THINKING AND REASONING, 2002, *8*(1), 69–95

Peirce, logic diagrams, and the elementary operations of reasoning

P.N. Johnson-Laird

This paper describes Peirce's systems of logic diagrams, focusing on the so-called "existential" graphs, which are equivalent to the first-order predicate calculus. It analyses their implications for the nature of mental representations, particularly mental models with which they have many characteristics in common. The graphs are intended to be iconic, i.e., to have a structure analogous to the structure of what they represent. They have emergent logical consequences and a single graph can capture all the different ways in which a possibility can occur. Mental models share these properties. But, as the graphs show, certain aspects of propositions cannot be represented in an iconic or visualisable way. They include negation, and the representation of possibilities *qua* possibilities, which both require representations that do not depend on a perceptual modality. Peirce took his graphs to reveal the fundamental operations of reasoning, and the paper concludes with an analysis of different hypotheses about these operations.

> Deduction is that mode of reasoning which examines the state of things asserted in the premisses, forms a diagram of that state of things, perceives in the parts of the diagram relations not explicitly mentioned in the premisses, satisfies itself by mental experiments upon the diagram that these relations would always subsist, or at least would do so in a certain proportion of cases, and concludes their necessary, or probable, truth.
>
> *Peirce (1.66[1])*

The American philosopher Charles Sanders Peirce (1839–1914) was a great logician. He discovered the central branch of logic known as the *predicate* calculus independently from Frege. Frege published in 1879; Peirce published in

[1]This paper follows the standard notation for citing papers in Peirce's (1931–58) *Collected works*, e.g., "1.66", refers to Vol. 1, paragraph 66, which is the first paragraph of a paper, "The logic of relatives", first published in 1883. For unpublished works, see www.peirce.org

Address correspondence to P.N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, NJ 08540, USA Email: phil@princeton.edu

This paper has benefited from the helpful suggestions and scholarship of Mike Oaksford, Jon Oberlander, and David Over. Any mistakes are the author's own. The research was supported in part by a grant (Number 0076287) from the National Science Foundation to study strategies in reasoning.

http://www.tandf.co.uk/journals/pp/13546783.html DOI: 10.1080/13546780143000099

1883 (3.328). Peirce also invented truth tables, higher-order predicate calculi, and systems of modal logic. However, his work has often been overlooked because his career in academia was tragically short, and because many of his papers were unpublished or unfinished in his lifetime. He developed an algebraic notation for logic similar to modern systems. He improved Venn's logic diagrams (see 4.357; Shinn, 1994). But he also devised two diagrammatic systems for logic, which were powerful enough to deal with sentential connectives such as "if" and "or", and quantifiers such as "all" and "some". That is, they corresponded to the predicate calculus. They also, he claimed, "[render] literally visible before one's very eyes the operation of thinking *in actu*" (4.6), and "put before us moving pictures of thought" (4.8). His work anticipated the semantic networks of artificial intelligence (Sowa, 1984), discourse representation theory in linguistics[2] (Kamp, 1981; Kamp & Reyle, 1993), and the theory of mental models in psychology (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). It has inspired Sowa's theory of conceptual graphs (Sowa, 1984, 1997), a text book on logic (Ketner, 1991), and the work of Jon Barwise and his students on diagrammatic systems of reasoning (e.g., Hammer, 1995; Shinn, 1999).

Cognitive psychologists, however, are largely ignorant of Peirce's diagrams. The present author knows of no psychological study of reasoning that describes them (*pace*, e.g., Anderson, 1996; Bauer & Johnson-Laird, 1993; Larkin & Simon, 1987; Lindsay, 1988; Schwartz & Fattaleh, 1972; Stenning & Oberlander, 1995); and he himself became aware of them only by chance. The diagram systems, or "logic graphs" as Peirce called them, have a reputation for being obscure and hard to understand. They are not mentioned in the standard histories of logic (e.g., Bochenski, 1961; Kneale & Kneale, 1962; Styazhkin, 1969). And Gardner (1982) in his excellent compendium of diagrammatic systems for logic describes the task of understanding them as formidable, and does not explain them. Yet they make possible proofs of all and only the valid theorems of the predicate calculus, and they are no more difficult to understand than the calculus itself. And, as Peirce himself realised, they are highly pertinent to the operations of the mind. Most other diagrammatic systems, such as Euler circles and Venn diagrams, lack the power of the full predicate calculus and are relevant to pedagogy or to computational implementations (see, e.g., Glasgow, Narayanan, & Chandrasekaran, 1995). Readers who want a detailed account of Peirce's graphs should consult Roberts (1973). He provides a magisterial survey, and proves the completeness and consistency of the "existential" graphs. In contrast, the aims of the present article are threefold: first, to give an account of the graphs that requires no prior knowledge of logic and avoids Peirce's jargon as much as possible; second, to consider their implications for mental

[2]Sowa (1997, p. 432) writes that Kamp's discourse contexts are isomorphic to Peirce's.

representations, particularly mental models with which they share many characteristics; and, third, to show how they illuminate the elementary operations of reasoning.

PEIRCE'S TWO SYSTEMS OF DIAGRAMS

Peirce devised two separate, though related, diagrammatic systems. His initial inspiration seems to have been chemical diagrams, and he devised diagrams of relations that are remarkably similar to modern semantic networks (see, e.g., Peirce, 3.469). He was also stimulated by Kempe's (1886) graphical notation for mathematical reasoning (see 3.423, and Peirce, 1992, p. 150). Peirce's first system of so-called "entitative" graphs was based on the assumption that each proposition written on a page represented a disjunctive alternative. The system is a precursor to methods of theorem proving developed for computer programs that depend on the so-called *resolution* rule of inference. These programs first transform all the premises containing connectives into disjunctions, because the resolution rule of inference works only on disjunctive premises. The rule has the following form: *p or q*, *r or not-q*, *therefore p or r* (see, e.g., Robinson, 1965). Peirce soon devised a second and superior system of "existential graphs". He published them in 1897, and he regarded them as his greatest contribution to logic (see 4.394, and Lecture 3 of Peirce, 1992).

In an existential graph, or henceforth *graph* for short, each atomic proposition is asserted by writing it on a sheet of paper representing the universe of discourse, where an *atomic* proposition is one that does not contain negation, connectives, or quantifiers. Thus:

Viv is in Rome Pat is in Rome

represents the *conjunction* of the two atomic propositions. To represent the negation of a proposition, it is written within a circle. Circles are typographically inconvenient, and so where possible in the present article they are represented by parentheses. Peirce used the same device. Hence, the following graph:

(Evelyn is in Venice (Leslie is in Venice))

stands for an outer circle containing both the proposition *Evelyn is in Venice* and an inner circle that in turn contains the proposition *Leslie is in Venice*. The graph therefore depicts the following proposition:

It is the not the case that both Evelyn is in Venice and Leslie is not in Venice.

