagramas de flujo. Lo que ha sucedido, en esencia, es una revolución metodológica por la que consideramos que para realizar experimentos y construir teorías no es suficiente el tener simplemente cajas con flechas entre ellas. Este es un importante cambio metodológico. La otra cosa que ha sucedido, ya no de tipo metodológico pero que pienso que es probablemente igual de importante, es que

mucha gente en diferentes disciplinas se ha dado cuenta de que, con frecuencia, está trabajando con los mismos problemas. Por ejemplo, si cogemos el estudio del razonamiento, cuando nosotros escribimos el libro, en inteligencia artificial no se estaba haciendo realmente mucho sobre razonamiento; no podíamos conseguir ideas aprovechables provenientes de la inteligencia artificial.

Sin embargo, ahora existen multitud de diferentes tipos de programas para razonamiento en inteligencia artificial. Tenemos, por una parte, toda una tecnología sobre los probadores unificados de teoremas que usan la regla de inferencia de resolución, y por otra parte el trabajo proveniente de Planner y otros lenguajes que implican la utilización de reglas de inferencia de contenido específico que últimamente han tenido un gran desarrollo comercial dentro del campo de los sistemas expertos. Además, actualmente está el trabajo realizado por filósofos interesados en el campo de la psicología del razonamiento como Gil Harman, de Princeton. De esta manera, hoy se encuentra una gama mucho más amplia de disciplinas relacionadas con lo que, podríamos decir, son los problemas de la psicología cognitiva y, al mismo tiempo, una gama mucho más amplia de personas implicadas en materias cognitivas.

2. *Ya en 1977, también con Peter Wason, en el libro de lecturas sobre pensamiento y razonamiento se hablaba de ciencia cognitiva. ¿Por qué ciencia cognitiva y no psicología cognitiva?*

R. Bien, aquel libro trataba de reflejar lo que por 1977 uno podía ver que eran en realidad esos dos diferentes desarrollos de los que he hablado. Así, si miramos los capítulos del libro encontramos un considerable número de ellos que tienen que ver con modelos computacionales y encontramos también que el libro contiene artículos de psicólogos, de gente de inteligencia artificial, de antropólogos, etc. En aquellos momentos la noción de una ciencia cognitiva había aparecido y había sido bautizada en América como tal. Lo que quiero decir es que la gente, de forma completamente deliberada, había decidido llamarla así y nosotros elegimos ese título para reflejar los dos desarrollos de los que he hablado anteriormente, la interrelación entre diferentes disciplinas y la existencia de un mayor número de gente implicado en ellas desde un punto de vista cognitivo.

3. *Sin embargo, la influencia del trabajo en otras disciplinas no es siempre positiva. Por ejemplo, después de la publicación con George Miller de Language and Perception, sostenías la necesidad de una psicolingüística. ¿Qué querías decir con esto?*

R. Bien, aquello fue algo que surgió de la siguiente manera. Para comprender la psicología del lenguaje es una idea realmente buena asumir que comprendemos y producimos oraciones porque tenemos una gramática de nuestro lenguaje en nuestra cabeza; muchos argumentos han sido proporcionados a este respecto por Chomsky y la esencia de su argumentación ha sido postular un tipo particular de gramática, la gramática transformacional, que debería ser hallada en nuestras cabezas. Ahora bien, lo que ha sucedido desde que aquellos argumentos fueron presentados, y estoy pen-

sando en particular en los dos primeros trabajos que Chomsky y Miller hicieron juntos, es que primero la gente se dio cuenta de que las llamadas reglas transformacionales no tenían ninguna realidad psicológica; así, había un grupo de investigadores del ITM, especialmente Jerry Fodor, Tom Bever y Meryl Garret, que decían «de acuerdo eliminamos las reglas transformacionales pero tenemos todavía una buena cantidad de datos a favor de la realidad psicológica de la estructura profunda». Bien, lo que sucedió a continuación fue que el propio estatus de la estructura profunda se convirtió en algo dudoso y, de hecho, actualmente no existen datos que, en forma decisiva y definitiva, demuestren que la gente recupera la estructura profunda de una oración cuando intenta comprenderla o que genera la estructura profunda de una oración cuando trata de producirla. Estos pensamientos estaban en mi mente a mediados de los años setenta y parecía como si uno necesitara en aquel momento una teoría psicológica del lenguaje, una psicolingüística que no estuviera basada en la lingüística. Lo que ha sucedido desde entonces ha sido muy interesante, ya que alrededor de cuatro o cinco lingüistas y sus colegas han propuesto justo el tipo de teoría que un psicólogo necesita; a saber, una teoría que no tiene ni reglas transformacionales ni estructura profunda; me estoy refiriendo a lo que a veces ha sido llamado gramática de estructura de frase generalizada. Entre este tipo de teorías que no tienen transformaciones están las propuestas por Gerald Gazdar y Michel Brame y la gramática funcional de Ron Kaplan y Joan Bresnan. Asimismo, están las ideas propuestas por Tony Adiz y Mark Steedman acerca de la forma canónica de las oraciones y las variaciones aceptables en ellas, sin utilizar tampoco las transformaciones.

Por tanto, aquel título está un poco anticuado aunque era casi un grito de desesperación. Dicho de otra manera, ahora tenemos teorías lingüísticas que desde el punto de vista de la producción y comprensión de oraciones tienen mucho más sentido.

4. *Centrándonos en el razonamiento deductivo. ¿Qué piensas acerca de la habilidad lógica de los humanos? ¿Existe una lógica mental?*

R. Es una pregunta muy oportuna dado el seminario que acabamos de tener. Pienso que hay que distinguir dos cuestiones completamente separadas. En primer lugar, si la gente tiende o no a ser lógica. En segundo lugar, suponiendo que sean lógicos cuál es el mecanismo que les permite serlo.

5. *De acuerdo, veamos la primera.*

R. En cuanto a si la gente es lógica, yo diría que sí. Pienso que en conjunto la gente es lógica, aunque por otra parte es realmente fácil, tanto en la vida cotidiana como en el laboratorio, encontrar situaciones en que la gente se comporta en forma no lógica, es decir, hacen juicios y toman decisiones irracionales; la gente realiza inferencias inválidas y el criterio no es lo que algún lógico o algún teórico de la decisión diga que es irracional, sino que en otras circunstancias tú puedes lograr que las mismas personas que cometen un error consideren que es ciertamente un error. Lo que quiero destacar es que se pueden observar inferencias inválidas bastante fácilmente. Ahora bien, existen diferentes psicólogos y filósofos que hacen el

siguiente tipo de argumento, dicen «bien, lo que sucede en estos ejemplos es que la gente aporta información extraña o no toman en consideración toda la información que les das»; tienen una teoría que sostiene que es la interpretación incorrecta de las premisas la que produce el abandono de los cánones lógicos.

6. *Sí, la teoría de Mary Henle.*

R. Exactamente. Pienso que este tipo de posiciones son falsas y me gustaría poner un ejemplo que sé que te es familiar. Si presentas a la gente silogismos en el laboratorio, encuentras que ciertos silogismos son muy fáciles y la gran mayoría de los sujetos realizan inferencias válidas, mientras que otros silogismos son muy difíciles y no encuentras gente o muy raramente encuentras gente que pueda con tales silogismos; tienden a realizar inferencias inválidas y no es en absoluto fácil, no digo que sea imposible, pero no es en absoluto plausible explicar su fracaso en términos de una interpretación incorrecta de las premisas. Resumiendo mi respuesta a la primera pregunta, pienso que todos poseemos una razonable competencia lógica; es decir, en principio sabemos lo que es razonar válidamente y, hablando en términos generales, cómo hacerlo en la práctica, pero todos nosotros estamos inclinados a cometer errores. Ahora bien, como ya dije si asumimos que tenemos algún grado de competencia lógica, entonces el problema es cuál es el mecanismo y si existe una lógica mental.

