KI - Künstliche Intelligenz (2019) 33:301–305
https://doi.org/10.1007/s13218-019-00607-z

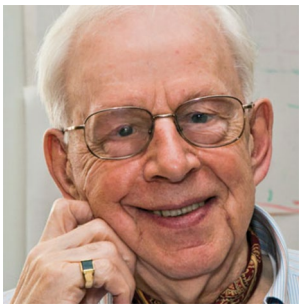
INTERVIEW

Thinking is Founded on Models of Possibilities

Interview with Phil Johnson-Laird, Princeton University/New York University

Marco Ragni[1]

Published online: 1 July 2019



Phil Johnson-Laird is emeritus Stuart professor of psychology at Princeton University and a visiting scholar at New York University. He is member of the National Academy of Sciences, and a fellow of the Royal Society. He developed the mental model theory with many colleagues. His publications include 14 books and over 300 paper about linguistics, thinking, and music. Marco Ragni (MR) spoke with him about Cognitive Reasoning on April, 2019.

MR: What makes human reasoning so special? It is neither logical nor illogical but demonstrates systematic patterns, how would you characterize it?

Human reasoning is special—at least in comparison with systems in AI—because it often combines deduction, induction, and abduction. For example, the first time I entered a coffee bar in Italy, I went up to the crush at the bar. When I got the attention of one of the two baristas, I said in awful Italian: "un cappuccino per favore". He gestured at me to go away. I tried again: he made the same gesture, and turned to the guy next to me to serve him. My accent was poor, I was obviously not an Italian, but Italians are not known to be xenophobic. So, I was doing something wrong (1: an abduction). What was it? The customer next to me handed over a slip of paper, and the barista gave it a small tear. Indeed, other customers were waving these pieces of paper at the barmen. So, perhaps all the people who wanted to be served had these pieces of paper (2: an induction). What were they? They were receipts for payments (3: an abduction). Yet, no-one was paying either of the two baristas. So, the customers were paying someone else (4: a deduction). I looked round the cafe, and there in the direction in which the barista had gestured was a woman behind a cash register, and people waiting in line to pay her. So, to get a coffee, you first pay the cashier, take the receipt to the bar and give it to a barista, who tears the slip to prevent its reuse, and makes your coffee (5: an abduction using the previous conclusions to simulate the sequence of events). It was unlikely that this bar was unique in its system of payment. If it had been then other customers would have been as bemused as I had been. Ergo, the bar's system was likely to be general at least in Padua, perhaps over all Italy (6: abduction). My working definitions are: for 'deduction,' inferences that do not increase the semantic information[1] in the premises, including those that knowledge provides; for 'induction,' inferences that increase semantic information but without introducing any concepts that are not already in the premises; and for 'abduction,' inferences that introduce new concepts from knowledge into explanatory conclusions.

My inferences took much less time to make—a few seconds—than to describe. But they illustrate a common feature of human reasoning. We leap around from inductions to deductions to abductions, and back again. The same pattern occurred in the thinking of John Snow as he figured out how cholera was communicated from one person to another, in the Wright brothers' thinking as they invented a controllable

✉ Marco Ragni
ragni@informatik.uni-freiburg.de

[1] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Georges-Köhler-Allee 52, 79110 Freiburg, Germany

[1] For the concept of semantic information, see Bar-Hillel & Carnap (1952).

heavy-than-air air craft, and in Gordon Welchman's thinking as a worked out how to break the Enigma machine's coding system at the outset of World War II.[2]

The processes of reasoning are seamless. Which makes it difficult to find out how they work. Yet, psychologists can isolate different sorts of reasoning in experiments. My colleagues have shown how abduction and deduction work together when children, and those who know nothing of programming, create informal algorithms for permuting the order of cars in a train (using a single siding in a track with the power of a Universal Turing Machine). And, here, is another sort of illustrative inference for you to try:

> If a pilot falls from a plane without a parachute then the pilot dies.
> This pilot didn't die, however.
> Why not?

People in both the West and East Asia divide up into two sorts depending on the inferences they make. Some abduce explanations, e.g.:

> The pilot fell into a deep snowdrift.
> He (sic) fell into a large bouncy cushion.
> The plane was on the ground so the pilot didn't fall far.
> The pilot was already dead.

Others make a straightforward deduction:

> This pilot did not fall from a plane without a parachute.

The two groups differ in personality. If you're open to experience and not highly conscientious—two well-known traits that can be assessed, you tend to make an abduction, but if you have their mirror-image traits, you tend to make the deduction.