In other words, this graph depicts the proposition: *If Evelyn is in Venice then Leslie is in Venice* granted an interpretation in which a conditional is true in any case except the one in which its antecedent *Evelyn is in Venice* is true and its consequent *Leslie is in Venice* is false.

The graphs exploit the fact that negation and conjunction allow us to represent any of the logical connectives. Indeed, Peirce (4.12) understood that a single connective, equivalent to *not both p and q*, suffices to represent negation and all the logical connectives. The credit for the discovery has usually gone to Sheffer (1913). The graphs can accordingly represent any proposition based on truth-functional connectives, i.e., connectives that form propositions that are true or false depending solely on the truth or falsity of the propositions that they interconnect. The logical connectives: "if-then", "or", and "and", have such idealised senses. For example, a proposition of the form:

p or q, or both

is true provided at least one of its constituent propositions (p, q) is true. Hence, a complex proposition such as:

If p then *either* q and not-r *or* s and t

has a truth-functional interpretation, which is represented in the following graph:

(p (q (r))(s t))

The parentheses represent negating circles and the lower-case letters denote atomic propositions.

Peirce (4.14) referred to the graphs representing sentential connectives as the "alpha" part of the system. He added to it a way to represent quantifiers in order to create the "beta" part, corresponding to the first-order predicate calculus. The calculus is "first order"—the term is Peirce's—because quantifiers range over individuals, but not properties. The graphs use lines to represent the existence of individuals. Each distinct individual has a corresponding line, which can split into several branches. To represent a relation among individuals, these lines must be linked to the predicate in the graph representing the relation. The diagram in Figure 1, for example, has three lines that represent three entities: there is a large cup between a small spoon and a pink saucer. As the diagram illustrates, a line to a predicate that takes only one argument, such as *is large*, represents a property of the entity that is denoted by the line. A predicate that takes two or more arguments, such as *is between*, represents a relation among entities, and it is important to link the entities to their correct position in the predicate. When a line branches—as the line in the figure to *is large* and *is a cup*—it represents a single entity with separate properties, i.e., an entity that is both large and a cup. Lines identifying the existence of a relation among three individuals cannot be constructed from graphs each containing one or two lines of existence, but

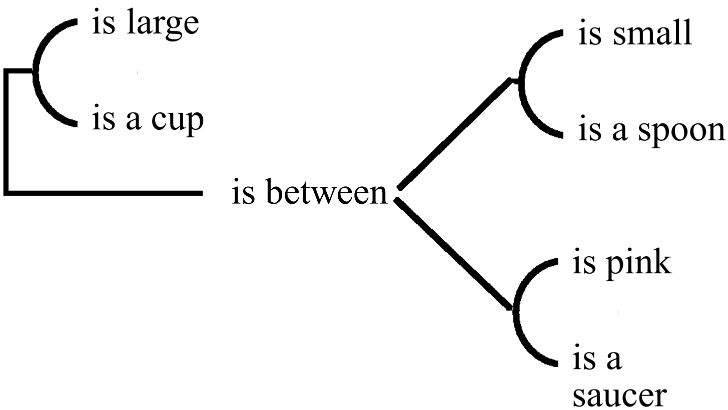


Figure 1. An existential graph of a set of assertions interrelating three individuals.

combinations of graphs with no more than three individuals suffice to graph every relation over more than three individuals (1.347, 4.445).

The logical force of a line in a graph is equivalent to the so-called “existential” quantifier, which can be expressed in English by the indefinite article, *a* or *an*, and by *some*, *something*, *someone*, etc. Hence, the proposition: *Something is tall*, has the graph:

—is tall

And the proposition: *Nothing is tall*, has the graph:

(—is tall)

which is equivalent to: *It is not the case that something is tall*. To represent the so-called “universal” quantifier, which is expressed by *every*, *everything*, *everyone*, as in: *Everything is tall*, graphs exploit the equivalence illustrated by this proposition and its paraphrase: *It is not the case that something is not tall*. The universally quantified proposition: *Everything is tall*, is accordingly represented by the following graph in which the line crosses into the inner circle:

(—(is tall))

which represents: *It is not the case that something is not tall.* Similarly, the graph for: *Something is tall and not thin*, is:

is tall —(is thin)

The universally quantified proposition: *Everything that is tall is thin*, has the negation of the preceding graph:

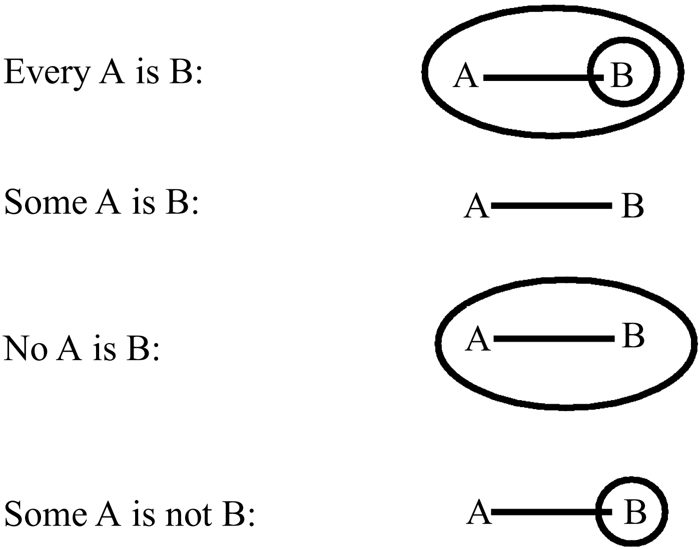


Figure 2. The graphs of the four main sorts of singly-quantified assertions.

(is tall —(is thin))

And the proposition that: *Nothing is tall and thin*, has the graph:

(is tall — is thin)

Figure 2 summarises the graphs of the four main sorts of propositions containing only a single quantifier, which occur in many sorts of inference, including the so-called *syllogisms*, first systematised by Aristotle. The graphs use capital letters to denote predicates.

The key to the interpretation of a line is that its outermost part—the one least enclosed by circles—determines its meaning. When the outermost part of a line is enclosed by an even number of circles, or no circles at all, it signifies the existential quantifier, *some*. When it is enclosed by an odd number of circles, it signifies the universal quantifier, *every*. Hence, the following graph:

(is a man —(loves—is a woman))

represents the proposition:

Every man loves some woman or other.