7. *Esa era la segunda parte de mi pregunta.*

R. Bien, pienso que la respuesta es que la mejor explicación de la forma en que la gente razona, aun cuando lo haga válidamente, no exige postular reglas formales de inferencia del tipo que se puede contrar en la lógica. Es interesante por qué tantos psicólogos, especialmente en el pasado, han querido explicar nuestra habilidad lógica diciendo que tenemos en nuestras cabezas reglas sintácticas formales de inferencia como el *modus ponens* y pienso que la razón de esto es, en parte, un accidente histórico. Sucede que cuando la lógica moderna comenzó su desarrollo, y estoy pensando en la obra de Frege en adelante sobre el cálculo cuantificacional y el cálculo proposicional, los lógicos que lo hicieron pensaron y, ciertamente, desarrollaron métodos puramente formales. Así, ellos estaban interesados en expresar todas las reglas de inferencia que permitieran derivar inferencias válidas. No fue hasta la obra de Tarski en los años treinta cuando los lógicos comenzaron a preocuparse por cómo podrían dar una interpretación semántica al cálculo lógico y no fue hasta los primeros sesenta, y estoy pensando en la obra de Saúl Kripke, cuando conseguimos tener una buena teoría semántica para las muchas y diferentes lógicas modales que pueden ser formuladas. Como vemos la lógica procedió en este caso de una forma un poco extraña; comenzó en primer lugar de forma puramente sintáctica, después desarrolló una semántica y, por tanto, la metalógica se interesó primero en ver cómo la noción de validez, que era un noción semántica, puede realmente ser captada por el cálculo formal. En otras palabras, el interés estaba en si una formalización particular es, utilizando una palabra técnica, completa; es decir, posee las reglas formales correctas que captarán todas

y sólo aquellas inferencias que la semántica del sistema diga que son válidas. Así, la validez significa simplemente que la conclusión debe ser verdad si las premisas son verdad. Hay que destacar que el uso de la palabra verdad es la que nos dice que es una cuestión semántica. Cuando utilizamos un cálculo formal con reglas de inferencia como el *modus ponens*, etc., entonces no estamos realmente manipulando cosas que son verdad y cosas que son mentira, lo que hacemos, simplemente, es manipular sintácticamente cadenas de símbolos para derivar expresiones y lo que la metalógica trata es de asegurar que la semántica y el mecanismo sintáctico se ajusten perfectamente. Por tanto, pienso que la razón por la que los psicólogos han quedado enganchados con la idea de que el único mecanismo que nos permitirá razonar válidamente es una regla formal de inferencia, es que cuando gente como Piaget comenzó a trabajar en estos asuntos era la única lógica que podían conocer.

Entonces, ¿por qué considero que postular una lógica mental es una mala idea? Bien, por varias razones. Primero, es un hecho muy bien conocido que el contenido de un problema de razonamiento puede afectar las conclusiones que la gente extrae y, obviamente, tal efecto es totalmente incompatible con cualquier simple aplicación directa de lógica formal, ya que lógica formal significa precisamente que se puede olvidar el contenido. De modo que esto plantea una dificultad superficial que, evidentemente, se puede intentar explicar si se cree en la lógica mental. Otra razón, que ya hemos visto, que supone una dificultad adicional es como mencioné en el caso de los silogismos, que la gente comete inferencias erróneas y es necesaria una explicación adecuada si están siguiendo reglas formales de inferencia.

8. *Estas son las dos razones que se suelen aducir en contra de las tesis de la lógica formal.*

R. Sí, pero existe otro punto que a menudo se olvida y que, si no me equivoco, no fue nunca apreciado por Piaget y sus colaboradores. Si tú das a alguien una pareja de premisas como «si llueve entonces ella se mojará» y «llueve», la inmensa mayoría de la gente extraerá la conclusión «ella se mojará». El lógico mental explicará esto diciendo que se ha utilizado el *modus ponens*. Pero de hecho la lógica sola no puede explicar la actuación de la gente ya que a partir de cualquier conjunto de premisas existen siempre un infinito número de conclusiones válidas que se siguen lógicamente. Así, la conclusión que se podría extraer es simplemente «llueve y ella se mojará» en la que se ha formado una conjunción y que es una conclusión perfectamente lógica, aunque un poco extraña. Vemos, por tanto, que existe un amplio consenso entre la gente en el tipo de conclusiones lógicamente válidas que extraen y se necesita algo más que la mera lógica para explicar este consenso. Es necesario explicar por qué la gente no extrae este tipo de conclusiones bastante extrañas que se siguen lógicamente. Tenemos, por tanto, tres principales problemas: los efectos del contenido, el hecho de que la gente razona inválidamente, y el hecho de que la lógica nos dirá sólo qué tipo de conclusiones se pueden extraer. Estas son las razones por las que no pienso que exista un lógica mental.

9. *Si no es mediante reglas formales de inferencia, ¿cuál es el mecanismo que utilizan los sujetos en su razonamiento? ¿Puede ser a través de reglas que contienen conocimientos específicos o esquemas?*

R. Bien, ésa es ciertamente una posibilidad y ya mencioné anteriormente en la entrevista el hecho de que los sistemas expertos son programas computacionales en los que el conocimiento se representa mediante un número muy grande de reglas condicionales que tienen contenido específico. Ciertamente una teoría semejante explicaría los efectos del contenido, pero dudo mucho que pudiera explicar el conjunto del problema, y la razón es la siguiente: si se tienen reglas específicas de inferencia, entonces se necesitaría un número endiablidamente grande de ellas para explicar la habilidad lógica de la gente. Pienso que lo que necesitamos es un sistema que sea sensible al contenido pero que no esté íntimamente ligado a él, de manera que pueda explicar cómo podemos realizar inferencias con materiales que nunca nos hemos encontrado anteriormente en nuestra vida. La gente es capaz de razonar en áreas en las que no tiene ningún conocimiento real y nosotros debemos ser capaces de explicar este hecho. Por tanto, debemos rechazar también el que todas las reglas de inferencia deban tener un formato de contenido específico.

10. *La teoría de los modelos mentales pone el acento en la influencia de la comprensión en el razonamiento, en la influencia de los factores semánticos.*

R. Sí, en esencia, la mejor explicación que conozco sobre cómo la gente razona, sostiene que ni utilizamos reglas formales de inferencia ni reglas de contenido específico, aunque tomemos en consideración el contenido cuando razonamos. La gente utiliza el siguiente principio semántico: cuando les presentamos unas premisas pueden construir una representación del estado de cosas descrito por las premisas; si hay alguna conclusión informativa, es decir, que no sea algo que haya sido dado explícitamente en las premisas, entonces existe al menos un posible candidato para realizar una inferencia. En ese momento la gente sabe que la única manera de que tal conclusión sea válida es que no exista otro posible modelo de las premisas en el que la conclusión sea falsa. En otras palabras, lo que estoy suponiendo, que en absoluto es algo incompatible con la noción general de lógica como empresa intelectual, es que la gente razona de forma semántica. Comprenden las premisas y extraen de ellas una conclusión semánticamente informativa, y entonces se aseguran de que la inferencia es válida realizando una prueba semántica directa de su validez mediante la búsqueda de contraejemplos, es decir, buscando otros modelos en los que las premisas sean todavía verdad pero la conclusión sea falsa. Así, el conjunto del proceso está guiado semánticamente, aunque no se usen ni reglas formales de inferencia como el *modus ponens*, ni reglas de contenido específico. Es como si no hubiera reglas de ninguna clase, o, si queremos llamarlas así, una clase de reglas definitivamente diferente, como en el caso del conocimiento del concepto de validez que capacita para la búsqueda de contraejemplos de las posibles conclusiones.

11. *Además de la habilidad para comprender el lenguaje y para construir un modelo mental del estado de cosas necesitamos, por tanto, una segunda habilidad que es la de buscar contraejemplos. Pero esta habilidad no parece ser ni general, ni fácil de adquirir y ¿no está en contradicción con la existencia de sesgos confirmatorios en el pensamiento humano?*

R. Bien, efectivamente pienso que no es fácil de adquirir, ni general y una parte de las razones por las que creo que esta teoría proporciona una buena explicación del razonamiento es porque si se consideran los errores que la gente comete en el razonamiento entonces se puede dar una buena explicación de los mismos en términos de una insuficiente búsqueda de contraejemplos. Esto sucede en forma muy clara en el caso de los silogismos.

