Mental Models, Model-theoretic Semantics, and the Psychosemantic Conception of Truth

Shira Elqayam[*]
School of Psychology, University of Plymouth (U.K.)

1 Cognitive Science and the Conception of Truth

Cognitive science research has been an established fact for decades. An interdisciplinary research domain, it involves mutual fertilisation between all disciplines which endeavour to elucidate the human mind — philosophy, psychology, linguistics, mathematics, neurology, and many others. Specifically, in the domain of deduction, philosophers and psychologists have both cooperated and contended over human rationality (or its lack whereof) as reflected in the empirical study of human deductive competence: [Cohen 1981], [Evans 2002], [Evans & Over 1996] are but a few prominent instances. (David Over, in particular, is a striking example of a philosopher who has 'crossed the lines' and conducts empirical work; also see the recent discussion on conditionals involving

∗. Correspondence concerning this article should be addressed to Shira Elqayam, School of Psychology, University of Plymouth, Drake circus, Plymouth, PL8 4AA, UK, e-mail: selqayam@plymouth.ac.uk.

This article is partially based on discussions from a Ph.D. dissertation approved by the senate of Tel-Aviv University. I thank Shulamith Kreitler for her inspiring guidance of the dissertation and her comments on it. I also thank Jonathan Evans and Simon Handley for many enlightening discussions.

philosophers, psychologists and linguists, *e.g.*, [Edgington 2003], [Over & Evans 2003].)

However, even enthusiastic cooperation has its limits, and there are yet many "grey areas" in which philosophy, mathematics and other formal disciplines have much to contribute to the psychological understanding of human cognition, and vice verse. One of the research domains in which cognitive psychology is surprisingly poorer than formal semantic theories is the domain of truth conception. Although in formal literature there is an abundance of truth theories (for review cf. [Cohen 1994], [Grayling 1997]), and although the concept of truth is fundamental in much of the debate in psychology of deduction (for review cf. [Evans, Newstead, & Byrne 1993], [Manktelow 1999]), a systematic effort to conceptualise and validate a psychosemantic truth theory has never been undertaken. From Wason's earliest studies (*e.g.*, [Wason 1960], [Wason & Johnson-Laird 1972]) to present times, in the many decades that the present paradigm of human deductive reasoning has taken its course, not one author has come up with a psychological truth theory.

This, then, is the subject of the present paper: a psychosemantic model of truth, anchored both in formal semantic conceptions of truth, and in psychological theories providing the necessary cognitive machinery for its formulation. Elsewhere I have presented this psychosemantic model of truth in detail ([Elqayam, in press], [Elqayam, submitted]); here I will only sketch its main outlines, and focus rather on the formal and theoretical conception of semantic theories — both philosophical and psychological - that formulate the basis and rationale for such a model.

2 Toward a Psychosemantic Truth Model

What does it mean, then, constructing a psychosemantic truth model? Theoretically, it could mean almost anything: a correspondence theory of truth, a coherence theory of truth, a semantic theory of truth. For a really good correspondence truth theory you need, in a phrase coined by Douglas Adams, the Ultimate Answer to Life, the Universe, and Everything - language, reality, the relationship between them — no wonder that Fodor [Fodor 1980] despaired of it all, calling for methodological solipsism. A coherence theory of truth is almost as bad. Although the Universe can stay out of the project, coherence itself is no mean feat.

However, the awesome dimensions of the task can and have been cut down to manageable bites. The task undertaken by *semantic* truth theories is much more modest — accounting for the behaviour of the truth

predicate. For instance, when Kripke presented his seminal 1975 paper, endeavouring to sketch "an outline of a truth theory"; all he meant, in fact, was to demonstrate how the truth predicate could be defined in a formal, semantically closed language. More specifically, he suggested that a formal language can retain its own truth predicate, thus keeping closer to natural language, and he demonstrated this by outlining a program for defining the truth predicate's potential extension and anti-extension by using truth-value gap schemes such as Kleene's [Kleene1952] or van Fraassen's [van Faassen1966], [van Fraassen 1969].

What I suggest in the present project is quite similar, only this time it regards natural language and psychological cognitive [1] processes rather than formal language and formal semantic procedures. I suggest that the *what* of a cognitive truth theory, that is, its computational level (in Marr's sense of the concept; cf. [Marr 1982]), is a definition of the truth predicate, the instrument for which is an outline of its cognitive extension and cognitive anti-extension; the *how* of a cognitive truth theory, *i.e.*, its algorithmic level explanation, is a semantic, meta-deductive theory that will show the meta-cognitive processes from which tactics and strategies for truth assignment — valuation — can be derived. [2] The theory needs not — indeed, it cannot — demonstrate a painstaking list of all tactics and procedures involved in valuation, at least not at the first stage. All it needs is a viable and robust account of the meta-cognitive processes from which it can be derived.