MR: Do you think formal logic (or any related formal system) is an adequate method for modelling cognitive reasoning?

One of my mentors, the late A.R. Jonckheere ('Jonck'), a brilliant psychologist and statistician, spent every summer in Jean Piaget's institute in Geneva. Early in my graduate career, he told me that the task for students of reasoning was to figure out which system of logic is in the head, and how it is formalized there. He was giving a succinct summary of how psychologists thought about reasoning circa 1964. It led to various of us, notably Lance Rips, Daniel Osherson, and Martin Braine, to follow Piaget's example and to propose theories of reasoning based on formal rules of inference. Formal logic is, of course, one of the supreme achievements of human reasoning, as is Gödel's proof of the incompleteness of the logic needed for arithmetic, and Turing's foundation for the theory of computability. So, our efforts to harness logic as the basis for reasoning were almost inevitable, notwithstanding Frege's argument against 'psychologism'. Contrariwise, it has been a hard slog to realize that logical form can be identified in everyday inferences only by taking meaning and knowledge into account. But, once you do that, there's no need for logical form, because people can reason from representations of meaning and knowledge—representations that we think of as mental models of the world. A crucial empirical discovery depended on pitting the length of formal proofs against the number of mental models: it was the latter that predicted the difficulty of inferences. Likewise, what ought to be difficult according to formal logic can be easy for human reasoners, and vice versa. Children can make inferences of this sort, for instance:

> More than half the people in the room speak English.
> More than half the people in the room speak German.
> So, at least one person in the room speaks both languages.

Yet, the treatment of 'more than half' in logic calls for the second-order predicate calculus. In contrast, an inference, such as:

> None of the French speakers in the room speaks English.
> All the English speakers in the room speak Italian.
> What follows?

is a tough inference even for adults to make.[3] Yet, it is from a tiny decidable subset of first-order logic corresponding to syllogisms. And, as you have shown, Marco, no normal modal logic seems adequate to explain human reasoning about possibilities.

MR: Can people make deductions if they don't rely on formal rules of inference from some sort of logic?

A logician once told me: 'People who don't know logic cannot make deductions', and when I presented data to the contrary, he accused me of intellectual dishonesty. In fact, a large literature of psychological experiments, starting over 100 years ago, shows that people who know nothing of logic can make valid deductions. Sudoku problems would hardly be popular otherwise. Three robust phenomena are, first, that people differ in their ability to reason, which correlates with their intelligence—not surprising, because many tests of intelligence call for deduction. Indeed, we know more about deduction than about intelligence. Second, inferences differ

[2] These three case histories are in my book, *How we reason*.

[3] The only definite conclusion about the two end terms that follows necessarily is: At least some of the Italian speakers do not speak French.

in their difficulty one from another. Third, many domains of deduction are computationally intractable, including, as you proved, Marco, two-dimensional spatial reasoning. What is crucial in all three of these phenomena is the processing capacity of working memory—the human memory system that holds the results of intermediate computations.

MR: How, then, do people reason?

The way they do so appears to be as follows. They envisage the possibilities to which the premises refer, and seek a parsimonious and new conclusion that holds in these possibilities. So, they construct a finite set of mental models of the possibilities. And insofar as feasible these mental models are 'iconic'—they have the same structure as the situations that they represent, and therefore embody relations that may not have been asserted as such in the premises. They are the source of a conclusion, and people realize that if it holds in all the models of the premises then it follows of necessity, whereas if it holds in only some of them then it follows only as a possibility—reasoners may even, if asked, estimate the probability of a conclusion from the proportion of models in which it holds, assuming their equipossibility. The only real knowledge of logic that untrained individuals seem to have is that an inference for which there is a counterexample does not follow of necessity. But, contrary to logic, they also think that nothing follows from many premises, including those that are contradictory.

This account is a thumbnail sketch of the theory of mental models, which is what my colleagues and I have been working on for some time, and which we have implemented in various computer programs (available at this website: https://mentalmodels.princeton.edu). The theory's most striking prediction, which emerged from one such a program, is the occurrence of systematic fallacies that can be as compelling as illusions. Here's an example:

Suppose you know that the following two exclusive disjunctions are both true:

> Either there's a seat in the balcony or else there's one in the stalls.
> Either there's a seat in the balcony or else there isn't one in the stalls.
> Is it possible that there's a seat in the balcony?