Estoy completamente de acuerdo en que la habilidad de buscar contraejemplos no es necesariamente fácil de conseguir, requiere una gran cantidad de capacidad de la memoria en funcionamiento. Pero hay otra razón por la que pienso que existe un sesgo confirmatorio, que es una pequeña parte de una larga historia que tiene que ver con la forma en la que, en mi opinión, la gente realiza inferencias inductivas. Si queremos definir lo que es una inferencia deductiva, podríamos decir que una inferencia deductiva no añade nunca información a las premisas. Si pensamos un poco en ello vemos que debe ser así ya que la conclusión tendrá que ser verdad en todos los estados de cosas en que las premisas puedan ser verdad, y en tal caso, entonces claramente no hemos añadido ninguna información nueva. Ahora bien, una de las características del razonamiento cotidiano es que muy a menudo no es posible realizar una deducción con alguna certeza; muy a menudo tenemos de hecho que ir más allá de la información de las premisas y extraer una conclusión que añade información semántica. Déjame poner un ejemplo que tú ya conoces pero que puede ser ilustrativo. Supongamos que formas parte de un jurado en un juicio por asesinato, en el que se ha establecido más allá de toda duda razonable que el sospechoso viajaba en el expreso a Edimburgo, cuando ocurrió el crimen en un cine, en el que la víctima fue apuñalada. Si te pregunto por la conclusión que extraerías de estos hechos, entonces me dirías que el sospechoso debe ser inocente, tal y como Tony Anderson y yo hemos hallado que casi todo el mundo dice, al menos inicialmente. Pienso que ésta es una interesante inferencia ya que probablemente es mucho más típica de la forma en que pensamos en la vida diaria que los silogismos, por ejemplo. Ante todo, la conclusión es informativa ya que ciertamente ni está explícita en las premisas, ni se sigue lógicamente de ellas como veremos a continuación. En realidad añade una gran cantidad de información semántica. Ahora bien, el punto concreto que quiero resaltar es que la teoría general sobre cómo evaluamos inferencias en términos de la construcción de modelos mentales, también se aplica en este ejemplo. Es decir, imaginemos un estado de cosas en que las dos premisas son verdaderas; cuando pensamos en trenes expresos y cines, y éste es el aspecto crucial por el que el contenido es importante, tienes que darte cuenta de que los trenes expresos a Edimburgo no llevan cines y sabemos que la gente realiza tal asunción porque cuando les preguntamos si están seguros de que el sospechoso es inocente dicen: «bien, yo parto de que en los expresos a Edimburgo no hay cines». También tienes que saber cosas tan evidentes como que una persona no puede estar en dos

sitios al mismo tiempo. Así, cuando construyes un modelo no sólo utilizas la información semántica contenida en las premisas, sino que aportas conocimiento general adicional y después extraes la conclusión. Además, cuando insistimos en preguntarles a nuestros sujetos en el experimento si están completamente seguros, entonces suelen ser capaces de pensar escenarios más elaborados en los que el sospechoso es culpable. Así, por ejemplo, sugerirán: «quizá el sospechoso no lo apuñaló personalmente, quizá empleó a alguna otra persona, un cómplice». Y cuando les decimos «eso es perfectamente posible pero de hecho no sucedió así», entonces pueden decir cosas como «quizá había un resorte preparado en el asiento del cine»; y si nosotros decimos «también podría haber ocurrido, pero de hecho tampoco sucedió así», nos dicen «podría haber algún tipo de mecanismo controlado por radio que apuñalase a la víctima»; o como dijo un sujeto «podría haber habido un bloque de hielo encima del asiento de forma que cuando el hielo se derritiera cayera sobre la víctima». Nosotros seguimos diciendo que no fue de esa manera y los sujetos continúan buscando escenarios aún más diabólicos. El más ingenioso de todos, que sólo fue propuesto por dos personas, consistió en la idea de que el sospechoso hipnotizaba a la víctima, dándole la sugestión posthipnótica de que se apuñalase a sí mismo en el momento en que apareciese determinada escena de la película. Como puedes ver es bastante ingenioso.

12. *El crimen perfecto.*

R. Exactamente el crimen perfecto. Ahora bien, cuando decimos a los sujetos «es posible que sucediera así, pero de hecho no fue de esa manera», la mayoría de los sujetos dicen «el sospechoso debe ser inocente, no puedo pensar otra cosa» y si les decimos «¿está usted seguro?», ellos responden «bien, confieso que no puedo estar seguro al cien por ciento» y esto nos muestra que evidentemente los sujetos en ningún caso realizan una inferencia deductiva. Como dije antes éste era un largo recorrido para explicar el por qué de los sesgos confirmatorios. La razón es que en la vida diaria es relativamente raro encontrarse en una posición en que tenemos toda la información que necesitamos para realizar una inferencia deductiva, lo que tenemos que hacer es ir más allá de la información dada para realizar conclusiones informativas que, por tanto, no son necesariamente válidas; y para hacer esto el verdadero trabajo consiste en imaginar la solución, más que usar lo que uno sabe como refutación de alguna conclusión. Es un proceso constructivo que requiere imaginación.

Existe otra razón a la que sólo me referiré muy brevemente a menos que quieras que me extienda más. Uno de los mayores logros inductivos que todos los seres humanos alcanza, o casi todos, es el aprendizaje de su lengua nativa; y esto lo logran teniendo relativamente poca información acerca de qué oraciones son no-gramaticales. Lo que quiero decir es que cuando uno es niño oye muchas oraciones no-gramaticales, pero raramente le indican a uno que una oración es no-gramatical. La mayor parte de las veces la información que uno recibe es si la comunicación tiene éxito; si el niño dice algo y la madre no le entiende, lo más probable es que la madre diga «lo que quieres decir es que...». De esta manera se puede argumentar

y alguna gente lo ha hecho así incluyendo Robert Berwick, que en términos generales la gente aprende la gramática de su lengua nativa simplemente enfrentándose a las expresiones, simplemente expresando oraciones bien formadas. Entonces, tendríamos que preguntarnos cómo sería esto posible, lo que nos llevaría a una teoría del aprendizaje informal que provoca una profunda aprensión y que es debida a un hombre llamado Dana Anglevin. Esta teoría muestra que si puedes ordenar tus hipótesis de forma que comiences con la más informativa acerca de algún concepto o gramática que sea consistente con los datos y puedes, por tanto, ordenarlas en términos de su grado de información semántica, entonces ciertamente puedes aprender sólo de los ejemplos positivos del concepto o gramática. Esto significa, en otras palabras, que va a existir un sesgo hacia los ejemplos positivos. De esta manera pienso que la razón por la que encontramos este sesgo hacia las hipótesis confirmadoras no es ningún tipo de prejuicio irracional, sino que proviene de un conjunto de cosas que ocurren en las inferencias cotidianas.

13. En diferentes ocasiones has argumentado contra la existencia de una lógica innata; ¿cuál es tu opinión sobre una fuerte base innata en el desarrollo del razonamiento?; ¿cómo podemos explicar la adquisición de la capacidad lógica humana?

R. Una lección general que, en mi opinión, se aplica en cualquier aspecto de la psicología cognitiva es que no podemos hacer nada sin postular constricciones innatas muy importantes. En la visión, por ejemplo, Marr elaboró muy buenos argumentos sobre el hecho de que la estereopsis no sería posible, a menos que existan en el sistema nervioso fuertes constricciones acerca de la naturaleza del mundo. Si consideramos lo que es posible aprender, por ejemplo, en el caso de la gramática a la que ya me he referido, está muy claro que no es posible aprender la gramática de una lengua nativa si no existen fuertes constricciones innatas en el mecanismo. Ciertamente, Chomsky diría que son constricciones para todos los posibles lenguajes, y puede ser que esté en lo cierto; muchos otros, como Piaget y el filósofo Hilary Putnam hablarían de constricciones más generales de tipo intelectual. No conozco la respuesta, pero estoy seguro de que a menos de que existan constricciones innatas el aprendizaje de la gramática, por ejemplo, no sería posible. En el caso del razonamiento, pienso que también tienen que existir constricciones innatas, especialmente en lo que se refiere a los aspectos semánticos. Pienso que el organismo tiene que haber sido capaz de comprender el concepto de verdad y falsedad antes de que podamos alcanzar el de inferencia válida; y me parece altamente probable que el concepto de verdad sea algo que la gente adquiere como resultado del desenvolvimiento de algún tipo de progreso evolutivo innato. Por tanto, no estoy contra la idea de factores innatos y de inclinarme hacia algún lado lo haría del lado racionalista, en el de Descartes, como opuesto al empirismo de Locke; aunque realmente no pienso que Locke estuviera tan absolutamente en contra de los mecanismos innatos como algunos racionalistas suponen, pero ése es otro asunto.