To be sure, such an agenda is not devoid of its own share of problems. There are problems to semantic truth theory, many of which are a product of just the feature utilised in the present study — the reduction or "deflation" of truth to the extension of the truth predicate [Grayling 1997]. As a working postulate for a cognitive truth model, though, the semantic conception of truth works well. It has the enormous advantage, already pointed up, of reducing the task to manageable proportions.

It should not be construed, however, as more than a working postulate. At this stage of empirical work, even this reduced requirement is quite formidable - so little has been done in the direction of forming a cognitive truth theory. It should also be pointed out, that such a project

1. It should be noted that, throughout this article, the term *cognitive* is meant to denote *psychologically cognitive*; the theoretical commitment of the present work is to empirical evidence of actual thought processes in logically untutored individuals.

2. For a related discussion of *psychologism* — logicians' belief that logic is a generalization of those inferences that people judge as valid [Johnson-Laird & Byrne 1991] - and its mirror-image twin, *logicism* — psychologists' belief that human deductive competence is inherently logical [Evans 2002] - in the context of a psychosemantic truth theory, see [Elqayam 2003].

is beneficial for any sort of future truth conception, be it correspondence, coherence, pragmatic, or any other. For instance, a cognitive correspondence model of truth will need an added component of verification, but it cannot forego the truth predicate component. [3]

The truth predicate is defined by its extension, i.e., the class of all true propositions, and by its anti-extension, which is the class of all entities that are not true propositions (false propositions or non-propositions). A cognitive theory of truth, then, consists of a characterization of the behaviour of the truth predicate in true and false propositions and in non-propositions under various cognitive conditions.

Of course, the truth theory in question does not necessarily have to be a semantic truth theory or a correspondence truth theory; it may as well be a coherence truth theory, the way Fodor suggests [Fodor 1980]. This is an empirical question. I suggest that the empirical findings shown below indeed establish that semantic truth theories are the appropriate normative model for the computational level of a cognitive truth model.

3 Mental model theory as a semantic theory — the theoretical background for a psychosemantic truth model

Of existing psychological theories of human deductive competence, the one explicitly committed to a semantic position is the theory of mental models [Johnson-Laird 1983], [Johnson-Laird & Byrne 1991]. The mental model theory suggests a theory of deductive competence that endeavours to explain various kinds of reasoning as a function of mental model representation. A mental model is 'an internal model of the state of affairs that the premises describe' [Johnson-Laird & Byrne 1991, 35]. It is an extensional representation of a situation, a state of affairs, or a premise, which consists of a finite number of tokens [Johnson-Laird 1983], [Johnson-Laird & Byrne 1991].

3. It does matter, however, that syntactic theories of deduction are explicitly committed to the coherence conception of truth [Fodor 1980], [Rips 1986], [Rips 1990], whereas semantic theories such as the mental model theory seems more inclined to the correspondence side of the dichotomy [Johnson-Laird & Byrne 1990], [Johnson-Laird & Byrne 1991], or, at any rate, to the correspondence interpretation of the semantic conception of truth. The semantic conception of truth seems primary to the model theory, and its correspondence interpretation is secondary. For discussion of the differences between semantic and syntactic psychological accounts of deduction see below.

Since its inception, the mental model theory [Johnson-Laird 1983], [Johnson-Laird & Byrne 1991] has been anchored in formal semantic account. As its computational model, the mental model theory has adopted *model-theoretic semantics*, which is evident in the theory of mental models at all levels, from the construction of a mental model and its manipulation to its validation process.

It should be noted, though, that although model-theoretic semantics is presented in some detail in Johnson-Laird's original account [Johnson-Laird 1983, chapter 8], he has not made an explicit commitment to it as a computational model, but used the more cautious term *precursor*. However, he did state that "model-theoretic semantics should specify *what* is computed in understanding a sentence, and psychological semantics should specify *how* it is computed" (ibid, 167; the italics are mine). This requirement effectively assigns model-theoretic semantics the role of a computational model. (Also see [Hodges 2001], for a discussion for mental models in the context of model theory; and [Johnson-Laird 2002], for a recent discussion of Peirce as a precursor or mental model theory in reasoning.)

According to the theory of mental models [Johnson-Laird 1983], [Johnson-Laird & Byrne 1991], reasoning consists of three stages: *comprehension, description, and validation.* At the *Comprehension* stage, reasoners construct models representing the state of affairs, or the situation, presented in the premises, using their knowledge of language and discourse, their knowledge of the world, their perceptions, or any other potentially relevant knowledge. The product of this stage consists of separate mental models, each representing one premise.

Since reasoners tend to be economic in their cognitive expenditures (*the economy principle*), these initial mental models do not constitute a full representation of all information contained in the premises. Normally, they represent only situations for which the premise holds true, and of these — only the ones mentioned explicitly in the premises. This is called the *principle of truth* [Johnson-Laird & Savary 1999]. For instance, initial model for the conditional is:

$$\begin{array}{ll} p & q \\ \ldots & \ldots \end{array}$$

where the dotted row stands for situations such as

$$\neg p \quad \neg q$$

which are not explicitly represented, at least not in the first stages of model construction. (For a more extensive and recent discussion of the representation of conditionals in mental models, cf. [Johnson-Laird &

Byrne 2002]; for a critique of this position cf. [Evans, Handley & Over 2003], [Evans & Over, in press]).

