Cognition 114 (2010) 253–265

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Conceptual illusions

Geoffrey P. Goodwin [a,*], P.N. Johnson-Laird [b]

[a] Department of Psychology, University of Pennsylvania, 3720 Walnut St., Solomon Lab Bldg., Philadelphia, PA 19146, USA
[b] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

ARTICLE INFO

Article history:
Received 20 January 2009
Revised 22 September 2009
Accepted 25 September 2009

Keywords:
Concepts
Boolean concepts
Mental models
Mental representation
Illusions

ABSTRACT

Many concepts depend on negation and on relations such as conjunction and disjunction, as in the concept: *rich or not democratic*. This article reports studies that elucidate the mental representation of such concepts from descriptions of them. It proposes a theory based on mental models, which represent only instances of a concept, and for each instance only those properties, affirmative or negative, that the description asserts as holding in the instance. This representation lightens the demands on working memory, but it also leads to predictable conceptual 'illusions' in which individuals envisage as instances of a concept some cases that in fact are non-instances, and vice versa. Experiments 1 and 2 demonstrated the occurrence of these illusions. Experiment 3 corroborated their results, and showed that the illusions can be alleviated in a predictable way by predicates with certain meanings. These findings cannot be easily explained by alternative theories.

1. Introduction

Many concepts in everyday life depend on combining existing concepts using negation, and such logical connectives as *and* and *or* (Bruner, Goodnow, & Austin, 1956). For example, the concept of a 'ball' in baseball is defined as: a pitch at which the batter does *not* swing *and* which does *not* pass through the strike zone. Systems based on these connectives, and those that can be defined in terms of them, are known as 'Boolean' in honor of George Boole, the logician who first formulated their algebra. Even concepts that are not based on a formal definition depend in part on Boolean connectives. Consider, for example, the concept of ownership conveyed by an assertion of the form, *x owns y*. On one analysis (Miller & Johnson-Laird, 1976, p. 560), the concept means in part: *It is permissible for x to use y, and not permissible for others to prevent x from using y*. Likewise, the concept of a leg, as in *a table's leg*, whether it depends on necessary conditions (e.g., Armstrong, Gleitman, & Gleitman, 1983; Fodor, 1998; Osherson & Smith, 1981), a prototype (Hampton, 1979; Posner & Keele, 1968, 1970; Rosch & Mervis, 1975; Smith & Medin, 1981), exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Palmeri, 1997), general knowledge (Keil, 1989; Murphy & Medin, 1985), or some other hybrid process (Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994; Smith & Minda, 1998, 2000) cannot be grasped without access to a Boolean system, e.g., *a leg has a maximal dimension and is rigid enough that it can support part of what it is a leg of*. A major question is, therefore, how are Boolean relations represented in the mind?

One view is that Boolean relations are represented in a mental language, which contains expressions of the form, e.g., *a and b*, *not a or not b*. These expressions make explicit the logical form of propositions, and the mind contains formal rules of inference for deriving inferences from them (e.g., Rips, 1994, 2002). Likewise, the acquisition of a concept calls for individuals to set up a decision tree that yields a correct classification of instances and non-instances of the concept (e.g., Hunt, 1962), or to find a minimal description consistent with the instances of the concept (Feldman, 2000, 2003).

* Corresponding author. Tel.: +1 215 746 3579; fax: +1 215 898 7301.
E-mail address: ggoodwin@psych.upenn.edu (G.P. Goodwin).