14. *Sin embargo, en el pasado has argumentado contra las teorías que postulan una lógica mental totalmente innata...*

R. Sí, sí, pero lo hacía porque era una lógica; lo que quiero decir es que hubiera argumentado en contra aunque hubiese sido adquirida de alguna manera, contra lo que argumentaba era contra el hecho de que el razonamiento dependiera de un proceso sintáctico. Pienso que el único lugar que la sintaxis ocupa en el razonamiento es como ayuda para comprender oraciones, pero no en términos de los mecanismos reales que se usan, es decir en términos de reglas sintácticas de inferencia. Por tanto, no estoy en contra de mecanismos innatos y pienso que un problema empírico interesante es establecer la línea de demarcación; en general, la psicología cognitiva como ciencia no ha tenido mucho éxito en el establecimiento de líneas precisas entre lo que tiene que ser innato y lo que puede ser aprendido, no sabemos suficiente al respecto.

15. *Dejando el difícil problema del innatismo, y centrándonos en las tareas de razonamiento y su relación con la memoria en funcionamiento; ¿cuál es el papel de la memoria en funcionamiento en la inferencia silogística? ¿podría explicar el efecto de la figura?*

R. Yo lo he defendido así, pero diría que me siento menos seguro acerca de este punto que de otros aspectos del razonamiento de los que he hablado previamente, como el uso de modelos, etc. Yo he sostenido que una razón por la que la gente comete errores y por la que ciertos problemas son difíciles, podría ser porque esos problemas exigen la construcción de varios modelos; y podría ser que fuera la memoria en funcionamiento la que produjera esta dificultad al no haber suficiente capacidad para considerar todos esos modelos, cuando tienes que considerarlos todos al mismo tiempo de forma que puedas ver lo que tienen en común. Actualmente me inclino por la opinión de que esto es correcto y que el efecto de la figura puede ser explicado de la misma forma. El efecto de la figura consiste en que cuando das a la gente premisas, por ejemplo, de la forma «algún A es B- todo B es C», entonces tienden a extraer la conclusión válida «algún A es C» en vez de la igualmente válida «algún C es A»; y la razón por la que podemos utilizar la memoria en funcionamiento para explicar esto es, simplemente, por el hecho de que el orden en el que la información entra en la memoria en funcionamiento es el orden en el que tiende a salir de la misma. Así, si te digo «algún A es B», «todo B es C» entonces extraerás una conclusión acerca de A siendo C, más que de C siendo A. Esto es, por tanto, consistente con la idea de que es en la memoria en funcionamiento donde surgen ciertos fenómenos. En primer lugar, produce un cuello de botella que hace difícil considerar diferentes modelos al mismo tiempo. En segundo lugar, parece tener esta característica de que la primera información en entrar tiende a ser la primera en salir.

Otra cosa que sería tentador asumir es que la razón por la que los niños, por ejemplo en el caso de la inferencia silogística, no lo hacen tan bien como los adolescentes o los adultos es por que no tienen una memoria en funcionamiento tan buena. Jane Oakhill y Debbie Bull hicieron estudios sobre las inferencias silogísticas de los niños y hallaron, por ejemplo, que resol-

vían correctamente alrededor del cuarenta por ciento de los problemas de un sólo modelo, mientras que un grupo de estudiantes universitarios brillantes conseguían un porcentaje de alrededor del 90 %. Podríamos decir que esto es algo que tiene que ver con el desarrollo de la capacidad de la memoria en funcionamiento. El inconveniente está en que no hay datos decisivos a favor de que la memoria en funcionamiento realmente se desarrolle. Déjame ser un poco más preciso con respecto a un argumento que se expone actualmente, por ejemplo por Alan Baddeley, y que consiste en que la capacidad básica del bucle de articulación verbal es más o menos constante, lo que sucede es que a medida que creces puedes aprender a articular más rápidamente, de forma que puedes almacenar más. En cualquier caso, aunque sospecho que esto es correcto, no puede explicar nuestros datos con niños porque no pienso que, en general, repitan las premisas. Por tanto, aunque pienso que proporciona una explicación razonable, muy bien puede no ser la última palabra y que no fuera la memoria en funcionamiento la fuente de las tendencias evolutivas. Aún es posible que no fuera la fuente del efecto de la figura o de la dificultad con los silogismos de tres modelos.

16. Volvamos a los problemas más actuales de la psicología cognitiva. Durante años has defendido la necesidad de los modelos computacionales en psicología. ¿Por qué necesitamos modelos computacionales?

R. Bien, existen varias razones. Déjame exponer primero lo que podríamos llamar las razones públicas, las que se pueden escribir e imprimir en los artículos de revistas. Es muy fácil si intentas explicar los fenómenos utilizando simplemente el lenguaje ordinario o el lenguaje técnico de los psicólogos, es muy fácil, digo, pensar que has explicado el fenómeno y no ser consciente de que tu explicación o bien es inconsistente, o bien tiene «agujeros vacíos» dentro de ella. Existen muy buenos ejemplos históricos de ello, que si quieres más adelante podemos volver sobre ellos. Déjame asumir por ahora que esto es verdad, que si intentas desarrollar una teoría en la forma tradicional en la que los teóricos trabajaban, entonces puedes producir una teoría que sea inconsistente o incompleta y, lo que es peor, puedes no darte cuenta de ello; puedes, por así decirlo, quedar embaucado por tu propia jerga. Es, por tanto, una buena idea intentar construir teorías que prevengan el que esto ocurra. Ahora bien, podrías utilizar las matemáticas y, ciertamente, estuvo en boga su utilización, como sabes, desde los tiempos de Hull hasta los años sesenta, y todavía continúa en algunos campos, muy correctamente por otra parte, para construir modelos matemáticos. Pero el problema más importante es que las técnicas de las matemáticas, las técnicas clásicas para problemas aplicados, y estoy pensando en el cálculo infinitesimal, etc., no tienen ninguna aplicación, tal como yo lo veo, a cuestiones como el razonamiento. La verdadera pregunta es, entonces, qué tipo de lenguaje teórico podríamos utilizar que fuera útil. Y es aquí donde hay otro factor muy importante. Supongamos que voy al psicoanalista y le digo que sufro una terrible parálisis en un brazo ya que no puedo moverlo, y el psicoanalista me dice «bien, evidentemente tiene usted una parálisis histerica» o aunque no me lo diga eso es lo que piensa, y continúa «y esto ha sido ocasionado porque usted tiene un conflicto edípico»; entonces proba-

blemente yo le diga al psicoanalista «¿qué quiere usted decir con conflicto edípico?» y el dirá «bien, ciertos asuntos causan tanta ansiedad en su mente consciente que usted los reprime en su inconsciente», y yo puedo decir «¿qué quiere decir con “represión” y qué quiere decir con “inconsciente”?»; y así podemos continuar durante un buen rato. Los libros y artículos psicoanalíticos están llenos de una enorme cantidad de jerga semejante, y, en cuanto a esto, lo mismo pasa en los artículos de psicología cognitiva. Y realmente uno siempre se pregunta, cuando se usan este tipo de jergas, si no se está suponiendo más de lo que se debería. Precisamente, la teoría de la computación fue desarrollada contando con este asunto, ya que la gente que lo hizo, y estoy pensando en el trabajo de Gödel y Turing en los años treinta, estaban principalmente preocupados con la siguiente idea. ¿Cuanto tenemos que suponer para ser capaces de tener un procedimiento efectivo para efectuar un cálculo? ¿Cuál es el mínimo que necesitamos suponer? Ellos respondieron esta pregunta en el campo de las matemáticas ya que explicaron detalladamente el mínimo que se necesita suponer para tener un procedimiento efectivo. Así, la teoría computacional estuvo dirigida a la cuestión de realizar el mínimo número de asunciones de forma que se pudieran explicar un amplio rango de posibilidades. Esto es exactamente la mismo que cualquier ciencia necesita hacer. De esta manera, si intentamos construir una teoría en términos computacionales y lo hacemos así, entonces tendremos una explicación manifiesta muy buena de lo que estamos suponiendo.