These provisions are congruous with Barwise' work on situational semantics, and specifically with his notion of a *partial situation* [Barwise & Perry 1983]. Barwise uses model-theoretic techniques to give a semantic interpretation of a sentence that consists of types of situations in which the sentence is true. Like a mental model, a situation is usually incomplete, specifying only some of the properties or relations attributed to individuals within it.

Even the basic mechanisms of mental models are steeped, then, in model-theoretic semantics, in particular situation semantics. The influence of model-theoretic semantics is most evident, however, in Johnson-Laird's approach to the truth conditions of a mental model [Johnson-Laird 1983, 247, 438–442; 1989, 473–474]. The theory of mental models defines a discourse as true if it has at least one mental model that satisfies its truth conditions that can be mapped into the real world model in a way that preserves the content of the mental model; i.e., a discourse is true if its mental model can be mapped into a mental model of the real world. This specification is strongly reminiscent of Tarski's Convention T, that a sentence 'p' is true in a language L if and only if p [Tarski 1944].

Even more specifically, the individuals represented in the mental model should "occur in the real world with the same properties and the same relations holding between them" [Johnson-Laird 1983, 441]. Montague grammar [Montague 1974] plays a clear part in this account, particularly his notion of "translation rules", which interpret a formal system by mapping it systematically into a particular domain. That is, a formal system can be perceived as relating to a particular domain insofar as it can be mapped into it isomorphically; just as a mental model can be perceived as relating to the world insofar as it can be mapped into it (or, to be precise, into its model) isomorphically.

Another version of model-theoretic semantics that influenced this particular analysis directly and explicitly is Kamp's work on discourse models. Kamp [cited by [Johnson-Laird 1983]) suggests that a text represented in a discourse model is true if there is a mapping of the individuals and events in the discourse model into the real world model in a way that preserves their respective properties and the relations between them. [4] Even more pertinent to current analysis is a very relevant observation,

4. This embrace of Montague and Kamp outlooks raises the question of intensional semantics. Whether the semantic account suggested by the model theory should be counted as intensional is moot. On the face of it, mental model theory does

which Johnson-Laird makes in passing, so to speak. He notes that the mapping formulation "applies only to assertions that have definite truth conditions" [Johnson-Laird 1983, 247]. Later in the book, he elaborates on this statement:

If a discourse has complete truth conditions, it is true with respect to the world if and only if it has at least one mental model that can be mapped into the real world. If a discourse has only partial truth conditions [. . .], it is false with respect to the world if it has no mental model that can be mapped into the real world. If its truth conditions are not fixed or known, then, to use Russell's aphorism about mathematics, we never know what we are talking about, nor whether what we are saying is true. Indeed, we cannot know. (442)

Although the term *truth-value gap* is not explicitly mentioned, this is manifestly an account of it. If a discourse has no truth conditions, nothing can be known of its truth status. This observation opens the door to the concept of discourse that has no truth-value, *i.e.*, a truth-value gap.

It should be noted though, that in more recent work, mental model theory view of truth-value gaps has changed from tacit approval to explicit rejection: A recent formulation of mental model theory in the context of conditionals [Johnson-Laird & Byrne 2002] has flatly rejected the possibility of many-valued systems, at least as far as conditionals are concerned.

All these, taken together, constitute quite a firm basis for construction of a truth model within the framework of the theory of mental models. There is convention T to provide an adequacy criterion; there

seem to offer a psychological intensional semantics, due to the theory's adoption of the Montague grammar principle of compositionality; indeed, the model theory has been explicitly characterised as anchored in possible-worlds semantics [Boden, 1988, 177]. The reality, however, is rather more complex, and mental model theory's attitude to possible worlds and intensional semantics has been ambivalent from its beginnings. There is a marked difference between the two major, definitive works of mental model theory [Johnson-Laird 1983], [Johnson-Laird & Byrne 1991], in their treatment of the topic, with most of the discussion contained in the earlier work. And although Johnson-Laird dedicates a relatively extensive discussion to possible worlds and intensional semantics [Johnson-Laird 1983, 56-61, 172–4], and assigns it the role of a computational model for a psychology of meaning, he also points out its intractable nature (*i.e.*, the impossible demands it sets for the human finite mind). In the case of conditionals logic, Johnson-Laird has persisted for two decades in rejecting intensional accounts, such as Stalnaker's [Stalnaker 1968], as intractable [Johnson-Laird 1983, 57–58] or simply impossible to compute [Johnson-Laird & Byrne, 2002, 652]. Indeed, the model theory approach to the conditional has been characterised as exhibiting a profound commitment to extensional propositional logic [Evans & Over, in press].

is an initial sketch of mapping rules; there even used to be preliminary provisions for a truth-value gap apparatus. However, in spite of all these prerequisites, the mental model theory has never come up with a fully developed truth model. It has remained, up until now, one of the myriad potentialities of the theory of mental models that are yet to be elaborated. Constructing a cognitive, psychosemantic truth model and applying it to the knight-knave paradigm may provide the means for gathering crucial empirical evidence in favour of the semantic approach to human reasoning in general, and the model theory in particular.