Most people say, 'Yes'. They envisage the possibilities for the first disjunction, which include the possibility of a seat in the balcony, and then they envisage the possibilities for the second disjunction, which also include the same possibility. So, they say, 'Yes'. They overlook what is false in these possibilities. Indeed, people have a bias to represent only what is true. The first possibility of a seat in the balcony co-occurs with the falsity of one in stalls, whereas the second possibility of a seat in the balcony co-occurs with the falsity of there *not* being one in the stalls. So, the two possibilities contradict one another. People balk at drawing any conclusion from a contradiction. So, it doesn't follow that there's a seat in the balcony.

MR: How far are we from a general theory of human reasoning?

The good news is that cognitive scientists are making progress. Here's an egocentric index: my first book on reasoning[4] proposed—with just one exception—no theory of either what cognitive reasoning computed or of how it made the computations. What the book showed in essence was: content matters. The one exception was an algorithmic account of how people select evidence to test general conditional hypotheses, such as: *If a card has an 'A' on one side then it has a '2' on its other side.* The people who carried out this 'selection' task, which Wason had devised, knew that each card had a letter on one side and a number on the other side. They seldom selected a card with a '3' on it. Yet, with an 'A' on its other side, it would have refuted the hypothesis. This oversight was so surprising—and seemed so irrational—that it launched hundreds of experiments. My algorithm—with your help—turned out to give a good account of what was going on. Naive individuals didn't realize in just one 'shot' at the task that it was important to try to falsify the hypothesis. Give them repeated trials with feedback, and they soon understand its importance.

Nowadays, psychologists have algorithmic accounts of sentential reasoning, modal reasoning, reasoning with monadic assertions, and reasoning about probabilities. So, the field has advanced. And the evidence suggests—not without controversy—that people reason by building models of possibilities, which are small finite alternatives (rather than the infinitely many 'possible worlds' in the semantics for modal logic). What is so far missing—though Sangeet Khemlani, you, and others, are making progress—is a single unified algorithm for cognitive reasoning.

MR: From your perspective—what are current limitations of AI approaches to explain human reasoning?

When smart people use their intuitions to develop algorithms, the results can be startling. But, psychologists learn after a few experiments that intuitions about human reasoning are often wrong. So, AI can be both brilliant and irrelevant to cognitive science. Its biggest discovery about cognitive reasoning was the need for systems to be able to advance tentative conclusions and to withdraw them should they turn out to be wrong, i.e., the need for nonmonotonic logic. Yet, in my view, these logics overlook a critical aspect of human reasoning. Here's an illustration. A friend and I were sitting outside a restaurant nearly opposite to Picasso's Château in Provence. Two other friends had gone to get the car, which

[4] The book was The Psychology of Reasoning, which Peter Wason and I published in 1972.

we'd parked elsewhere. And we inferred that they'd be back in about 10 min. After 20 min, there was no sign of them. An AI nonmonotonic logic would allow us to withdraw our conclusion, and to amend our premises. Well, we did withdrew our conclusion, but by far the most important part of our thinking was to come up with a plausible explanation of what had happened to our friends. We needed it in order to decide *our* best course of action, i.e., whether to walk to the car or to stay where we were. We inferred that our friends had had difficulty in starting the car—this problem had happened before. Such abductions are not part of nonmonotonic logics. Ours enabled us to infer that our best course of action was to stay where we were, and to wait. After another 5 min or so, sure enough, the car came spluttering into view—it had needed a tow to get it started.

AI programs for automated theorem-proving often implement first-order logic, and embody unification and resolution. They can be helpful to mathematicians and to programmers, but they are not implementations of cognitive reasoning. One of its facets, as I mentioned earlier, is a susceptibility to making systematic fallacies. If AI researchers ever seek to simulate human reasoning, they cannot base their systems on standard logic, or only on their intuitions about everyday reasoning. Philosophers and linguists have taken to carrying out experiments to test their ideas. Perhaps it's time for artificial intelligencers to do the same.

MR: Nowadays machine learning with neural networks seems to be the dominant technology in artificial intelligence. Do you think it helps to understand or mimic human reasoning? How far?

Indeed, AI's great success story is the development of deep learning, even though "adversarial" attacks show that programs implementing it can be fooled in ways that humans are not. Nonetheless, it triumphs are striking and revolutionary. But they have pushed AI approaches to reasoning into the background. If you test the reasoning ability of Alexa, Siri, and other electronic helpers, they fail miserably. Try asking them "at what time will it be hottest today?" and it is beyond them to say. I posed this simple inference to them:

You're a machine, and all machines are fallible. So, what follows?