Para resumir un asunto sobre el que se puede escribir por extenso, si construyes una teoría en términos computacionales estas construyendo un modelo de la misma, un modelo de trabajo que si funciona entonces puedes estar seguro de que no estás suponiendo demasiadas cosas y que es consistente. Ahora bien, aunque todo esto que pienso es perfectamente verdadero y legítimo, sin embargo, quizá no es suficiente para convencer a los escépticos. Pienso que la mejor razón para construir un modelo computacional, y a propósito espero que esto se pondrá de manifiesto en tu último modelo sobre silogismos, es que cuando construyes tales modelos, en la medida en que consigues hacerlos más cercanos a la realidad psicológica, te proporcionan a menudo nuevas ideas sobre la teoría. Déjame que vuelva sobre este punto, ya que incluye una noción bastante sutil. No pienso que un programa computacional sea una teoría. La forma ideal en que se debe proceder en ciencia cognitiva es desarrollar una teoría sobre un fenómeno o un campo particular, como la estereopsis en la visión o el razonamiento silogístico, y expresarla en lenguaje ordinario. Entonces se deben hacer dos cosas completamente distintas. Por una parte debes someter la teoría a una comprobación experimental; de esta manera es la teoría la que estás comprobando cuando haces experimentos. Por otra, desarrollas un programa computacional que es un modelo de la teoría y, en mi opinión, sería realmente raro que lo que tú quisieras fuera comparar directamente lo que la gente hace con la actuación del programa. La razón proviene en parte del hecho de que es extremadamente difícil evaluar la bondad del ajuste entre ambos y, además, el ajuste puede ser falsamente bueno. Puede ser bueno por razones incorrectas. Por tanto, soy en general bastante contrario a la idea de simular la actuación humana mediante programas computacionales. Lo que quiero decir es que la gente que lo intenta tiene la gran dificultad

de cómo evaluar los méritos de la simulación. Pienso que es preferible considerarlos como dos desarrollos separados de la teoría, por una parte la comprobación experimental y, por otra, el programa computacional.

El aspecto verdaderamente interesante en la construcción de un programa computacional es que puedes cambiar la teoría; aunque el hecho de construir modelos computacionales no asegura el que consigas buenas intuiciones teóricas y, de todas formas, es un error intentar construir un programa computacional de toda una teoría completa. Sería un programa enorme y cuando lo terminaras no lo entenderías. Es mucho mejor coger pequeñas partes de la teoría y construir modelos de ellas con el propósito de conseguir ideas nuevas sobre esas partes de la teoría. Así, en mi opinión, aunque la consistencia y la completitud están muy bien, creo que es mucho más como los puntos de vista del pintor americano Jackson Pollock. ¿No sé si conoces a Jackson Pollock el expresionista abstracto?

17. *Sí, creo que sí.*

R. Bien, él dice sobre la pintura que lo que importa no es el producto final, sino el proceso que te lleva a producirlo. En realidad no sé si esto se aplica a toda la pintura, pero estoy convencido que se aplica a la construcción de programas computacionales que modelan teorías. No es tanto el programa final en lo que estás interesado, sino en cómo has llegado hasta allí y en cómo tus ideas sobre la teoría cambian. Es en gran medida un proceso dialéctico; refleja en cierto modo el proceso dialéctico de realizar experimentos para corroborar una teoría. Lo que quiero decir es que los experimentos interesantes son aquellos que nos obligan a cambiar la teoría, no a abandonar toda esperanza en ella, sino a cambiarla, a modularla y mejorarla. Lo mismo pasa con los programas computacionales, y pienso que estas dos diferentes tareas son muy importantes para lograr buenas teorías. Quieres que tus teorías estén conectadas a los hechos y, al mismo tiempo, que sean consistentes; es decir, que cumplan las dos concepciones teóricas sobre la verdad, la teoría de la correspondencia y la teoría de la coherencia. Como dije el subproducto de la búsqueda de la coherencia es el logro de buenas ideas. Hasta puedes conseguir buenas ideas acerca de posibles nuevos experimentos cuando estas mirando cómo funciona el programa, intentado calcular qué es lo que pasa cuando existe algún error en el mismo.

18. *Centrándonos en las tendencias recientes en ciencia cognitiva, en «Mental Models» ya defendías que los procesos mentales ocurren en paralelo. Esta idea, aunque no nueva, está siendo crecientemente admitida. ¿Por qué necesitamos postular procesos en paralelo?*

R. Pienso que existen una gran cantidad de argumentos. Quizá lo más sencillo sea considerar lo que sucede cuando la gente habla, cuando conversa como lo estamos haciendo nosotros. Sabemos que para producir una oración razonable tienes que tener una idea, yo diría un modelo de la situación, y tienes que formularla. Sabemos que tienes que proyectar tal idea en palabras y en una sintaxis apropiada. Cuando has conseguido una representación de la cadena fonológica correspondiente a las palabras en unas posiciones sintácticas apropiadas, entonces tienes que proyectarla en un con-

junto de instrucciones para mover el aparato articulatorio. Cuando mueves los articuladores sabemos que suceden varias cosas en paralelo. Una de las razones por la que es tan difícil construir un programa computacional para comprender el lenguaje es porque si consideramos el habla desde el punto de vista ondulatorio, entonces en ningún momento el habla contiene información sobre sólo un sonido. Lo que quiero decir es que es muy fácil pensar que los sonidos del habla, cuando pronuncias una palabra, están encadenados juntos como las cuentas de un collar o como las letras en el alfabeto, pero de hecho esto es completamente erróneo. Los sonidos del habla ocurren simultáneamente, son más como diferentes trenes corriendo uno al lado de otro en carriles separados, de forma que en un momento determinado no sólo necesitas controlar el que tu aparato vocal haga varias cosas simultáneamente, la laringe puede tener que vibrar, la lengua que moverse a una posición determinada, todas ellas en paralelo; sino que retrocediendo en la línea de producción se están escogiendo palabras, colocándolas en una correcta disposición sintáctica, etc. Ahora bien, no somos conscientes de casi nada de esto, normalmente somos conscientes sólo de que tenemos una idea y de las palabras que acabamos de expresar. Si tuviéramos que pensar deliberadamente qué estábamos diciendo en cada paso del proceso, no podríamos hacer una cosa semejante en una forma tan rápida. Un contraste bastante instructivo al respecto es, por ejemplo, la diferente velocidad con la que podemos hacer distintas tareas. Tomemos a un músico especialmente hábil, un músico moderno de jazz como Dizzy Gillespie que toca la trompeta. Contando con que un instrumento musical obliga a producir una nota tras otra, Dizzy Gillespie es capaz de improvisar un solo a una velocidad de doce o trece notas por segundo. Aunque nos sorprenda como muy rápido, no es tan rápido como la producción espontánea del habla en la que se pueden producir los sonidos del habla, reconociendo que se solapan, a una velocidad de veinticinco por segundo. Y la razón por la que podemos hacerlo mucho más rápido que un músico tocando un instrumento, es porque el aparato vocal nos permite hacer algunas cosas simultáneamente. Así, el procesamiento en paralelo tiene forzosamente que estar implicado cuando hacemos cosas en paralelo y nos capacita para hacerlas de esa manera. Existen, por supuesto, muchas otras ventajas.

19. *Las teorías conexionistas sobre sistemas distribuidos de memoria están basadas en la computación en paralelo. En tu último libro Cognitive Science has escrito que estas teorías suponen una revolución. ¿Podrías aclarar este punto?*

R. Pienso que mucha gente las consideran así; quiero decir que seguramente no soy el único que considera que son revolucionarias. En mi opinión son realmente revolucionarias quizá por razones que algunos teóricos podrían pensar que no lo son tanto. Para explicar esto déjame explicar primero cómo lo verían estos teóricos. Ellos dirían que sería maravilloso tener una teoría de los procesos cognitivos que fuera un poco más cercana a cómo funciona el cerebro y cómo están conectadas las células del mismo. De esta manera, proponen mecanismos que implican, como tú señalabas, procesamiento en paralelo entre un gran número de procesos muy simples.

Aunque existen diferentes versiones sobre lo que estos procesos deberían hacer, básicamente todos reciben una cierta cantidad de *input* y modulan realmente tal *input*; es decir, realizan algo que transmite desde el simple procesador cierta cantidad de activación. Puede ser una activación positiva, o bien una inhibición, o bien que la célula no haga nada, pero básicamente todo lo que la célula hace es acumular activación, modularla y alimentar algún nivel de actividad. Lo que estos investigadores han mostrado es que cuando se construyen sistemas de este tipo se puede dar una explicación bastante buena de fenómenos que de otra manera podrían ser considerados como un poquito enigmáticos.