4 The Knight-Knave Paradigm – Empirical Research of Semantic Concepts

Does it mean, then, that a cognitive truth model simply has to enumerate cognitive truth tables? If so, there hardly seems a need for one to exist. After all, psychological research in propositional logic is almost a century old (for review cf. [Evans *et al.* 1993], [Manktelow 1999]). There is a vast accumulation of empirical and theoretical work concerning the psychological behaviour of various logical connectives - inclusive and exclusive disjunction, conditionals - you name it. There is even some sort of truth-value gap apparatus, in the form of the so-called "defective" truth tables of the conditionals [Evans et al. 1993], [Evans & Over, in press]). Could it be that the extension of the truth predicate is already fully described in the psychological literature discussing deductive reasoning?

Moreover, a certain branch of psychological research in deductive reasoning specifically deals with the concept of truth. I refer to the research in meta-deductive reasoning, pioneered by Rips [Rips 1989], and later taken over by proponents of the mental model theory (*e.g.*, [Johnson-Laird & Byrne 1990], [Johnson-Laird & Byrne 1991]).

Rips [Rips1989], [Rips1994] used puzzles modelled after the ones presented and developed by Smullyan [Smullyan 1978]. These puzzles introduce the delightful island of knights and knaves, in which each of the inhabitants is either a knave, who tells nothing but lies, or a knight, who only tells the truth. The readers of Smullyan's book — and with them, the participants in Rips' experiments, and many later experiments in meta-deduction — are invited to have a stroll over the island, throughout which they keep meeting the island's inhabitants, witnessing their conversations, and overhearing the assertions they make. Here are some typical examples:

1. I am a knave.

2. I am a knight.

3. Either I am a knave or my friend here is a knight.

The first thing that comes to mind while reading proposition 1 is its paradoxical nature. In fact, it is the knight-knave version of the *Liar* paradox (*e.g.*, [Barwise & Etchemendy 1987], [Barwise & Moss 1996], [Kripke 1975], [Martin 1987], [Martin 1984]; [Tarski 1944]). The second proposition, albeit not paradoxical, is not so straightforward either, and constitutes a knight-knave version of the *Truth-teller.* [5]

Since Rips' original contribution, meta-deduction has been dedicated to analysing the way that individuals test concepts of truth and falsity and make deductive conclusions about them. Two competing models of deductive reasoning strive to elucidate the processes underlying meta-deductive. The first one is the syntactic position, which suggests rule-based inference, whose representative in the meta-deductive reasoning literature is Rips' Natural Deduction system [Rips1989], [Rips 1994]. The other position is the semantic one, represented by the mental model theory already outlined above (*e.g.*, [Johnson-Laird & Byrne 1990], [Johnson-Laird & Byrne 1991]).

Rips based his analysis of the knight-knave paradigm on his theory of natural deduction [Rips 1983], [Rips 1994], which advocates mental inference rules, such as *modus ponens*. According to this theory, logical rules constitute a psychological primitive that, applied to propositions represented in working memory, can be used to derive long chains of proofs. The shorter the proof chain, and the more available its rules — the easier is the reasoning task, and the faster and more accurate its performance. Rips adapts this theory to the knight-knave paradigm by adding to the basic propositional rules of his theory four more rules, which define the concepts *knight* and *knave* respectively.

Johnson-Laird and Byrne [Johnson-Laird & Byrne 1990], [Johnson-Laird & Byrne 1991] introduced an alternative explanation based on the theory of mental models. As we already saw, relative to the natural deduction position advocated by Rips, the mental model theory is on the other extreme of the syntax-semantics polarity. It does not accept that deductive competence can be based on formal propositional rules, but endeavours to explain reasoning as a function of mental model representation. Johnson-Laird and Byrne [Johnson-Laird & Byrne 1990],

5. The solution to proposition 3 is that both the speaker and his friends are knights.

[Johnson-Laird & Byrne 1991] assert that the strategy suggested by Rips [Rips 1989] — *i.e.*, a comprehensive testing of the hypothesis chain - is only one out of many strategies the reasoners may develop to deal with the knight-knave puzzles. Moreover, it is a highly unlikely strategy, since it creates a considerable cognitive load. They assert that the meta-cognitive competence involved in observing the cognitive process itself and drawing conclusions about it enables participants in the knight-knave paradigm to develop strategies that are more parsimonious. (For more recent discussions of strategies in mental model theory cf. [Johnson-Laird, Savary, & Bucciarelli 2000], [Van der Henst, Yang, & Johnson-Laird 2002].) They the proceed to re-analyse Rips' data and demonstrate that puzzles that could be solved using one of the simpler strategies were easier to solve than puzzles which could only be solved by the full chain strategy. Byrne and Handley [Byrne & Handley 1997] have subsequently obtained evidence for the use of several strategies for these problems. This approach was so efficient that all subsequent research in the knight-knave paradigm has been done within the general framework of the mental model approach (e.g., [Byrne & Handley 1997]; [Byrne, Handley & Johnson-Laird 1995], [Schroyens 1997], [Schroyens, Schaeken, & d'Ydewalle 1996], [Schroyens, Schaeken, & d'Ydewalle 1999]).