In contrast, the graph:

(is a man —(loves —))— is a woman

represents the proposition:

There is some woman whom every man loves.

Interpretation accordingly starts with all the lines that are unenclosed, next those that are enclosed once, those that are enclosed twice, and so on. In this way, graphs cope with what logicians refer to as the scope of quantifiers. The first of the two preceding propositions has a universal quantifier that includes the existential quantifier within its scope, as in the “Loglish” sentence: for every x, if x is a man then there is some y such that y is a woman and x loves y. “Loglish” is a language that Jeffrey (1981) invented, and that is half way between natural language and the algebraic notation of the predicate calculus. The second of the preceding propositions has an existential quantifier that includes the universal quantifier within its scope: there is some y such that y is a woman and for every x if x is man then x loves y.

In order to make inferences with graphs, Peirce (4.505) uses no axioms but instead five rules of inference. His system is therefore a precursor to so-called *natural* deduction systems, which are formalised with no axioms but only rules of inference, and which have greatly influenced psychologists committed to formal rules in the mind (e.g., Braine & O’Brien, 1998; Rips, 1994). Shinn (1999) has reformulated Peirce’s rules, exploiting the symmetries in odd versus even numbers of circles, and other visual features in the graphs. The resulting rules produce the same transformations as Peirce’s rules, but bring together common effects on the visual features of the graphs under the same rule. Shinn argues that her system is more efficacious. However, the present paper outlines Peirce’s rules (4.505–4.509) because its concern is not with an efficient tool for reasoning, but the psychological implications of his original system.

The first rule of inference is:

1. The rule of *erasure* and *insertion* (4.505): any graph enclosed by an even number of circles, including the case of no circles at all, can be erased; and any new graph can be added to an area enclosed by an odd number of circles.

Erasure allows the graph:

p (q (r)), which represents: p & not both q & not r

to be transformed into:

(q (r)), which represents: not both q & not r or into:

p.

It allows a line to be severed where it is evenly enclosed. Hence, the graph:

A —(— B), which represents: Something is A & not B

can be transformed into:

A—— —(— B), which represents: Something is A & something is not B

A further erasure yields:

—(— B), which represents: Something is not B.

Erasure, in effect, allows a conjunction to be simplified into one of its constituents. Insertion allows the graph:

(p), which represents: not p

to be transformed into:

(p q), which represents: not both p & q.

The rule accordingly parallels a rule for inclusive disjunctions: *p*, ∴ *p or q or both*. Likewise, within an odd number of circles two separate lines can be joined. Insertion accordingly allows the graph:

A —(— —B), which represents: Something is A & nothing is B

to be transformed into:

A —(——B)), which represents: Something is A & not B.

The second rule of inference allows for graphs to be replicated and removed:

2. The rule of *iteration* and *deiteration* (4.506): any graph with its lines of identity can be copied again in any area within all the circles that enclose the original graph or within additional circles; and any graph can be erased if it could have been the result of iteration.

Iteration allows the graph:

p (q), which represents: p & not q

to be transformed into:

p p (q), which represents: p & p & not q

or into:

p (p q), which represents: p & not both p & q

or into:

p (q q), which represents: p & not both q & q.

What iteration does not allow is the initial graph above to be transformed into:

p q (q), which represents: p & q & not q.

In this case, the iterated q is outside a circle containing the original graph, q. Iteration also allows the transformation of:

(A—B (—C))

into:

(A—B (A—B —C))

One consequence of iteration is that a line can have a new branch added to it provided that the loose end does not cross a circle, although it can abut one. This step allows the transformation from the preceding graph to:

(A—B (A—B —C))

Another consequence of iteration allows a loose end to be drawn inwards through circles, and then to be joined to the corresponding line in an iterated graph. This step allows the transformation from the preceding graph to:

(A—B (A—B —C))

Iteration also allows two loose ends of the innermost parts of lines to be joined together. Deiteration has the effect of reversing the operations of iteration, and so all the preceding transformations can be reversed.

Rule 3 allows the assertion of premises:

3. The rule of *assertion* (4.507): Any graph taken to be true can be drawn with no enclosing circles.

This rule allows each premise to be written as a graph as illustrated in Figure 3.

The effect of drawing two circles round a graph is equivalent to double negation. Because a proposition such as: *not not p*, is equivalent to *p*, the fourth rule allows double circles to be inserted or removed:

4. The rule of *biclosure* (4.508): Two circles can be inserted around any graph or removed from any graph. There must be nothing between the two circles whose significance is affected by the circles.

Biclosure allows the transformation from:

A—B

to:

A—((—B))

and back again.

The final rule allow inessential changes to graphs:

5. The rule of *deformation* (4.509): any part of a graph can be deformed in any way as long as the connections of the parts are not altered.

Table 1 summarises the five rules of inference. Peirce added various other conditions on the rules, but their general principles should be clear. Their use can be illustrated by showing the derivation of a simple inference (of a form known as *modus ponens*):

1. p (p (q)), which uses the rule of assertion (rule 3) to represent the two premises: p; if p then q.

TABLE 1
The five rules of inference for manipulating graphs in order to derive valid conclusions

1. *Erasure* and *insertion*: any graph enclosed by an even number of circles, or none, can be erased; and any new graph can be added to an area enclosed by an odd number of circles.
2. *Iteration* and *deiteration*: a graph can be copied again in any area within all the circles enclosing the original; and any graph can be erased if it could have been the result of the rule of iteration.
3. *Assertion*: Any graph taken to be true can be written with no enclosing circles.
4. *Biclosure*: A double circle can be inserted around any graph or removed from any graph.
5. *Deformation*: any part of a graph can be deformed in any way as long as the connections of the parts are not altered.

2. (p (q)), using the rule of erasure (rule 1).
3. ((q)), using the rule of deiteration (rule 2).
4. q, using the rule of biclosure (rule 4).

Figure 3 presents another illustration of the rules in a graphical derivation of a syllogism of the form:

All A are B.
All B are C.
Therefore, All A are C.

Peirce went on to extend the system to deal with higher-order and modal logic (4.510), but we will not pursue these developments. With a little practice—as Hilary Putnam remarks in his introduction to Peirce (1992)—existential graphs are straightfoward to use. This facility is striking for a graphical system that accommodates the whole of the first-order predicate calculus.