Una de las cosas que pasan, o una de las cosas que tú puedes efectuar en sistemas de este tipo es que ya no tienes un sistema que trata con símbolos explícitos y con reglas explícitas que los manipulan. Por ejemplo, cuando estás pensando en escribir un código en Lisp, estás pensando en un símbolo explícito, una variable que tendrá un particular valor en un momento determinado; y además escribes reglas que básicamente, implican condicionales, etc. Lo que los conexionistas, o la gente del procesamiento distribuido en paralelo, han mostrado es que se puede idear un sistema que no posea símbolos explícitos y que no haga uso de reglas explícitas; más bien, lo que este sistema tiene, como ya dije, son estos procesadores que modulan la actividad y para, por ejemplo, asociar un patrón de *input* con otro lo que se puede hacer es preparar el sistema para establecer las necesarias fuerzas de conexión entre cada par de procesadores que están conectados, de forma que dándole una gran cantidad de ejemplos de los tipos de cosas que se quiere aparear, construya una representación distribuida. Es un poco como un holograma; es decir, cuando miras un holograma, el verdadero negativo, cada parte del mismo representa la escena completa. Lo mismo pasa con estos sistemas de procesamiento distribuido. Una asociación concreta entre ciertos patrones de *input* y ciertos patrones de *output* no está representada por un único procesador, sino que está literalmente distribuido entre todos ellos; inversamente cada procesador participa en todas las asociaciones que el sistema es capaz de mediar. De esta manera, la clase de cosas más interesantes que este sistema puede hacer es, ante todo, el que puede aprender usando principios muy simples. Además, es capaz de explicar fácilmente fenómenos como los aspectos reconstructivos de la memoria. Así, si te doy algunas pistas; por ejemplo, te digo que estoy pensando en un actor, que es un hombre muy inteligente, que es también un político muy importante, que lidera una de las más poderosas civilizaciones del mundo, entonces puedo ver que te has dado cuenta de que estaba pensando en Reagan. Toma nota de que deliberadamente te he dado una pista falsa y, sin embargo, has sido capaz de juntar estas pistas muy rápidamente y llegar hasta el nombre de Reagan. Este tipo de cosas, especialmente el que yo te diera una información errónea, son extremadamente difíciles de explicar si se opera con símbolos formales, pero una representación distribuida en paralelo las explica fácil y sencillamente.

Entonces, ¿por qué piensa la gente que esta teoría es revolucionaria? En mi opinión, algunos de los teóricos del procesamiento distribuido en paralelo consideran que es revolucionaria porque han conseguido una teoría que maneja los problemas de aprendizaje mediante redes, maneja bien los aspectos reconstructivos de la memoria, y es un poco más cercana a lo que

sabemos sobre cómo funcionan las células del cerebro. En realidad, permanece tan alejada del funcionamiento del cerebro, como lo hace un algoritmo concreto respecto a lo que sabemos acerca de las señales cerebrales.

¿Por qué pienso yo que es revolucionaria? Bien, aunque no necesariamente sea una revolución completamente beneficiosa, la razón es porque nos hace darnos cuenta en una forma muy llamativa del contraste entre la manera en que tendemos conscientemente a pensar en las cosas, y cómo van a ser representadas mediante el funcionamiento de todos esos procesos paralelos, de los que somos totalmente inconscientes. Cuando pensamos en las cosas conscientemente tendemos a hacerlo en términos de símbolos y reglas explícitas, y en el pasado cuando se elaboraron teorías acerca de los procesos inconscientes se producían, en cierto sentido, simplemente el mismo tipo de teorías que, se decía, funcionaban inconscientemente. De esta manera, se produjeron modelos que implicaban símbolos explícitos y manipulaciones explícitas de símbolos que pueden ser encontrados, tanto en la gente que estaba interesada en el inconsciente en el sentido de Freud, como entre la gente que estaba interesada en la noción de Helmholtz sobre el inconsciente, por ejemplo, los procesos inconscientes que subyacen a la inferencia perceptiva, etc. Esta tendencia a construir teorías que implican símbolos explícitos proviene simplemente de que es la forma en que pensamos y, normalmente, pensamos conscientemente en cómo elaborar, construir tales teorías. Puede estar muy bien, puede ser absolutamente correcto, y así lo he defendido yo, el pensar que la forma en la que funciona la conciencia es en términos de reglas y símbolos explícitos. Por ejemplo, si estás revisando una oración en español para asegurarte de que el verbo concuerda con el sujeto, entonces estás pensando en una regla completamente explícita, estás pensando en forma completamente explícita. No todas las cosas que suceden son explícitas, pero en el nivel alto sí lo son. Estos procesos explícitos dependen de otros procesos que son inconscientes y que, ciertamente, pueden estar mediados en forma distribuida. Sin embargo, cuando la gente piensa en el habla, en los sonidos del habla, lo hace como si cada sonido fuera siguiendo uno detrás de otro, y esto se ha hecho así hasta que en los años cincuenta y sesenta Al Liberman mostró que era un error; la causa de esto es que cuando escribimos una palabra, escribimos una letra detrás de otra. ¿Por qué escribimos una letra detrás de otra? Porque es la forma natural de operar con símbolos. Cuando pensamos en los conceptos tendemos a considerarlos como poseyendo condiciones definidas y precisas. La razón es que cuando intentamos pensar explícitamente en los conceptos definiéndolos como si fuéramos abogados o científicos o matemáticos, tendemos a otorgarles las condiciones necesarias y suficientes para una definición precisa. Esta forma de pensar es la manera natural en la que pensamos conscientemente acerca de las cosas, y lo que la gente del procesamiento distribuido en paralelo ha mostrado es que puede ser un error adoptar esta clase de teorías cuando consideramos los procesos inconscientes, porque estas teorías son muy diferentes. De modo que para mí podría ocurrir que los actuales algoritmos de aprendizaje fueran todos erróneos, o aún que los procesos reconstructivos sean erróneos en los detalles o que, y de esto estoy seguro, las unidades de procesamiento sean completamente diferentes de las que usa el cerebro. Al fin y al cabo es esta

cuestión de obligarnos a reconsiderar la forma en que forjamos nuestras teorías sobre los procesos inconscientes, lo que creo que es la verdadera revolución. La parte negativa de esto es que, ciertamente, es muy difícil pensar conscientemente en los tipos de representaciones que estas teorías utilizan. Lo que quiero decir es que cuando teorizamos acerca de estos procesos, podemos no ser conscientes de que podemos necesitar una endemoniada cantidad de trabajo para comprender nuestras teorías, y que lo mejor que podemos hacer es simplemente elaborar algún programa de computador que arme la representación distribuida en paralelo y que uno pueda decir, bien, así es como yo pienso que se hace. Esto va a ser un poco aburrido, quiero decir para nosotros los teóricos.

20. *Volviendo a la metáfora computacional, con frecuencia se oye decir entre la gente y aun entre algunos psicólogos, que el problema de la utilización de modelos computacionales es explicar las conductas y procesos psicológicos superiores, como la creatividad o la capacidad de elegir libremente, la libertad. En otras palabras, la metáfora computacional serviría sólo para las conductas determinísticas. Sin embargo, tú has estudiado estos procesos desde el punto de vista computacional. ¿Podrías resumir tu opinión acerca de estos temas?*

R. Desgraciadamente aquí no he seguido mis propias recetas. No he seguido mi consejo y, de hecho, no he construido ningún modelo explícito, de modo que mis concepciones al respecto tienen un carácter mucho más especulativo. En mi opinión la mayor parte de la gente valora la idea de tener libre albedrío y, por supuesto, muchos filósofos han argumentado que tenemos libre albedrío y, por tanto, algo como la ciencia cognitiva sería imposible. Pienso que éste es exactamente el argumento que utilizó Descartes. Muchos psicólogos han respondido a tal argumento diciendo, simplemente, que el libre albedrío es una ilusión y que no poseemos semejante cosa; supongo que Skinner es la persona que en forma más evidente ha realizado esta argumentación. Lo que yo he intentado mostrar es que realmente poseemos libre albedrío, en qué podría consistir y cómo puede ser compatible con la consideración del cerebro como un dispositivo computacional.