Observing the knight-knave paradigm and its accumulation of empirical evidence, and looking back at decades of empirical research in propositional logic, one can reasonably ask: what can a psychosemantic truth model add that we do not already know? It seems to be all covered by past research.

The answer is that this extensive corpus of research does not constitute a systematic survey of the extension of the truth predicate. There is a theoretical discrepancy between the research in propositional logic and the research in the knight-knave paradigm. This discrepancy is reflected by the lack of any empirical research of truth table production or comprehension in which the atomic propositions presented in the experiments are not simply true or false. The truth table research has normally stuck to the principle of bivalence in the test materials it presented to experiment participants, accepting deviations from it only on the level of participant responses. The much-celebrated "defective" truth tables of the conditional, for instance (see [Evans *et al.* 1993] for a review), consisted of participants constructing or evaluating conditionals in which both antecedent and consequent were determinate; the truth-value gaps were caused by participants' responses, which typically were indeterminate whenever the antecedent was false. (For a recent review of the conditionals literature and particularly the defective truth-tables,

cf. [Evans & Over, in press].) Even if there were any exceptions to this, they remained neglected, and were not taken into consideration in the discussion of the knight-knave paradigm.

I have demonstrated elsewhere [Elqayam 2003], that whatever constitutes an "error" in the knight-knave paradigm cannot be analysed without adopting some sort of a normative system, and many-valued logics are just not good enough for the job. As very different truth valuations can be supported by various normative systems — none of which has any sort of ascendancy over the others — the specific normative system chosen for evaluating participants' responses can have crucial implications for an experiment's results. One has to adopt a truth theory as a normative system, if there is to be any hope of making sense of the results obtained in this paradigm.

The knight-knave paradigm originally seemed to offer a unique opportunity to tackle the role of semantic concepts such as truth and falsity in a theory of reasoning directly rather than by derivation. Rips' main achievement in his 1989 article was precisely that he seemed to demonstrate that a theory of reasoning can do away with semantic concepts even while accounting for semantic contents in the experimental materials.

However, the response suggested by the mental models approach was not as strongly formulated as might have been expected. Indirectly, Johnson-Laird and Byrne [Johnson-Laird and Byrne 1990] did demonstrate the necessity of semantic concepts to any theoretical model endeavouring to explicate the knight-knave paradigm. They showed that Rips' original data could be re-interpreted by a mental model account, and indeed that mental models had better explanatory power than mental logic for this particular set of data; they also introduced the notion of meta-logical reasoning and demonstrated its applicability to the knight-knave paradigm. A theory of meta-logical thinking presupposes a semantic approach: it cannot be formulated without notions of truth and falsity [Johnson-Laird & Byrne 1990, 70].

This approach, persuasive as it is, falls short of a *direct* demonstration that semantic concepts are indeed essential to understanding human reasoning. The knight-knave paradigm offered a unique opportunity for supporting the model approach with a different type of evidence, much more closely related to computational-level considerations; and this opportunity has not been fully utilised or even acknowledged. As already noted, a major difference — if not the primary one — between the model approach and formal rule theories is that the former is semantic whereas the latter are basically syntactic. This semantic commitment is expressed

in no uncertain terms throughout Mental Models [Johnson-Laird 1983], and keeps running through mental model literature, the most recent link in this chain being the [Johnson-Laird 2003] paper on Peirce.

It has been pointed out [Elqayam 2003] that Rips' explicit motivation for the paradigm was his computational formalism [Fodor 1980]; *i.e.*, the commitment to reasoning as an un-interpreted formal system devoid of semantic meaning. This, however, is precisely why the knight-knave paradigm offers the opportunity to address issues of truth and reference within the paradigm, and that directly rather than indirectly. However, the way the paradigm developed within the model theory missed this opportunity by a narrow margin. It accepted unquestioningly the insupportable normative assumptions of Rips' work, rather than trying to answer the challenge headlong, by looking into the very properties of truth and reference. The focus on strategies, while interesting and productive in itself, misses the real potential of the paradigm, which is to tackle basic semantic notions.

I maintain that a direct approach has never been suggested, much less carried out, due to the scarcity of the theoretical background that would have enabled it. Although, as we saw, in previous work Johnson-Laird had laid the groundwork for a truth model, none of these features availed the theory of mental models when it came to deal with the knight-knave paradigm. The missing link is the truth model itself. Without a full-fledged psychosemantic model of truth, the necessary crucial predictions could not have been derived. Had it been formulated, the mental model approach to meta-deduction might have been quite different, and a lot more forceful.