Alexa said: Um, I'm not sure.
Siri said: I don't really like these arbitrary categories, Philip.

Yet, 9-year olds can infer correct conclusions of this sort.

MR: For some problems such as the Wason Selection Task there exist more than 16 theories how humans reason—which is devastating for any science. One answer to eliminate theories is to come up with criteria and benchmarks. What do you expect of good cognitive theories?

The late David Marr gave an excellent answer to this question—presaged by Lord Adrian, the Nobel prize-winning neurophysiologist. A good cognitive theory should include, first, an account of what the brain is computing, and, second, an account of the algorithm that it uses to make the computations. A great cognitive theory will also explain the neuronal processes implementing the algorithm. Because there are infinitely many ways in which to carry out any computable function, it makes sense to express the algorithmic theory in the vernacular, but to develop a computer program that executes its processes—to ensure that the theory works. You and I were surprised by how few of the 16 theories of Wason's selection task were expressed as algorithms, and at least one of the theorists has argued that they shouldn't be, because it makes them harder to fit to data. Perhaps. But, the bane of psychology is theories that take too much for granted—from Freud on the dreamwork to Piaget on intellectual development. And an algorithm is an obvious antidote. So, my criteria for cognitive theories are threefold. First, they should describe what the brain is computing and how it does the computations. Second, they should yield novel, even unexpected, predictions. And, third, they should fit the data with as few parameters as possible. So, theories should rely on well-understood processes, they should be surprising, and yet they should match reality.

MR: Is there hope to learn form neuroscientists about how to handle large amount of knowledge and how to use it for reasoning?

Knowledge is critical to human reasoning, and humans accumulate knowledge throughout their life times. Several scientists have tried to estimate how much. But, only Tom Landauer, who died in 2014, took into account how much they are also likely to forget. His estimate was a net gain of 10^9 bits in a lifetime. That's equivalent to the contents of about a hundred books. Depressed by this small number, I once asked him whether were any grounds to raise his estimate. 'Yes, for you,' he said, 'I'll make it 10^{10} bits!' Neuroscientists get big data from brain imaging studies—information from vast numbers of voxels (3D pixels)—and so they have automated the process of analysis in order to discern which regions of the brain underlie particular tasks. A common assumption on the part of logically-oriented students of reasoning is that one can treat knowledge as a set of assertions. But, in what language? Natural language is too ambiguous; the language of 'thought' is terra incognita; and neural activity has yet to be related to the contents of knowledge. So, my assumption is to treat knowledge as represented in models, accessible via their contents, but in ways as yet largely unknown.

MR: Hilbert proposed twenty-three problems that were responsible for great progress in mathematics. Do you see similar challenging problems for human reasoning?

A complete solution to how any sort of cognition works seems to depend on solutions to all of them—cognitive reasoning depends on language, knowledge, inferential processes, consciousness, and recursive theories. So, there are five challenges:

Problem 1: *Devise an algorithm that learns a compositional semantics for a natural language from examples of its utterances and their referents.* A compositional semantics composes the meaning of a sentence from the meanings of its parts according to their syntactic relations, and so on, down to the meanings of its morphemes (cf. Tarski on predicate calculus; and Montague on natural language). No-one knows how humans acquire a compositional semantics for their native tongues.

Problem 2: *Devise an algorithm that represents knowledge from descriptions—to keep inputs simple—and that retrieves from it what is pertinent to given inferences.* For example, the conditional:

If it's raining then it isn't pouring

refers to two exhaustive possibilities:

It is raining and not pouring.
It is not raining.

The reason is that we all know that *pouring* means *raining heavily*, and so it can't be pouring but not raining.

Problem 3: *Devise a single algorithm that unifies the apparently different processes of reasoning needed to solve tractable everyday and scientific problems.* It should be able to solve my problem in the Italian coffee bar, and, say, John Snow's problem of inferring a single mechanism for the transmission of cholera given that the disease can leap great distances.

Problem 4: *Devise an algorithm that accounts for those aspects of cognitive reasoning that humans are conscious of, and those that they are not conscious of.* The aim is to elucidate the role of conscious reflection about inferences in devising a logic.

Problem 5: *Devise an algorithm that can construct theories of cognitive reasoning.* Just as evolution is the primordial recursive process—it applies to its own outputs, so cognitive science is the first recursive discipline. A theory of cognitive reasoning should account for its own origins.

Thanks for the chance to answer your questions, Marco.