¿Qué es lo que realmente nos da la posibilidad de tener libre albedrío? Bien, el argumento es un poco complicado, pero intentaré explicarlo. Nosotros podemos reflexionar sobre cómo vamos a tomar una decisión. Muy a menudo en la vida real no reflexionamos en absoluto, simplemente elegimos una cosa u otra como, por ejemplo, continuar hablando contigo en vez de salir del despacho y tomarme una taza de café. El que continúe hablando contigo no es algo sobre lo que haya tomado una decisión, y aun cuando termine la entrevista y me levante y vaya a tomarme una taza de té o cualquier otra cosa, yo podría de hecho hacerlo sin pensar conscientemente «¿qué debo hacer a continuación?, sé que voy a ir y tomarme una taza de té». Puedo hacerlo sin más. Pero existen ocasiones en que realmente nos paramos y pensamos y nos decimos a nosotros mismos «¿qué voy a hacer a continuación?». Lo que es más importante, si nos enfrentamos con una decisión muy difícil no sólo pensamos en lo que vamos a hacer a

continuación, sino que pensamos en cómo vamos a tomar la decisión sobre qué hacer a continuación; y tenemos a nuestra disposición muchas diferentes maneras de tomar la decisión. De modo que por una parte podemos intentar tomar decisiones que sean prudentes y racionales, podemos calcular las probabilidades de los diferentes resultados que pueden suceder, podemos calcular también el valor que, dada esa decisión, tienen para nosotros todos los diferentes resultados, y así sucesivamente; y entonces podemos intentar o bien maximizar el valor subjetivo esperado, o bien hacer alguna otra cosa para tomar tal decisión. Esta sería una manera de decidir hacer algo racional. Otra cosa que podríamos hacer sería decir «bien, no quiero tomar una decisión racional, voy a elegir al azar» y tomar una decisión arbitraria, que nosotros consideremos que es arbitraria. Ahora bien, sabemos como psicólogos que la gente no es realmente muy buena tomando decisiones genuinamente al azar, aunque, sin embargo, tomen decisiones que para ellos son arbitrarias. Si queremos estar realmente seguros de que tomamos una decisión arbitraria, entonces podemos tirar un dado o echar a cara o cruz una moneda. Hay que resaltar el que el hecho de que podamos pensar en cómo tomar una decisión, nos permite proponer formas de tomar decisiones que nos liberan de cualquier nicho ecológico concreto, de modo que podemos hacer cosas que, desde el punto de vista de nuestra supervivencia, pueden ser consideradas como locuras. Podemos decidir arriesgar nuestras vidas, podemos decidir el plantearnos si queremos jugar a la ruleta rusa o no; y para tomar tal decisión podemos tirar una moneda al aire, si sale cara jugaremos a la ruleta rusa. Por tanto y, hasta donde llega mi conocimiento, al contrario que otras especies, como resultado de nuestra capacidad de reflexionar sobre cómo tomar una decisión podemos librarnos de las presiones evolutivas de la selección natural, podemos arriesgar nuestras vidas. Por cierto, esto es algo que escritores como Dostoyevski consideraban como el mayor bien que los seres humanos poseen, algo que siempre es dejado de lado por los científicos que intentan comprender la mente humana, el que podamos comportarnos a nuestro antojo, por capricho. La razón de esto, me gustaría resaltar, es porque podemos reflexionar sobre cómo tomar una decisión. Ahora bien, no sólo podemos reflexionar sobre cómo tomaremos una decisión, sino que podemos subir un nivel más, podemos reflexionar sobre qué método utilizaremos para tomar una decisión. Así, supongamos que tienes que tomar una decisión difícil, tienes que decidir si quieres permanecer en tu actual trabajo o prefieres cambiar. Entonces te dices a ti mismo. ¿cómo voy a tomar esta decisión?, y puedes muy bien decidir realizar una lista de todas las ventajas y desventajas o puedes tirar una moneda al aire. Pero antes de hacer cualquiera de las dos cosas puedes preguntarte cómo vas a elegir el método de tomar la decisión; y puedes decir «tengo cuatro diferentes maneras de tomar una decisión: hacer una tabla de decisión, lanzar una moneda, pedir consejo a mi mujer, pedir consejo a los colegas. Son cuatro maneras diferentes, ¿cómo voy a escoger entre ellas?». Puedes reflexionar al respecto y decir «bien la ventaja de elegir uno de los métodos racionales es que es improbable el que lamente haber decidido mi futuro mediante un método semejante». Entonces, cuando reflexionas sobre cómo vas a escoger entre métodos de elección, de hecho podrías reflexionar sobre tal decisión.

21. *Una especie de cadena de niveles de decisiones y metadecisiones.*

R. Exactamente, y de hecho no tiene fin la construcción de esta torre de posibilidades. Ahora bien, por supuesto en el nivel superior se toma siempre una decisión pasiva, ya que si piensas en tal decisión conscientemente entonces ya no es la decisión de nivel superior; la decisión de nivel superior es aquella en la que estás pensando conscientemente. En cualquier caso, esto es la primera mitad de la historia y una porción bastante abstracta y difícil de algo semejante al razonamiento filosófico. En esencia, los humanos poseemos libre albedrío no porque podamos hacer cosas al azar o por capricho, sino porque podemos reflexionar sobre los métodos de elección y porque podemos aún reflexionar sobre cómo escogeremos entre ellos; y si decidimos elegir entre ellos utilizando un método A, podríamos haber reflexionado sobre los métodos de elección que nos permitieron seleccionar el método A. Y así sucesivamente, «ad infinitum». Ahora bien, ¿qué es lo que esto significa desde el punto de vista computacional? Lo más importante que significa es que tenemos que tener la habilidad de construir modelos de nuestra propia actuación. Podemos considerar la construcción de modelos de nuestra propia actuación desde el punto de vista de la solución de problemas. Supongamos que te doy una serie de problemas para resolver, y tú los vas resolviendo uno a uno hasta que llegas a uno bastante difícil. Entonces puede que te pares a pensar y te digas a tí mismo «¿qué he hecho cuando he resuelto los problemas anteriores?». Construyes un modelo de tu actuación y, quizá, como resultado de algún «insight» que alcanzas por ello, vuelves al nivel inferior de la resolución del problema y lo resuelves. De esta manera, construyes únicamente un modelo de nivel uno. En realidad puedes construir un modelo con un nivel superior adicional, ya que cuando has resuelto un grupo de problemas, te has atascado con uno y te has preguntado «¿qué he hecho cuando he resuelto los problemas anteriores?», puedes responderte a ti mismo «espera un momento, este es uno de los problemas que resuelvo construyendo un modelo de mi actuación, y cuando lo hago tiendo a concentrarme demasiado en los problemas fáciles». Ahora has construido un modelo de tí mismo construyendo un modelo de ti mismo al resolver problemas; es decir, el mismo tipo de jerarquía al que nos referíamos antes. Por tanto, poseemos esta habilidad de construir modelos de nuestra actuación e incrustarlos dentro de nuestro nivel actual de operación; esto supone una habilidad recursiva ya que al construir un modelo de nuestra actuación e incrustarlo dentro de lo que estamos haciendo actualmente, entonces podemos construir un modelo de todo ello y así sucesivamente, creando un nuevo nivel en la jerarquía. De este modo, pienso que el mecanismo que necesitamos para mostrar lo que el libre albedrío parece ser es muy similar al mecanismo que necesitamos para explicar la llamada metacognición.

22. *¿Podríamos explicar la autoconciencia de la misma manera?*

R. Sí, en mi opinión el mismo mecanismo que utilizamos para explicar la autorreflexión en general y el hecho de que podamos reflexionar sobre nuestras autorreflexiones es también el mecanismo que nos sirve para explicar la autoconciencia. En este momento en que estoy hablando contigo,

normalmente no soy consciente de mí mismo, estoy concentrado únicamente en lo que estoy diciendo. Pero puedo parar y volverme consciente del hecho de que soy yo, Phil Johnson-Laird hablando contigo Juan Ma-druga y, en cierto sentido, distanciarme un poquito de lo que está pasando. Acabo de construir algo parecido a un modelo de mi actual actuación y, de hecho, no creo que pueda hacerlo muy bien, pero existe alguna gente, por ejemplo, ciertos místicos, que sostienen esta misma idea, el volverse conscientes de sí mismos siendo autoconscientes. En resumidas cuentas, si tenemos acceso a modelos de nuestra propia actuación, ya sea pasada o actual, y podemos incrustarlos dentro de lo que estamos haciendo actualmente, entonces poseemos algo que es casi una habilidad recursiva que, pienso es lo que subyace cuando tenemos intención de hacer algo o cuando ejercitamos nuestro libre albedrío, nuestra autorreflexión o autoconciencia. Como ves por algo decía que era una historia complicada.

23. Una de las críticas más usuales contra la psicología cognitiva ha sido la de no considerar suficientemente los aspectos emocionales de la conducta humana. ¿Qué piensas sobre el papel de la psicología cognitiva en el estudio de las emociones? ¿Cuál es tu opinión acerca de la relación entre emoción y cognición?