A psychosemantic truth model, then, has become a necessity for cognitive science. Its lack is sorely felt in psychological theories of deduction; and it may shed some light on formal accounts of truth.

5 The Collapse Illusion Hypothesis

I have outlined the suggested psychosemantic model of truth elsewhere [Elqayam, in press], [Elqayam, submitted], and here I will just sketch its main principles and some of the major findings. It is based on complementary semantic approaches: at the formal, computational level, semantic theories containing a partial semantic closure apparatus ([Kripke 1975], [Barwise & Etchemendy 1987], [Barwise & Moss 1996]); and at the algorithmic level, the mental model theory ([Johnson-Laird 1983], [Johnson-Laird & Byrne 1991]). A complementary account is also of-

fered, by the meaning system (e.g., [H. Kreitler & S. Kreitler 1976], [S. Kreitler & H. Kreitler 1990], [S. Kreitler, in press]), which provides the apparatus for dealing with content effects, which should not be neglected in a psychological model.

The major assertions of the psychosemantic truth model can be summarized into one principle: *The partial semantic construction principle* of truth predicate behaviour. Partial semantic construction means that the semantic information reflected in defining the extension of the truth predicate has various gaps, which prevent its full definition.

I will focus on one major hypothesis and the empirical findings that validate it: the *collapse illusion hypothesis*, which is the crucial and most important hypothesis of the psychosemantic truth model. The concept of collapse is based on Rescher's account of collapsed truth tables, which are multi-valued truth tables nested within one another in containment relationship [Rescher 1969]. The partial semantic closure of the truth predicate means that the *Truth-teller* has the potential to be collapsed, *i.e.*, to evoke an illusion that it has a determinate truth-value (T or F). To use Kripke's terms [Kripke 1975], this illusion reduces the *Truth-teller* from a maximal fixed-point status from to a minimal fixed-point status, thus turning it into a seemingly grounded proposition. The *Liar*, on the other hand, is resilient to this sort of illusion since it is a paradoxical proposition. Collapse gives the reasoner the opportunity to conduct the whole valuation process at the minimal fixed point, thus collapsing his multi-valued truth-tables into bivalent, easier ones. Hence, many-valued truth tables in which the indeterminate constituent is the *Liar* were predicted to be less collapsed than many-valued truth tables in which the indeterminate constituent is the *Truth-teller*, and thus weaker in Kleene's sense [Kleene 1957]; or to borrow van Fraassen's term [van Fraassen 1966], [van Fraassen 1969], more radical.

The collapse illusion seems to be compatible with a major observation of the theory of mental models, namely, the *principle of truth*. According to the principle of truth, because reasoners form as parsimonious representations as they can, initial mental models only contain those elements that represent true possibilities of the connective, and of these, only true literals — *i.e.*, elements that match named constituents of the proposition [Johnson, Laird & Byrne 1991], [Johnson, Laird & Savary 1999]. This means that although formally the *Truth-teller* can be collapsed into F just as probably as into T,[6] cognitively the T illusion is much more probable, since reasoners tend to represent the named

6. Indeed, Kripke emphasizes that the assignment of a truth-value to the *Truth-teller* is arbitrary [Kripke 1975, 73].

constituents (in this case, the truth of the proposition).

6 How do untutored reasoners handle the Liar?

Three experiments conducted on college and university students have repeatedly demonstrated just this pattern. Consider the typical finding as shown on figure 1.

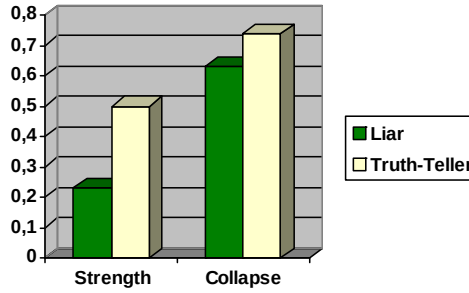

Figure 1: *Truth table aggregated scores of strength and collapse as a function of indeterminate constituent (Liar vs. Truth-teller)*

The function expressing truth-table strength for any participant was computed as:

$$P_S = \frac{\sum F_S}{\sum F_I}$$

where P_S is the computed strength value, $\sum F_I$ is the sum-total of all indeterminate truth functions that the participant has valuated, and $\sum F_S$ is the sum-total of all indeterminate truth functions that the participant has valuated as determinate (T or F).

Similarly, the function expressing truth-table collapse was:

$$P_R = \frac{\sum F_R}{\sum F_I}$$

where P_R is the computed collapse value, $\sum F_I$ is the sum-total of all indeterminate truth functions that the participant has valuated, and $\sum F_R$ is the sum-total of all indeterminate truth functions that the participant has valuated the same as the matching determinate truth-function.