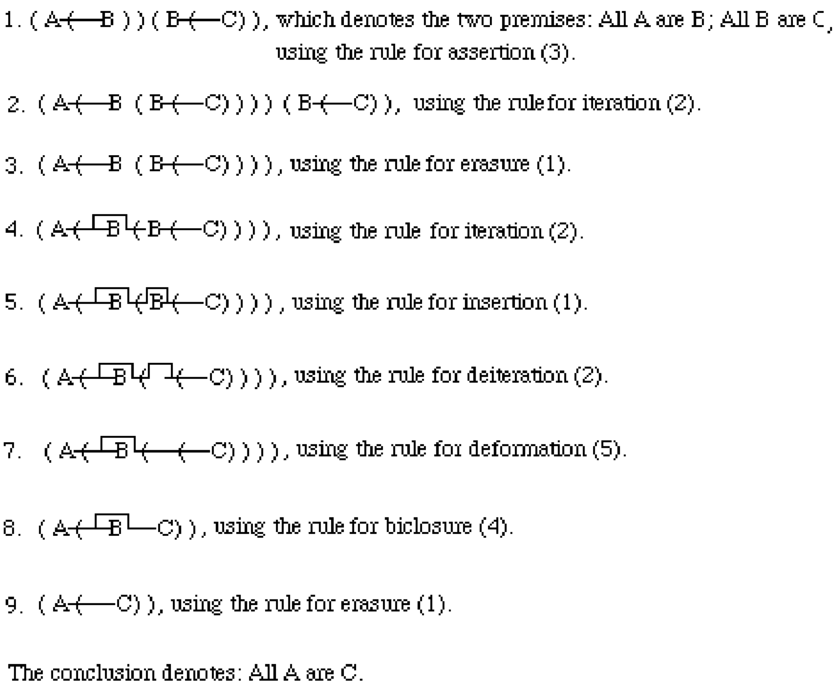


Figure 3. A graphical derivation of a syllogism.

THE IMPLICATIONS FOR MENTAL REPRESENTATIONS

The purpose of most logic diagrams, such as Euler circles and Venn diagrams, is to make reasoning easier. That was *not* Peirce's aim. He wanted instead to improve the *analysis* of reasoning—to take it to pieces and to reveal its elementary steps. He took his existential graphs to be a model of "the operation of thinking" (4.6), not necessarily conscious thinking, though consciousness, he claimed, is a necessary part of deductive reasoning. The graphs are pertinent to current psychological theories of reasoning. These theories are of two main sorts. One sort postulates the use of language-like representations of propositions and the use of formal rules (e.g., Braine & O'Brien, 1998; Rips, 1994) or content-specific rules (e.g., Cheng & Holyoak, 1985) to manipulate these representations. The representations resemble the syntactically structured notations of modern logic, which can be traced back to Peirce's algebraic notation. The other sort of theory postulates that linguistic representations of the meaning of propositions are used to construct mental models of the situation under description (e.g., Johnson-Laird & Byrne, 1991). Reasoning is based on these representations, which can be traced back to Peirce's graphs. They have implications, which this section considers, both for models and for mental representations in general.

The first implication concerns the structure of Peirce's diagrams. He distinguished three properties of signs in general, which included thoughts (e.g., 4.447). They could be *icons*, such as visual images that represent entities in virtue of structural similarity; *indices*, such as an act of pointing, that represent entities in virtue of a direct physical connexion; and *symbols*, such as verbal desciptions, which represent entities in virtue of a conventional rule or habit. These properties can co-exist, and so a photograph is both iconic and indexical. Peirce argued that diagrams should be as *iconic* as possible (4.433). He meant that there should be a visible analogy between a graph and what it represents: the parts of the diagram should be interrelated in the same way that the entities that it represents are interrelated (3.362, 4.418, 5.73). Hence, the existential graphs are more iconic than the entitative graphs (4.434). This feature of iconicity is of fundamental importance to mental representations. It is central to the theory of mental models: "A *natural* model of discourse has a structure that corresponds directly to the structure of the state of affairs that the discourse describes" (Johnson-Laird, 1983, p. 125). Early formulations of iconicity were Maxwell's (1910) analysis of diagrams and Wittgenstein's (1922) "picture" theory of meaning, with its key proposition 2.15: "That the elements of the picture are combined with one another in a definite way, represents that the things [in the world] are so combined with one another." Peirce anticipates both Maxwell and Wittgenstein. His concept of an iconic representation contrasts, as he recognised, with the syntactical symbols of language.

The second implication of the graphs derives from their iconic nature. As Peirce argued (2.279, 4.530), the inspection of a graph reveals truths to be discerned over and above those of the propositions that were used in its construction. In an ideal system, the logical consequences of a set of premises can be "read off" from the representation of the premises. Mental models of spatial and temporal relations have exactly this characteristic (Johnson-Laird, 1983, p. 136). That is to say, their logical consequences are *emergent* properties of representations. For example, the propositions:

The circle is to the right of the triangle.
The triangle is to the right of the square.

support the following model (Johnson-Laird & Byrne, 1991):

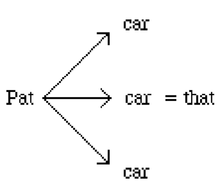
[□ △○]

where the square brackets indicate a model that is accessible in terms of spatial co-ordinates. The model yields the conclusion that *The circle is to the right of the square*, which is valid because it holds in all the models—in this case the single model—that satisfy the premises. Yet this conclusion was not a proposition used in the construction of the model. Moreover, there is no need for an axiom or meaning postulate to capture the transitivity of a relation such as: *to the right of*. The meaning of the relation, which is used to construct and to interpret models, suffices for the inference. This characteristic of spatial and temporal models has been implemented in computer programs (Johnson-Laird, & Byrne, 1991; Schaeken, Johnson-Laird, & d'Ydewalle, 1996).

Discourse about a domain that does not seem to be iconic can be represented in models from which logical consequences emerge. For example, the following inference depends on an abstract relation:

Pat likes all cars.
That is a car.
Hence, Pat likes it.

Yet, if the model theory is right, individuals make this inference by constructing a model of the premises that has the following structure:



where the arrow represents the relation of liking. The conclusion that Pat likes the car (referred to by "that") can be read off from this model. Such a model is only a step away from the corresponding existential graph of the two premises:

((Pat — likes —) — is a car) that — is a car

This example illustrates an important difference between mental models and graphs. Mental models represent a *set* of entities using a corresponding *set* of mental tokens of their predicates, whereas graphs do not. The number of tokens in a mental model is small, but arbitrary—as reasoners similarly assume when they construct external models (Bucciarelli & Johnson-Laird, 1999). The diagram of the mental model just given uses words, but that does not imply that models are made up of words. It is important to bear in the mind the distinction between the diagram and what it purports to represent, namely, a model in the mind. You understand, for instance, what it means for a person to like all cars. Your understanding can be represented in your mind in a way that is independent of your language. Indeed, if such a thought comes to you, it may not always be easy to put it into words. Moreover, if you are bilingual and speak, say, both English and Italian, then you are liable to build the same mental model from the corresponding assertion in Italian.