R. Antes de nada pienso que emociones y cogniciones son distintas. Pero pienso también que las emociones pueden tener un efecto causal sobre cómo pensamos y cómo pensamos puede tener un efecto causal sobre nuestras emociones. Por tanto, pienso que son entidades mentales diferentes aunque interactúen causalmente y antes de ir más allá y decirte cómo en mi opinión se encajan las emociones, me gustaría resaltar primero que creo que son muy importantes en la vida cognitiva y que existe un tipo de comunicación en ambos sentidos entre ellas y lo que podríamos llamar cogniciones. Keith Oatley, Tony Anderson, Valentina d'Urso y yo hemos desarrollado una teoría computacional sobre las emociones que esencialmente asume el tipo de arquitectura paralela de la que hemos hablado previamente pero que, al contrario de las teorías conexionistas, asume que lo que se transmite en ciertos casos de un procesador a otro posee un contenido simbólico explícito. Las emociones constituyen un canal completamente separado de comunicación que existe primariamente para rellenar el siguiente hueco. Por una parte, los organismos muy simples viven sus vidas casi enteramente en términos de patrones innatos de conducta y, asimismo, nosotros poseemos también patrones innatos de conducta, reflejos como el parpebral, etc. Estas conductas son muy fáciles de computar, el sistema nervioso sencillamente realiza una simple computación y automáticamente las elicit, pero estas conductas son muy poco flexibles. En el otro extremo, los seres humanos podemos realizar complicadas computaciones inferenciales para hacer frente a problemas muy complicados. Los problemas, por ejemplo, que surgen cuando estás negociando con algún colega del trabajo sobre un asunto social delicado. Ahora bien, la dificultad con este tipo de cálculos es que, de hecho, consumen mucho tiempo y que si tu vida dependiera de ellos posiblemente no tendrían tiempo para finalizar. En esta línea, Keith Oatley y yo hemos propuesto que la selección natural ha generado un compromiso entre ambos tipos de conductas, por el que ciertas

clases importantes de acontecimientos predisponen a ciertos tipos de conductas que no son ni tan rígidas como los reflejos, ni consumen tanto tiempo en su realización como las computaciones inferenciales complejas. De esta manera, lo que señala la conexión entre el tipo de acontecimientos y las, por así decirlo, restricciones en el tipo de cosas que probablemente vas a hacer, son las señales emocionales. Estas señales emocionales permiten que cuando te sientes feliz te sientas como inclinado a hacer ciertas cosas, o que cuando te sientes triste te sientas como encerrado en ti mismo, sin hacer nada, o cuando estás enfadado te sientas como agresivo, etc. Por ejemplo, cuando estás enfadado es bastante difícil hacer cosas constructivas, agradables y placenteras; por el contrario, la naturaleza te encamina en la realización de ciertas clases de cosas que si las consideramos desde el punto de vista de la evolución de las emociones podemos encontrarlas en los ruidos modos sociales de otras especies como los chimpancés. Por tanto, las emociones, desde un punto de vista evolutivo, son más antiguas que nuestras computaciones inferenciales complejas, son las señales internas que nos predisponen a realizar ciertos tipos de conductas.

Otro aspecto de singular interés es el que las emociones sean comunicadas mediante clases de señales muy semejantes en diferentes especies; es decir, cuando comunicamos nuestras emociones a otros miembros de nuestra especie, lo hacemos sin usar estructuras simbólicas complejas en las que tenemos que extraer porciones de significado de diferentes partes y combinarlas; no son como las oraciones del lenguaje, no son inferenciales, computacionalmente complejas. Los mecanismos que subyacen son de tal tipo que podemos crear emociones en otra gente mediante simples vocalizaciones o gritos. Una vez más, pienso que hay una manera de incluir las emociones dentro del marco general cognitivo o computacional y, ciertamente, existe en la actualidad una cantidad creciente de trabajo realizado por científicos cognitivos que están interesados en las emociones. Aunque la idea de intentar integrar la cognición con la emoción es, en cierto sentido, bastante antigua, sin embargo, yo espero que la realización o el intento de desarrollo de modelos computacionales nos capacitará para hacer cosas nuevas o para tener algunas ideas nuevas al respecto.

24. Quizá la más importante crítica contra la psicología cognitiva haya sido la de la tendencia que ésta tiene a las miniteorías rebuscadas. ¿Qué piensas sobre la proliferación de miniteorías? ¿Cuál es en tu opinión sobre el estado actual de salud de la psicología cognitiva?

R. Concretamente pienso que eso es verdad; es precisamente el punto que sostiene Allan Newell en su artículo «You can't play 20 questions with nature and win», sobre la tendencia existente entre los psicólogos cognitivos a concentrarse en un determinado paradigma experimental para realizar interminables experimentos dentro del citado paradigma, y avanzar teorías sólo sobre lo que sucede en esos experimentos. Mi opinión es que quizá uno de los subproductos del desarrollo de los modelos computacionales es que la gente haya comenzado a preocuparse de cómo podría trabajar el sistema. La gente empieza a preocuparse de lo que podría llamarse la arquitectura cognitiva y pienso que es una decisión muy sensata. Creo que probablemente es un error para los psicólogos el quedar enganchados en un

determinado paradigma experimental, yo me confieso culpable de haberlo hecho, aunque sólo fuera porque tal paradigma podría no tener demasiada validez ecológica. Los ejemplos típicos de esto para mí serían conseguir que la gente aprenda listas de sílabas sin sentido o conceptos que son definidos únicamente en función de un número de atributos. El aprendizaje de sílabas sin sentido resulta estar demasiado alejado del aprendizaje ordinario que depende decisivamente del significado de lo que uno está aprendiendo y el aprendizaje de conceptos que están definidos puramente en términos de conjunción o disyunción de atributos resulta ser o demasiado relevante acerca de cómo nos enfrentamos a los conceptos, que no hay muchos conceptos de ese tipo en la vida normal. Lo que quiero decir es que una mesa no es sólo la conjunción de una parte superior y unas patas, el aspecto decisivo es que las patas tienen que soportar la mesa; hay aquí una relación y entonces necesitas algo semejante a un lenguaje para describir conceptos que van más allá del cálculo proposicional que sólo permite conexiones.

En cuanto al estado de salud, pienso que es muy difícil decirlo, aunque ciertamente creo que las cosas van mejor ahora que cuando escribí mi artículo sobre psicolingüística. Pienso que estamos en una fase muy excitante; yo diría que en los últimos diez años, más o menos, las cosas han tomado un camino que si a finales de siglo, y puedes volverme a preguntar entonces si todavía estoy por aquí, no han sucedido toda una serie de cosas comparables a, por ejemplo, el trabajo de David Marr sobre la visión, entonces estaré muy desilusionado e inclinado a pensar que quizá es lógicamente posible, pero que, simplemente, sería demasiado esperar que un organismo sea capaz de comprender su propio funcionamiento interno.

25. *Muchas gracias.*

Bibliografía Básica (*)

- JOHNSON-LAIRD, P. N. (1977). Psycholinguistics without linguistics. En N. S. Sutherland (ed.). *Tutorial Essays in Psychology*, vol. 1, Hillsdale, N. Jersey: Erlbaum.
- JOHNSON-LAIRD, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.
- JOHNSON-LAIRD, P. N. (1987). A taxonomy of thinking. En R. J. Sternberg y E. E. Smith (eds.). *The Psychology of Human Thinking*. Cambridge: Cambridge University Press.
- JOHNSON-LAIRD, P. N. (1987). The development of reasoning. En G. Butterworth y P. Bryant (eds.). *Causes of Development: Interdisciplinary Perspectives*. Hassocks, Sussex: Harvester Press.
- JOHNSON-LAIRD, P. N. (1988). *The Computer and the Mind*. Cambridge, Mass.: Harvard University Press.
- JOHNSON-LAIRD, P. N., y WASON, P. C. (1977). (eds.). *Thinking: Readings in Cognitive Science*. Cambridge: Cambridge University Press.
- MILLER, G. A., y JOHNSON-LAIRD, P. N. (1976). *Language and Perception*. Cambridge: Cambridge University Press.
- NEWELL, A. (1973). You can't play 20 questions and win. En W. G. Chase (ed.). *Visual Information Processing*. N. York: Academic Press.
- OATLEY, K., y JOHNSON-LAIRD, P. N. (1987). Towards a Cognitive Theory of Emotions. *Cognition and Emotion*. 1(1), 29-50.
- WASON, P. C., y JOHNSON-LAIRD, P. N. (1972). *Psychology of Reasoning: Structure and Content*. Londres: Batsford. Versión española en editorial Debate (1980).

(*) En esta bibliografía hemos incluido únicamente aquellas publicaciones que se citan en la entrevista y algunas de las más recientes del profesor Johnson-Laird, relacionadas directamente con los temas tratados.