A repeated-measure MANOVA performed for the strength variable produced a significant main effect of indeterminate constituent type (*Liar* versus *Truth-teller*), in which tables for truth functions containing the Liar were considerably and significantly weaker than tables for truth functions containing the *Truth-teller* ($\bar{X} = .2296$ versus $\bar{X} = .5000$ respectively; $F(1, 139) = 1146.526$, $p < .0005$).

Similarly, a repeated-measure MANOVA performed for collapse produced a significant main effect of indeterminate constituent type (Liar versus Truth-teller), in which tables for truth functions containing the Liar were significantly less collapsed than tables for truth functions containing the Truth-teller ($\bar{X} = .6295$ versus $\bar{X} = .7392$ respectively; $F(1, 139) = 18.524$, $p < .0005$).

7 The implications of a psychosemantic truth model

If there is one predicate that can be said to be the most central predicate in semantics and reasoning, the truth predicate is a major candidate for this role. A cognitive truth model, elaborate enough to enable operational definitions and empirical testing, as well as the possibility of validating or refuting, is necessary for many branches of cognitive research. In spite of this obvious need, theoretical contributions until now have been fragmentary at best. Strictly speaking, there has not been until now a cognitive truth model elaborate enough for theoretical work, and explicit enough for empirical testing. Constructing such a model has become a necessity for cognitive science. The suggested model endeavours to fill up this gap.

The suggested model and the empirical findings related to it contribute to understanding the cognitive architecture of truth valuation in several ways:

First and most basically, at the purely descriptive level, the *collapse illusion effect* is a contribution in its own right. It demonstrates that, for untutored individuals of normal intelligence, there is a fundamental difference between various types of propositions that constitute truth-value gaps; or, in Kripke's terminology [Kripke 1975], that the distinction between paradoxical ungrounded propositions and non-paradoxical

ungrounded propositions is psychologically viable. Logically naïve individuals just do not handle the *Liar* as they handle the *Truth-teller*: the former stays indeterminate, whereas the latter is collapsed into a determinate value; it is perceived, in effect, as if it were grounded - or, more specifically, as if it were simply true.

This novel effect is quite robust, having been replicated over three different studies using different populations and different methods. Indeed, it is so robust, that it may well tap something very basic in the semantic makeup of human cognition. This basic feature may be the ability to spot paradoxes. Why this ability is so basic, remains to be studied.

In terms of differentiating between different approaches in psychology of reasoning, the collapse illusion offers a unique opportunity. It is a semantic phenomenon; the difference between the *Liar* and the *Truth-teller* is a semantic difference, anchored in semantic truth theories at the computational level of explanation. Semantic parameters of truth and reference, then, make all the difference in this effect. There is some irony in this conclusion, considering that the knight-knave paradigm, whose adaptation was utilised in the collapse illusion paradigm, has first been contrived for the sole purpose of establishing just the opposite assertion, namely, that semantic conceptions of truth and references were superfluous for a theory of reasoning. The collapse illusion effect stands in stark contrast to this syntactic position, and is very difficult for syntactic theories to explain away. Syntactic theories would have predicted no differences between various sorts of truth-gaps; from a purely syntactic viewpoint, any truth-value gap, whatever its source, has the same status and produces the same sort of truth tables. Thus, the collapse illusion effect scores significantly for the semantic side of the semantics-syntax controversy in deductive reasoning.

References

Barwise, Jon & Etchemendy, John

1987 *The Liar: An Essay on Truth and Circularity*, Oxford, UK: Oxford University Press.

Barwise, Jon & Moss, Lawrence

1996 *Vicious Circles: On the Mathematics of Non?Wellfounded Phenomena*, Stanford, CA: Center for the Study of Language and Information.

Barwise, Jon & Perry, John

1983 *Situations and Attitudes*, Cambridge, Mass.: Bradford Books / MIT Press.

Boden, Margaret A.

1988 *Computer Models of Mind*, Cambrigde, UK: Cambridge University Press.

Byrne, Ruth M.J., & Handley, Simon J.

1997 Reasoning strategies for suppositional deductions. *Cognition*, 62, 1–49.

Byrne, Ruth M.J., Handley, Simon J., & Johnson-Laird, Philip N.

1995 Reasoning from suppositions. *The quarterly journal of experimental psychology*, 48A, 915–944.

Cohen, L. Jonathan

1981 Can human rationality be experimentally demonstrated? *Behavioral and brain sciences*, 4, 317–370.

Cohen, Yael

1994 *Semantic Truth Theories*, Jerusalem: Magness Press.

Elqayam, Shira

2003 Norm, error and the structure of rationality: The case study of the knight-knave paradigm. Semiotica, 147, 265–289.

In press: The meanings of paradox: Meaning preferences in multivalued truth assignment.

In Shulamith Kreitler (Ed.), *The meanings of meaning*, Cambridge University Press.

Submitted: The collapse illusion effect: a semantic illusion of truth and paradox.

Edgington, Dorothy

2003 What if? Questions about conditionals. *Mind & language*, 18, 380–401.

Evans, Jonathan St.B.T.

2002 Logic and human reasoning: An assessment of the deduction paradigm. *Psychological bulletin*, 128, 978–996.

Evans, Jonathan St.B.T., Handley, Simon J. & Over, David E.

2003 Conditionals and conditional probability. *Journal of experi-*

mental psychology: Learning, memory and cognition, 29, 321–335.

Evans, Jonathan St.B.T., Newstead, Stephen E., & Byrne, Ruth M.J.

1993 *Human reasoning: The psychology of deduction*, Hove, UK: Lawrence Erlbaum Associates.

Evans, Jonathan St.B.T., & Over, David E.

1996 *Rationality and reasoning*, Hove, UK: Psychology Press. In press: If. Oxford, UK: Oxford University Press.

Fodor, Jerry A.

1980 Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and brain sciences*, 3, 63–110.

Fraassen, Bas C., van

1966 Singular terms, truth-value gaps, and free logic. The Journal of Philosophy, 63, 481–495.

1969 Presuppositions, supervaluations, and free logic. In K. Lambert (ed.), *The Logical Way of Doing Things*, London: Yale University Press.

Grayling, Anthony C.

1997 *An Introduction to Philosophical Logic*, Oxford, UK: Blackwell.

Hodges, Wilfrid

2001 Model Theory. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2001 Edition,

URL = <http://plato.stanford.edu/archives/win2001/entries/model-theory/>.

Johnson-Laird, Philip N.

1983 *Mental models*, Cambridge, UK: Cambridge University Press.

2002 Peirce, logic diagrams, and the elementary operations of reasoning. *Thinking and reasoning*, 8, 69–95.

Johnson-Laird, Philip N. & Byrne, Ruth M.J.

1990 Meta-logical puzzles: Knights, knaves, and Rips. Cognition, 36, 69–84.

1991 Deduction. Hove, UK: Lawrence Erlbaum Associates.

2002 Conditionals: A theory of meaning, pragmatics, and inference. *Psychological review*, 109, 646–678.

Johnson-Laird, Philip N., & Savary, Fabien

1999 Illusory inferences: a novel class of erroneous deductions. *Cognition*, 71 191–229.

Johnson-Laird, Philip N., Savary, Fabien, & Bucciarelli, Monica

2000 Strategies and tactics in reasoning. In W. Schaeken, G. DeVooght, A. Vandierendonck, and G. d'Ydewalle (Eds.), *Deductive Reasoning and Strategies*, Mahwa, NJ: Lawrence Erlbaum.

Kleene, Stephen C.

1952 *Introduction to Metamathematics*, Groningen: Wolters-Noordhoff.

Kreitler, Hans, & Kreitler, Shulamith

1976 *Cognitive Orientation and Behavior*, New York: Springer.

Kreitler Shulamith (Ed.)

In press: *The Meanings of Meaning*, Cambridge University Press.

Kreitler, Shulamith, & Kreitler, Hans

1990 *The Cognitive Foundations of Personality Traits*, New York: Plenum.

Kripke, Saul

1975 Outline of a Theory of Truth. *The journal of philosophy*, 72, 690–716.

Manktelow, Ken

1999 *Reasoning and Thinking*, Hove, UK: Psychology Press.

Marr, David

1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco: Freeman.

Martin, Robert L. (Ed.)

1978 *The paradox of the Liar*, (2^{nd} ed.). Ohio: Ridgeview.

1984 *Recent Essays on Truth and the Liar Paradox*, Oxford, UK: Oxford University Press.

Montague, Richard

1974 *Formal Philosophy*, New Haven: Yale University Press.

Over, David E. & Evans, Jonathan St.B.T.

2003 The probability of conditionals: the psychological evidence. *Mind & language*, 18, 340–358.

Rescher, Nicholas

1969 *Many-Valued Logic*, New York: McGraw-Hill.

Rip, Lance J.

1989 The psychology of knights and knaves. Cognition, 31, 85–116.

1994 *The Psychology of Proof.* Cambridge: MIT Press.

Schroyens, Walter

1997 Meta-propositional reasoning about the truth or falsity of propositions. *Psychologica Belgica*, 37, 219–247.

Schroyens, Walter, Schaeken, Walter & d'Ydewalle, Géry

1996 Meta-propositional reasoning with knight-knave problems: The importance of being hypothesized. *Psychologica Belgica*, 36, 145–169.

1999 Error and bias in meta-propositional reasoning: A case of the mental model theory. *Thinking and reasoning*, 5, 29–65.

Smullyan, Raymond M.

1978 *What is the Name of this Book?*, Englewood Cliffs: Prentice-Hall.

Stalnaker, Robert

1968 A theory of conditionals. *American philosophical quarterly monograph series*, 2, 98–112.

Tarski, Alfred

1944 The semantic conception of truth and the foundations of semantics. *Philosophy and phenomenological research*, 4, 341?375.

Wason, Peter C.

1960 On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12, 129–140.

Wason, Peter C., & Johnson-Laird, Philip N.

1972 *Psychology of Reasoning: Structure and Content*, London: Batsford.