0010-0277/$ - see front matter
doi:10.1016/j.cognition.2009.09.014

An alternative possibility, however, is that individuals represent Boolean concepts in mental models. The model theory postulates that human reasoning depends, not on logical form, but on mental models of possibilities (Johnson-Laird, 2006; Legrenzi, Girotto, & Johnson-Laird, 2003). Individuals use the meaning of expressions and their knowledge to envisage what is possible, and they represent each distinct sort of possibility in a mental model. A conclusion is valid provided it holds in all models of the premises, and it is invalid if there is a counterexample, that is, a model in which the premises hold but not the conclusion. Mental models differ from other proposed sorts of mental representation, such as expressions in a mental language, because models are as iconic as possible: their structures correspond to the structure of what they represent. They can likewise unfold in time kinematically to simulate sequences of events (Goodwin, Bucciarelli, & Johnson-Laird, 2009). But, they do also contain some symbols that are not iconic, such as a symbol for negation (see Peirce, 1931–1958, Vol. 4, for an account of icons and symbols). The model theory provides an explanation of how individuals make deductions, inductions, explanatory abductions, probabilistic inferences, and inferences to default conclusions that hold in the absence of evidence to the contrary (see Johnson-Laird, 2006, for a review).

The model theory extends naturally to the representation of concepts: mental models represent the different sorts of instance of a concept. And a key assumption for concepts, which parallels an assumption about reasoning (e.g., Johnson-Laird & Savary, 1999), is the principle of *conceptual truth*:

> Mental models represent only the instances of a concept, and in each instance they represent only those properties, or their negations, that the description ascribes to the instance.

This principle minimizes both the processing load on working memory, and yields parsimonious representations. But, the principle is subtle, because it applies at two levels. At the first level, mental models represent only the instances of a concept, not its non-instances. At the second level, a mental model represents only those properties, or their negations, that the description asserts as holding in an instance. This point can be best explained by way of an example. Consider a concept based on an exclusive disjunction, such as:

Red or else not square.

Here, and for the rest of the paper, 'or else' refers to an exclusive disjunction, i.e., *A or else B* rules out the possibility of both *A* and *B* holding, whereas 'or' refers to an inclusive disjunction in which both *A* and *B* can hold. According to the principle of conceptual truth, the concept above has two mental models shown here on separate lines:

```
red
      ¬ square
```

where '¬' denotes the symbol for negation. Each model represents a different sort of instance of the concept. One sort consists of instances that are red, and the other sort consists of instances that are not square. Mental models accordingly do not represent non-instances of the concept, such as instances that are *red and not square*, or *not red and square*. Likewise, each instance represents only what the description asserts to hold within it. Hence, the first model does not represent that *not square* does not hold in this sort of instance, i.e., these instances are red squares. And the second model does not represent that *red* does not hold in this sort of instance, the instances are neither red nor square. An alternative description of the concept is accordingly: *red if and only if square*, but individuals do not normally realize that this equivalence holds.

We have written a computer program (in Common LISP) that implements the principle of conceptual truth for any Boolean concept. The program takes as input a description of a concept, which may contain negations, conjunctions, inclusive and exclusive disjunctions, and various other connectives, and its output is a set of mental models that represent the possible sorts of instance of the concept. The program also constructs fully explicit models, which represent the status of all properties in all instances. As we discovered in the output of the program, the mental models of a concept do not always correspond to the fully explicit models of the concept. If individuals follow the principle of conceptual truth they should accordingly make systematic misunderstandings of certain concepts. In order to elucidate this prediction, we consider the workings of the program in more detail.

The program uses a grammar to parse an input description, and each rule in the grammar has a corresponding semantics, so that the parser controls the process of interpretation too. Given, say, the following description:

a or b, and c

the program first constructs mental models of the inclusive disjunction, *a or b*, to yield three sorts of instance:

```
a
      b
a     b
```

It then forms a conjunction of each of them with a model of *c*:

```
a         c
      b   c
a     b   c
```

The mechanisms for forming conjunctions of models are summarized in Table 1. Because any Boolean connective can be defined in terms of negation and conjunction, the program's mechanisms for other connectives can, in effect, be reduced to combinations of conjunctions and negations. The mechanisms in Table 1 contain some subtleties. If one model represents an instance containing the property, *a*, and another model represents an instance containing its negation, ¬ *a*, their conjunction yields the empty (or null) model of a self-contradictory and therefore impossible instance. But, what happens if two particular mental models to be conjoined contain no items in common? Examples illustrating this case occur with this description of a concept based on two exclusive disjunctions: