Cognition 119 (2011) 403–418

Contents lists available at ScienceDirect

Cognition

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

The mental representation of causal conditional reasoning: Mental models or causal models

Nilufa Ali [a], Nick Chater [b], Mike Oaksford [c,*]

[a] Department of Psychology, Anglia Ruskin University, Cambridge, UK
[b] Behavioural Sciences Group, Warwick Business School, Warwick University, Coventry, UK
[c] Department of Psychological Sciences, Birkbeck College, University of London, London, UK

ARTICLE INFO

Article history:
Received 6 May 2010
Revised 2 February 2011
Accepted 4 February 2011

Keywords:
Causal models
Bayesian Networks
Conditional inference
Mental models
Discounting
Augmenting

ABSTRACT

In this paper, two experiments are reported investigating the nature of the cognitive representations underlying causal conditional reasoning performance. The predictions of causal and logical interpretations of the conditional diverge sharply when inferences involving *pairs* of conditionals—such as *if* P_1 *then* Q and *if* P_2 *then* Q—are considered. From a causal perspective, the causal direction of these conditionals is critical: are the P_i *causes* of Q; or symptoms *caused by* Q. The rich variety of inference patterns can naturally be modelled by Bayesian networks. A pair of causal conditionals where Q is an effect corresponds to a "collider" structure where the two causes (P_i) converge on a common effect. In contrast, a pair of causal conditionals where Q is a cause corresponds to a network where two effects (P_i) diverge from a common cause. Very different predictions are made by fully explicit or initial mental models interpretations. These predictions were tested in two experiments, each of which yielded data most consistent with causal model theory, rather than with mental models.

1. Introduction

What is the nature of the cognitive representations underlying human conditional inference? Conditional inference is reasoning about what is often represented in English by *if…then*, such as that embodied in the logical inference rule, Modus Ponens (MP), e.g., if *the key is turned* (*P*), then *the car starts* (*Q*), *the key is turned*, therefore, *the car starts*. This question is of fundamental importance due to the centrality, (i) of the conditional to any theory of reasoning, be it logical, philosophical or psychological, (ii) of hypothetical, "what if," reasoning to all kinds of cognitive processes, from decision making and planning to moral reasoning (Evans, 2007), and (iii) of conditional knowledge in any account of the contents of long term memory.[1] The study of human verbal reasoning has provided various answers to this question, from formal syntactic rules (Braine & O'Brien, 1998; Rips, 1994), semantic representations like mental models (Johnson-Laird & Byrne, 1991; Johnson-Laird, 1983), to probabilistic approaches that argue that *if P then Q* should be read as the conditional probability, i.e., $\Pr(\textit{if P then Q}) = \Pr(Q|P)$ (Evans & Over, 2004; Oaksford & Chater, 2007; Oaksford, Chater, & Larkin, 2000). The latter approach is consistent with the idea that conditionals are mentally represented in something like a Causal Bayes Net (Glymour & Cooper, 1999; Pearl, 1988, 2000). It is this hypothesis that we explore in the experiments presented in this paper.

* Corresponding author. Address: Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK. Tel.: +44 (0) 20 7079 0879.
E-mail address: mike.oaksford@bbk.ac.uk (M. Oaksford).

[1] For example, as Kowalski (2010, p. 253), the founder of logic programming, argues that "conditionals of one kind or another are the dominant form of knowledge representation in Artificial Intelligence," from which cognitive science has typically drawn much of its theoretical inspiration.

0010-0277/$ - see front matter
doi:10.1016/j.cognition.2011.02.005

The majority of current research on conditional reasoning has focused on causal conditionals (Cummins, 1995; Evans, Handley, & Bacon, 2009; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007), where typically the antecedent (*P*) of a conditional describes a cause of the effect described in the consequent (*Q*), as in the above example of the key being turned and the car starting. However, the possible inferential effects of the causal structures described have not been systematically explored in any detail (although see, Sloman & Lagnado, 2005). We argue that patterns of human inference with the conditional may depend crucially on these causal relations represented as causal models (Sloman, 2005).

Sloman, Barbey, and Hotaling (2009) have recently shown how Causal Bayes Nets (Glymour & Cooper, 1999; Pearl, 1988, 2000), or "causal model theory" provides an account of the interpretation of various causal terms, such as, "cause," "enable," and "prevent," which is better empirically confirmed than the account provided by mental models theory (Goldvarg & Johnson-Laird, 2001). The idea is that the mental representations in working memory, which represent the meaning of these terms, is something like the structural element of a Causal Bayes Net (see also, Chater & Oaksford, 2006), which determines the kinds of causal relations believed to exist (Griffiths & Tenenbaum, 2005). This contrasts with the probabilistic or parametric element of these networks, which determines the perceived strength of those relations (Griffiths & Tenenbaum, 2005). Previously, Sloman and Lagnado (2005) presented evidence that similar representations probably do *not* underlie people's reasoning with causal conditionals. In their experiments people responded differently when a causal relation between two events was described as, e.g., "*turning the key* causes *the car to start*" as when it was described as "if *the key is turned* then *the car starts*," where the antecedent is a cause and consequent is its effect. Recently, we have suggested that this evidence is inconclusive (Ali, Schlottmann, Shaw, Chater, & Oaksford, 2010; Oaksford & Chater, 2010a), largely because other manipulations may have had the unintended effect of introducing strength differences between the two locutions while leaving the underlying causal structure the same. Moreover, Sloman and Lagnado (2005) explored quite complex causal scenarios, unlike the much more simple inferences studied in the conditional reasoning task. Here we investigate the effects of casual structure in simple inferences involving pairs of conditional premises that have been much studied in the conditional reasoning literature since Byrne's (1989) classic studies.

In particular, we explore whether effects on reasoning could be observed that could only be predicted on the assumption that the causal conditionals used are represented by the underlying causal structures suggested by causal model theory. Sloman et al. (2009) compared the predictions of the causal model approach with mental models theory and with a recent account of the mental representation of causation based on force dynamics (Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010). To our knowledge, force dynamical models have not been applied to reasoning data in general or to the discounting and augmentation effects, which are the focus of this research, in particular. Consequently, we only compare causal model theory and mental model theory because the latter provides the only other general theory of the mental representations involved in reasoning from which alternative predictions can readily be derived. We first briefly review the literature on causal conditional reasoning before outlining some simple inference problems for which causal model theory (Glymour & Cooper, 1999; Pearl, 1988, 2000; Sloman et al., 2009) makes clear predictions that are not compatible with surface logical form. We then derive contrasting predictions for these cases from mental models theory (Byrne, Espino, & Santamaria, 1999; Goldvarg & Johnson-Laird, 2001; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). We then assess the predictions of each theory in two experiments.

2. Causal conditional reasoning

Since Cummins, Lubarts, Alksnis, and Rist's (1991) seminal paper (see also, Cummins, 1995), a great deal of research has been carried out using causal conditionals, to the extent that it would not be too much of an exaggeration to say that the bulk of what we currently believe about conditional reasoning has been driven by the use of causal materials. We will not review this large literature in any detail but rather focus on the theoretical implications that have been drawn from it. The main conclusion is that when provided with a causal conditional people automatically retrieve information about the whole semantic frame in which a causal conditional is embedded (Markovits & Potvin, 2001). In particular, information about *enabling conditions* and *alternative causes* is accessed and effects are observed similar to those found when such information is provided explicitly (Byrne, 1989; Byrne et al., 1999) as in (1) and (2):

If you turn the key the car starts (1)
If battery is not flat the car starts

If you turn the key the car starts (2)
If you hot-wire it the car starts

So for example, when the second conditional premise in (1)—describing an enabling condition (not having a flat battery)—is included, people draw fewer Modus Ponens inferences to the conclusion that the car starts when the further categorical premise, "the key is turned," is added. Cummins et al. (1991) and Cummins (1995) observed similar effects for materials pre-tested for possible disabling conditions even when the second premise was not explicitly presented. These results suggest that people automatically construct a richer representation of a causal conditional that includes not only the dependency described by that conditional but also possible enablers and alternative causes as Markovits and Potvin (2001) suggested.

There is now a body of evidence showing that the efficiency of retrieval of enabling conditions and alternative causes directly affects inference (De Neys, Schaeken, & d'Ydewalle, 2003a, 2003b; Markovits & Quinn, 2002); that the frequency of enablers or alternative causes matters