Can the logical consequences of all propositions be modelled in an emergent way? No one knows. An apparent counterexample is based on the following sort of premises:

> Imagine a world of mirrors in which there are three persons: Pat, Viv, and Les, and in which everyone can see anyone who can see someone. Pat can see Viv. What follows?

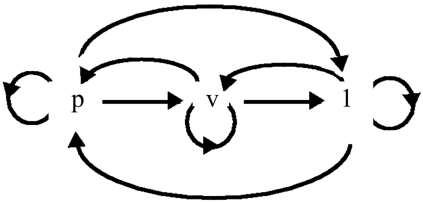
Few people can grasp at once the consequences. What seems to happen is that reasoners construct a sequence of models, which starts with the following model of the categorical premise:

1. $p \longrightarrow v \qquad l$

where “p” represents Pat, “v” represents Viv, “l” represents Les, and the arrow represents the relation of seeing. As Pat can see someone, reasoners can use the universal premise to update the model to:

2. $$\begin{tikzcd} p \arrow[r] & v \arrow[l, bend right] & l \arrow[ll, bend left] \end{tikzcd}$$

This model yields the conclusion that everyone can see Pat. So, everyone can see someone, and the universal premise can be used once more to update the model:

3. 

This model yields the final conclusion that everyone can see everyone, including themselves. Perhaps a super-intelligence could leap immediately to this valid conclusion. For human reasoners, however, it seems that the universal premise must be repeatedly used to update the current model until the process can go no further. This claim was corroborated in an unpublished study carried out by Paolo Cherubini and the author.

The third implication of graphs contrasts with the "possible worlds" that logicians invoke to cope with the semantics of formal languages (e.g. Kripke, 1963). A possible world corresponds to a complete description of a world, and so a proposition divides the infinite set of possible worlds, which includes the real world, into two: those worlds in which the proposition is true and those worlds in which it is false. This division captures the *meaning* of the proposition. Possible worlds are at the heart of both the semantics of modal logic, and, some claim, the semantics of natural language. As a psychological theory, however, they are implausible because when you understand a proposition such as:

The circle is to the right of the triangle

your mind cannot hold the infinite set of possible worlds in which the proposition is true (Partee, 1979). These worlds vary in such matters as the distant apart of the two shapes, their sizes, colours, and so on. But, as Peirce realised, a graph is analogous to what it represents "differing from the objects themselves in being stripped of accidents". The graph represents just those characteristics of the objects that bear on reasoning (4.233). Similarly, the late Jon Barwise (1993) pointed out that a mental model captures what is common to *all* the different ways in which a possibility might occur. Hence, a mental model of the spatial proposition just given represents what is common to any situation in which a circle is to the right of a triangle, but it represents nothing about size, colour, distance apart, and other such "accidents". The model can be updated to take into account information about these properties and relations. But no matter how many details are added to a description, it is always consistent with many distinct states of affairs.

The graphs have implications for a matter under debate: are any mental representations amodal, i.e., not in any sensory modality? Markman and Dietrich (2000, p. 472) write: "Theoretical arguments and experimental evidence suggest that cognitive science should eschew amodal representations." This view has much to commend it. In the past, theorists may have resorted to amodal representations too quickly and underestimated the power of perceptual representations (see Barsalou, 1999, for a similar case). But, as we will see, the moral of Peirce's graphs is that not everything can be represented in an iconic way, and hence there remains a need for amodal representations. Certainly, you can have an iconic representation of a circle to the right of a triangle: you can have a visual image of their spatial arrangement. But, even if we put to one side the question of abstract predicates, there remain at least two elements of meaning that cannot be iconic. The next two implications of the graphs concern them.

The fourth implication of the graphs is that you cannot have an iconic representation of negation. Hence, no visual image can capture the content of a negative assertion, such as:

The circle is *not* to the right of the triangle.

You might form a visual image, say, of a large red cross superimposed on your image of the circle to the right of the triangle. But you have to *know* that the function of the red cross is to symbolise negation, and you have to *know* the semantics of negation—that it reverses the truth value of the corresponding un-negated assertion. And you cannot have a visual image, or any other sort of perceptual representation, that captures either the function of the symbol or its semantics (see Wittgenstein, 1953). Negation cannot be perceived. Hence, the use of circles to represent negation in graphs cannot be perceived. You must know that circles signify negation: you cannot perceive this fact merely by inspecting graphs. Likewise, to interpret graphs you must know the semantics of negation: you cannot perceive this semantics merely by inspecting the graphs. Graphs are not pure icons, but hybrid representations containing what Peirce called symbolic elements (see Newell, 1990).

One putative defence of the perception of negation is that one envisages a *contrast* class. Given the assertion, say:

I did not travel to Manchester by *bike*

reasoners infer that I travelled there by some other appropriate mode of transport. Likewise, given:

The circle is *not* to the right of the triangle.

one might envisage a specific contrast case of the circle *to the left of* the triangle. Individuals do indeed envisage such cases on the basis of the meaning of assertions and their general knowledge (see, e.g., Oaksford & Chater, 1994; Oaksford & Stenning, 1994; Schroyens, Schaeken, & d'Ydewalle, 2001). But the only way in which such classes can capture the full meaning of the negative is if one envisages all the alternative possibilities, e.g., all the possible modes of transport other than by bike by which the speaker might have travelled to Manchester. When the negated term is binary, then there is just a single possibility, e.g., if the door is not open then it is closed. Otherwise, it is necessary to envisage a set of disjunctive possibilities. But, as we will see, such a set itself calls for an amodal representation.

Like graphs, mental models therefore use a symbol to designate negation. Given a simple affirmative assertion, such as:

The circle is to the right of the triangle.

individuals can construct a mental model of the spatial arrangement between the two objects. The model captures what is common to all the different ways in which the possibility might be realised. Given the corresponding negative assertion:

The circle is *not* to the right of the triangle

as we have seen, there is a disjunction of different contrast *classes* in which the assertion would be true. In one class, the circle is above the triangle; in another class, the circle is to the left of the triangle; and so on. Each of these different classes can in turn be realised in infinitely many ways. The theory, however, postulates that individuals can capture all the different classes using negation as an operator on a model:

¬ [△ ○]

where "¬" denotes negation. Individuals have the option of imagining the different classes of possibilities that would be instances of this negated model. That is, they can construct a model of the circle above the triangle, a model of the circle to the left of the triangle, and so on. But they are only likely to use this strategy when a binary meaning or knowledge yields a single possibility. The *falsity* of a negative assertion, such as:

The circle is not to the right of the triangle

is compatible with the only possibility represented in the model of the un-negated proposition. Hence, with true assertions, simple affirmatives are compatible with one class of possibilities and their negations are compatible with many classes of possibilities; whereas with false assertions, simple affirmatives are compatible with many classes of possibilities and their negations are compatible with one class of possibilities. This pattern may account for the well-established interaction in the verification of assertions: people are faster to evaluate true affirmatives than false affirmatives, but faster to evaluate false negatives than true negatives (see, e.g., Clark & Chase, 1972; Wason, 1959).

The fifth implication of the graphs concerns disjunctive possibilities, and the need for their symbolic rather than iconic representation. Peirce abandoned his first system of "entitative" graphs because it treated separate assertions written in a graph as disjunctive alternatives. Thus, the graph:

Viv is in Rome   Pat is in Rome

represents the disjunction: Viv is in Rome or Pat is in Rome, or both. To represent a conjunction in this system it is necessary to use negating circles:

((Viv is in Rome) (Pat is in Rome))

The graph represents: *It is not the case that Viv is not in Rome or Pat is not in Rome*, i.e., *Viv is in Rome and Pat is in Rome*. Peirce argued that his second system of "existential" graphs was more iconic because it treated separate assertions in a graph as conjunctions. But, to represent a disjunction in this system, it is necessary to use negating circles:

((Viv is in Rome) (Pat is in Rome))

which represents: *It is not the case both that Viv is not in Rome and that Pat is not in Rome*. As Peirce realised, there is a one-to-one relation between his two systems: entitative graphs take disjunctions as fundamental, whereas existential graphs take conjunction as fundamental. One passes from one system to the other by way of the following equivalences between conjunctions and inclusive disjunctions:

A & B ≡ not (not-A or not-B)
A or B ≡ not (not-A & not-B)

In fact, both systems rely on a convention: you cannot *perceive* which system of graphs you are looking at. There is no perceptual or iconic system for the representation of disjunctive possibilities. They call for an amodal system of representation.

Mental models contrast with both of Peirce's graphical systems. The evidence suggests that logically naïve individuals represent a disjunction such as: *Viv is in Rome or Pat is in Rome or both*, by considering each of its possibilities separately. That is, they construct three mental models representing these possibilities:

Viv is in Rome	
	Pat is in Rome
Viv is in Rome	Pat is in Rome

Each row in this diagram represents a separate model of a possibility, and the clauses in the premises are represented in a model only when they are true in the corresponding possibility. The theory accordingly takes possibilities as primitive, and each mental model represents a possibility, which is treated as a conjunction of individuals, their properties, and interrelations. Hence, a *set* of mental models represents a disjunction of possibilities. In other words, entitative graphs take negation and disjunction as primitive, existential graphs take negation and conjunction as primitive, and mental models take all three—negation, conjunction, and disjunction—as primitive. The evidence supports the model theory: people understand disjunctions by representing alternative possibilities (e.g., Johnson-Laird & Byrne, 1991).

Both models and graphs are computationally intractable in the technical sense (see Cook, 1971). That is, reasoning makes bigger and bigger demands on time and memory as the number of distinct constituent propositions in an inference increases. These constituents are atomic propositions, which contain neither negation nor sentential connectives. The computational demands can increase so that no system can yield a result, not even a computer as big as the universe running at the speed of light. A set of, say, 100 atomic propositions allows for 2^{100} possibilities, which is a vast number, and in the worst case a test of a deduction calls for check of the truth or falsity of all of them. If one possibility could be checked in a millionth of a second, it would still take over 40 thousand million million years to examine them all. As long as a system is rich enough to reason with sentential connectives, no way round this barrier to efficient computation exists—excluding some as yet undiscovered magical effect in a hypothetical quantum computer. Systems in artificial intelligence have been proposed that give up the expressive power of sentential reasoning in favour of computational tractability, but they must eschew negation, disjunction, and existential quantifiers (Levesque, 1986), that is, precisely those elements that are not iconic. This price is too high for a plausible psychological theory of reasoning.

THE IMPLICATIONS FOR ELEMENTARY OPERATIONS OF THE MIND

What are the fundamental operations of thought? The question is worth raising because Peirce believed that existential graphs provide the answer, and because recent psychologists have sought an answer to it too (e.g. Rips, 1994). Of course, the notion that thought can be decomposed into operations is not accepted by everyone. However, it seems plausible. And if thinking is based on fundamental operations, then it is analogous to computation. The theory of computability reveals distinct ways to construct the set of computable functions. These distinct ways are based on different fundamental operations, and yet the systems turn out to be equivalent to one another in what they can compute. For instance, the apparatus of Turing machines—reading and writing symbols on a tape, and shifting the tape and the state of the machine—yields infinitely many computable functions. The theory of recursive functions shows that a small number of different functions and a small number of different ways of combining them are sufficient to compute anything that is Turing-machine computable. Yet another set of fundamental operations—those that underlie an unrestricted transformational grammar—can also compute anything that is Turing-machine computable. The operations of a Turing machine can in turn be expressed in the first-order predicate calculus. Indeed, in this way, one can prove that no general procedure can exist in the calculus for determining the invalidity of an inference (see Boolos & Jeffrey, 1989). The critical question is: which of these various systems corresponds most closely to the fundamental operations of reasoning? This section of the paper accordingly examines Peirce's answer and the answer from recent accounts of the psychology of reasoning.

In standard accounts of the predicate calculus, proofs depend on three stages: the substitution of the names of real or hypothetical individuals in place of quantified variables, the use of rules of inference for sentential connectives to derive conclusions about the resulting propositions, and finally the restoration of quantifiers in them. As an example, consider the derivation of the following inference:

Every mother has a child.
Elizabeth is a mother.
Therefore, Elizabeth has a child.

In Loglish, the inference is as follows:

1. For every x, if x is a mother then there is some y such that y is a child of x.
2. Elizabeth is a mother.
   Therefore, there is some y such that y is a child of Elizabeth.

The first stage is to eliminate the quantifiers, that is, to instantiate them by replacing them with names. The universal quantifier ranges over everyone in the universe of discourse, and so we can replace it in the first premise with "Elizabeth":

3. If Elizabeth is a mother then there is some y such that y is a child of Elizabeth [from line 1, using a rule for universal instantiation].

A similar rule allows us to replace the existential quantifier, although we must be careful about the temporary name of the hypothetical individual that we substitute for it. This name should not have occurred already in the inference.

Hence, we rewrite the previous line as:

4. If Elizabeth is a mother then Charles is a child of Elizabeth. [from line 3, using a rule for existential instantiation, where “Charles” is the temporary name].

We have now eliminated the quantifiers from the premises, and can proceed to the second stage of reasoning based on sentential connectives:

5. Charles is a child of Elizabeth [from lines 2 and 4, using the rule of modus ponens: if p then q; p; therefore, q].

The third and final stage is to restore quantifiers, using rules of generalisation. Because “Charles” is a temporary name instantiating an existential quantifier, we must restore an existential quantifier:

6. There is some z such that z is a child of Elizabeth.

This assertion is the required Loglish conclusion corresponding to: Elizabeth has a child.

In some psychological theories, the three stages are collapsed into one by rules of inference that take quantified premises and deliver quantified conclusions directly (see Braine & O’Brien, 1998). But this idea will not work for many deductions, because they call for a series of deductive steps in the second stage, after the elimination of quantifiers. The rules in Braine’s system are accordingly incomplete in that they fail to account for many valid deductions that individuals are able to make.

Rips (1994) has pointed out that the number of possible instantiations is an exponential function of the number of variables, and so proofs soon become intractable. He hoped to minimise these problems by eschewing an explicit representation of quantifiers (p. 186). In his system, which is known as PSYCOP, the work of quantifiers is carried out by names and variables in a way akin to certain automated theorem provers in artificial intelligence. For example, PSYCOP represents the sentence:

Every mother has a child

in the following way:

IF Mother(x) THEN Child-of(ax, x)

where “x” stands for a universally quantified variable, and “ax” stands for a *temporary* name with a value dependent on the value of x (i.e. a so-called Skolem function). In addition to variables and temporary names, Rips allows for permanent names, such as "Elizabeth" in the example. He describes the elaborate transformations needed to get from the usual algebraic notation of the predicate calculus to his quantifier-free representations, but suggests that there may be a more direct route to them from sentences in natural language (1994, p. 94). He then introduces four rules for matching one sentence to another in his quantifier-free notation. These rules make reference to an important distinction in PSYCOP between sentences in a derivation, i.e., premises or conclusions derived from them, and subgoals, which PSYCOP sets up in trying to prove a given conclusion. The four rules are as follows:

1. A variable, x, in a subgoal can match another, y, in a sentence, where both x and y derive from universally quantified variables, because logic is not affected by the particular variable representing a universal quantifier.
2. A temporary name in a subgoal can match a temporary or permanent name in a sentence.
3. A permanent name in a subgoal can match a variable in a sentence, i.e., if a sentence applies to all entities in the universe of discourse then it applies to the particular individual in a subgoal.
4. A temporary name in a subgoal can match a variable in a sentence.

These rules are formulated with constraints to prevent invalid inferences. But there is no rule to match a name (permanent or temporary) in a sentence to a variable, x, in another sentence. Hence, the rules do not suffice to draw the conclusion:

Elizabeth has a child

in the example just given. The problem with the required matching rule is that it could lead to a futile sequence of steps ad infinitum (Rips, 1994, p. 193). Hence, the inference can be made only by guessing the conclusion, setting up a subgoal to prove it, and then using rule 3 to make the required step. Rips proves that his system does not lead to invalid conclusions. But it is incomplete in that there are valid inferences that cannot be derived within it. Despite its motivation, the quantifier-free notation does not lead to a computationally tractable system. Studies of automated theorem proving have shown that the problem of intractability re-emerges elsewhere (in the number of different possible uses of a rule of inference). Intractability is inevitable in any complete system for reasoning based on sentential connectives (Cook, 1971). Indeed, human inferential ability rapidly collapses as inferences depend on an increasing number of atomic propositions.

Like the algebraic notation for the predicate calculus, the preceding formal rule theories rely on variables. Substitution of a value for a variable is therefore a potential candidate for a fundamental operation in the process of reasoning. If Peirce is right, however, there is no such fundamental operation. His graphs make no use of variables. They represent individuals as lines in graphs—a representation that Sowa (1997) has argued is more natural. Peirce therefore argued that only two operations are fundamental in reasoning: insertion and deletion (4.374). As Table 1 shows, the rules of inference for graphs insert something into a graph (the rule of insertion and the rule of iteration), or delete something from a graph (the rule of erasure and the rule of deiteration), or both (the rule for double circles).

Mental models are similar to Peirce's graphs. They too contain no variables, and therefore require no substitution of values for variables. The fundamental operations of reasoning based on mental models are insertion (the addition of entities, properties, or relations to models), and deletion (the elimination of models when they are combined with other, inconsistent, models). In the case of reasoning based on quantifiers, the theory also proposes that individuals search for alternative models. In modelling this search in a computer program (see Bucciarelli & Johnson-Laird, 1999), the program refutes conclusions by constructing counterexamples, i.e., alternative models of the premises in which the conclusion is false. To construct counterexamples, the program uses the insertion of new entities into models, and two further fundamental operations—breaking the representation of a single entity into two, and joining the representations of two separate entities into one. The first operation applies to models that contain, for instance, an entity of the form:

artist	beekeeper	chemist

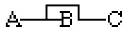
which represents a single individual satisfying the three predicates. Provided that the operation does not violate the meaning of the premises, the representation can be broken into two:

artist	beekeeper	
	beekeeper	chemist

The result represents two separate individuals, one an artist and beekeeper and the other a beekeeper and chemist. This operation refutes, for instance, the following invalid inference:

Some artists are beekeepers.
Some beekeepers are chemists.
Therefore, some artists are chemists.

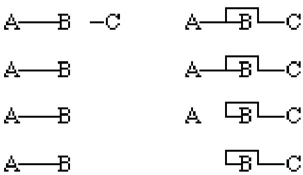
The operation of joining two individuals into one, in essence, reverses the previous step, although the circumstances of its use can be different. These two operations run in parallel with Peirce's system. Given the following graph, which corresponds to Something is A, B, and C:

1. 

the rule of iteration allows the transformation to:

2. 

Successive applications of the rule of erasure then allow the following sequence of transformations:

3.
4.
5.
6.


The result (Something is A and B, and something is B and C) is the same as the operation on the mental model. The difference between the operations on graphs and models concerns their constraints, because the two systems differ in how they represent sets of entities.

There are major discrepancies between the computer program implementing the model theory and human performance in constructing external models, e.g., the program uses only a single strategy, whereas humans use a variety of strategies. However, the program's fundamental operations for constructing models and counterexamples turn out to be remarkably similar to those used by the participants. A major resemblance is that both the program and human reasoners tend to construct initial models that satisfy both the premises and the conclusion. After this step, reasoners refute an invalid conclusion by carrying out one of the following operations: they create a new individual by inserting a new token, they break a token of an existing individual into two, and they join tokens of two separate individuals into one.

CONCLUSIONS

Peirce's existential graphs are remarkable. They establish the feasibility of a diagrammatic system of reasoning equivalent to the first-order predicate calculus. They are therefore a precursor to recent systems of diagrammatic reasoning (see, e.g., Barwise & Etchemendy, 1992). Peirce, however, took pride not in the pedagogical applications of the graphs but in the way they illuminated the mental processes of reasoning. They anticipate the theory of mental models in many respects, including their iconic and symbolic components, their eschewal of variables, and their fundamental operations of insertion and deletion. Much is known about the psychology of reasoning (see, e.g., Evans, Newstead, & Byrne, 1993). But we still lack a comprehensive account of how individuals represent multiply-quantified assertions, and so the graphs may provide a guide to the future development of psychological theory.

Manuscript received 25 January 2001
Revised manuscript received 6 June 2001

REFERENCES

Anderson, M. (1996). Reasoning with diagram sequences. In J. Seligman & D. Westerståhl, (Eds.), *Logic, language and computation, Vol. 1* pp. 35–47. Stanford, CA: Center for the Study of Language and Information.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660.

Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences, 16*, 337–338.

Barwise, J., & Etchemendy, J. (1992). Hyperproof: Logical reasoning with diagrams. In N.H. Narayanan, (Ed.), *AAAI Symposium on Reasoning with Diagrammatic Representations*, (pp. 80–84). Stanford, CA: Stanford University.

Bauer, M.I., & Johnson-Laird, P.N. (1993). How diagrams can improve reasoning. *Psychological Science, 4*, 372–378.

Bochenski, I.M. (1961). *A history of formal logic*. Notre Dame: Notre Dame Press.

Boolos, G.S., & Jeffrey, R.C. (1989). *Computability and logic*. (3rd Edn.) Cambridge: Cambridge University Press.

Braine, M.D.S., & O'Brien, D.P. (Eds.) (1998). *Mental logic*, Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Bucciarelli, M., & Johnson-Laird, P.N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23*, 247–303.

Cheng, P.N., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

Clark, H.H., & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

Cook, S.A. (1971). The complexity of theorem proving procedures. *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, 151–158.

Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Gardner, M. (1982). *Logic machines and diagrams*. (2nd Edn.). New York: McGraw-Hill.

Glasgow, J., Narayanan, N.H., & Chandrasekaran, B. (1995). *Diagrammatic reasoning: Cognitive and computational perspectives*. Cambridge, MA: MIT Press.

Hammer, E.M. (1995). *Logic and visual information*. Stanford, CA: Center for the Study of Language and Information.

Jeffrey, R. (1981). *Formal logic: Its scope and limits*. (2nd Edn). New York: McGraw-Hill.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge: Cambridge University Press; Cambridge, MA: Harvard University Press.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Kamp, H. (1981). A theory of truth and semantic representation. In J.A.G. Groenendijk, T.M.V. Janssen, & M.B.J. Stokhof, (Eds.), *Formal methods in the study of language* (pp. 277–322). Amsterdam: Mathematical Centre Tracts.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*. Dordrecht: Kluwer Academic Publishers.

Kempe, A.B. (1886). Memoir on the theory of mathematical form. *Philosophical Transactions of the Royal Society of London*, *177*, 1–70.

Ketner, K.L. (1991). *Elements of logic*. Lubbock, TX: Texas Tech Press.

Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford: Clarendon.

Kripke, S. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica*, *16*, 83–94.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, *4*, 317–345.

Levesque, H.J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*, 81–108.

Lindsay, R.K. (1988) Images and inference. *Cognition*, *29*, 229–250.

Markman, A.B., & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Science*, *4*, 470–475.

Maxwell, J.C. (1910). Diagram. *Encyclopaedia Britannica* (11th Edn.). (Vol. D, pp. 146–149). New York: The Encyclopaedia Britannica Company.

Newell, A. (1990) *Unified theories of cognition*. Cambridge. MA: Harvard University Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.

Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 835–854.

Partee, B.H. (1979). Semantics—mathematics or psychology? In R. Bäuerle, R. Egli, & A. von Stechow, (Eds.), *Semantics from different points of view*. Berlin: Springer-Verlag.

Peirce, C.S. (1931–1958). *Collected papers of Charles Sanders Peirce*. (8 Vols). Edited by C. Hartshorne, P. Weiss, & A. Burks. Cambridge, MA: Harvard University Press.

Peirce, C.S. (1992). *Reasoning and the logic of things: The Cambridge Conferences Lectures of 1898*. Edited by K.L. Ketner, Cambridge, MA: Harvard University Press.

Rips, L.J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Roberts, D.D. (1973). *The existential graphs of Charles S. Peirce*. The Hague, The Netherlands: Mouton.

Robinson, J.A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery*, *12*, 23–41.

Schaeken, W.S., Johnson-Laird, P.N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.

Schroyens, W., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking and Reasoning*, *7*, 121–172.

Schwartz, S.H., and Fattaleh, D.L. (1972). Representation in deductive problem solving: The matrix. *Journal of Experimental Psychology*, *95*, 343–348.

Sheffer, H.M. (1913). A set of five independent postulates for Boolean algebra, with application to logical constants. *Transactions of the American Mathematical Society*, *14*, 481–488.

Shinn, S-J. (1994). Peirce and the logical status of diagrams. *History and Philosophy of Logic*, *15*, 45–68.

Shinn, S-J. (1999). Reconstituting Beta graphs into an efficacious system. *Journal of Logic*, *Language and Information*, *8*, 273–295.

Sowa, J.F. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.

Sowa, J.F. (1997). Matching logical structure to linguistic structure. In N. Houser, D.D. Roberts, & Evra, J.V. (Eds.), *Studies in the logic of Charles Sanders Peirce* (pp. 418–444). Bloomington, IN: Indiana University Press.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cognitive Science*, *19*, 97–140.

Styazhkin, N.I. (1969). *History of mathematical logic from Leibniz to Peano*. Cambridge, MA: MIT Press. [Originally published in Russian in 1964.]

Wason, P.C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *21*, 92–107.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations* [Trans. G.E.M. Anscombe]. New York: Macmillan.