A proof of the consistency of the mental model theory of sentential reasoning

P. N. Johnson-Laird

The mental model theory of sentential reasoning gives a psychological account of how human reasoners make deductive inferences that depend on sentential connectives, such as *if*, *or*, and, *and* (for an account of the theory, see, e.g., Khemlani, Byrne, & Johnson-Laird, 2018). Critics sometimes ask whether the theory is consistent. They inherit the worry from logicians. In classical sentential logic, an inconsistent set of sentences is devastating: any conclusion whatsoever then follows in a valid deduction. Inconsistency is much less devastating in the model theory: nothing follows from it. But, any theory of reasoning has to allow that the premises of an inference may be inconsistent. This point is important, and so henceforth the claim that a theory is *consistent* means that it yields neither inconsistency – a conclusion of the sort, *B and not B* – when its premises are consistent, nor consistency when its premises are inconsistent. A capital letter, *B*, here and elsewhere can have as its value any elementary sentence or any compound sentence containing one or more negations or sentential connectives.

So, is the model theory of sentential reasoning consistent in the preceding sense? The theory postulates that an intuitive process in system 1 evaluates a given conclusion, but that a subsequent deliberative process in system 2 sometimes corrects this evaluation. Does this difference between the two systems make the theory inconsistent? Not as long as the two systems have separate analyses. On the one hand, intuitions are the source of erroneous inferences in some cases: they can err in treating inconsistent premises as consistent, and vice versa – thereby creating compelling illusory inferences. On the other hand, deliberation in ideal circumstances should not err: it is normative and should be consistent.

Deliberation can build fully explicit models of the premises and of the conclusion, and then compare the two sets of models to evaluate the conclusion. The following summary of the proof aims above all to be comprehensible rather than parsimonious. It shows first that the process of building models is consistent – a proof that depends on several lemmas, and then that the comparison required for the evaluation of conclusions is also consistent. The model theory postulates that human reasoning depends on a parser that uses a grammar and a lexicon to interpret sentences. It interprets them in a compositional way in which each rule in the grammar has a semantic principle for constructing interpretations. It is not clear whether the linguistic component in human beings is consistent; it differs from one speaker to another; and some logicians have argued that natural language is itself inconsistent (e.g., Tarski, 1956). If so, then no theory of human reasoning, whatever its basis, can or should be consistent. Yet, as the proof will show, the linguistic component in the model theory of sentential reasoning is consistent.

There are three sorts of model:

nil, which represents an inconsistent set of sentences,

T, which represents a tautological set of sentences, which refer to all possibilities,

and models such as this example of a single model:

Possibly: △ ¬ ○ □

A symbol such as '△' represents an elementary (aka "atomic") sentence, as expressed in 'there is a triangle,' which does not contain negations or sentential connectives, and the symbol '¬' is for sentential negation, as expressed in, 'it is not that there is a circle'. So, a model is a finite conjunction of elements that each represent elementary propositions or their negations, and so the preceding model corresponds to the sentence:

Possibly (there is a triangle, and there is not a circle, and there is a square).

Its status is as a possibility that holds in default of knowledge to the contrary, unless the set as a whole contains only one model, in which case its status is as a factual claim. Each model in the deliberative system makes explicit every elementary proposition to which the clauses in the premises refer. Any conjunction of two or more models of possibilities, such as:

Possibly: △ ○

Possibly: ¬ △ ¬ ○

is equivalent to an assertion of a conjunction of possibilities, such as:

Possibly (there's a triangle and circle) and possibly (there isn't a triangle and isn't a circle)

The two models of possibilities are consistent with one another, because both possibilities could hold. So, what matters for consistency is that deliberation yields nil from inconsistent premises, T from tautological premises, consistent models from consistent premises, and cannot yield an individual model of this sort:

Possibly: ○ △ ¬ ○ □

This model is inconsistent, because it contains both ○ and ¬ ○, which cannot co-occur in the same possibility.

What follows is an outline of the proof of system 2's consistency. The proof is based on mathematical induction, and it proceeds in a series of lemmas. However, it is necessary to explain how deliberation works in order to demonstrate its consistency.

The process of interpretation yields the models for a sentence in sentential reasoning. The parser uses a grammar in which each syntactic rule has a linked semantic principle. When the parser identifies a simple sentence, which contains neither negation nor a sentential connective, the grammar's linked semantic principle elicits an elementary model of it from the lexicon. So, given the sentence, which may be part of another, such as:

there is a triangle

the semantic principle yields an elementary model:

△

When the parser identifies a compound sentence such as, *Sentence connective Sentence'*, e.g., the exclusive disjunction:

There is a triangle or else there is a square,

where 'or else' denotes a connective, it has already built models for *Sentence* and for *Sentence'*. The semantic principle linked to the parse, *Sentence connective Sentence'* applies the function for the meaning of the connective 'or else' to the relevant sets of models, namely, the models: △and □, respectively. This way of treating compound sentences is recursive. So for a sentence such as:

There is a triangle and, there is a circle or there is a square

the parser's final step elicits a semantic principle to apply the function for 'and' to the model of *a triangle* and the models for *a circle or a square*. Some connectives, such as: *If ... then ...,* have a different sort of grammatical structure, but it can be treated as *Sentence connective Sentence'*.

When the parser identifies the negation of a compound sentence, such as:

It is not that both there is a triangle and there is a square

it elicits the semantic function for 'not' and applies it to the models of *there is a triangle and a square*. None of these processes can introduce an inconsistency provided that the semantic functions for *not* and for the connectives cannot, either. Hence:

Lemma 1: *The parser, the grammar and its compositional principles, which construct models, are consistent if each of the semantic functions in the lexicon is consistent.*

Lemma 2: *Each of the semantic functions in the lexicon is consistent.* The lexicon for sentential reasoning concerns negation, *not*, and the sentential connectives, *if*, *if and only if*, *or*, *or else*, and *and*. The open classes of nouns, verbs, etc. cannot introduce an inconsistency into sentential reasoning, because it does not concern them. The model theory takes the meaning of conjunction (*and*) and sentential negation (*not*) to be fundamental, and it uses them to define all the other connectives. For instance, an exclusive disjunction *A or else B* is defined for deliberation as the following conjunction of two sets of default possibilities, each shown on a separate line, where henceforth the article omits the symbol, 'possibly':

A and not-B

Not-A and B

Capitals in Roman font, such as A, denote the set of models for the sentence, *A*, which can be an elementary clause or a compound.

Table 1 presents the model theory's definitions of the main sentential connectives. The conditional invokes presuppositions, but because they are possibilities too, and they can be ignored except for negation. Inspection shows that the definitions in Table 1 are consistent provided that the meanings of negation *(not)* and conjunction (*and)* are consistent.

Table 1: Deliberation's definitions of five sentential connectives in terms of negation and conjunction, where *A* and *B* stand for simple or compound sentences, and A and B stand for their sets of models in semantic definitions that are conjunctions of default possibilities. A single model for a sentence refers, not to a possibility, but a fact.

Name of compound	Sentence	Semantic definition of connective
Conjunction	*A and B*	A and B
Conditional	*If A then B*.	A and B
		Not-A and not-B
		Not-A and B
Biconditional	*If and only if*	A and B
	A then B.	Not-A and not-B
Inclusive disjunction	*A or B*.	A and B
		A and not-B
		Not-A and B
Exclusive disjunction	*A or else B*.	A and not-B
		Not-A and B

Lemma 3: *Negation is consistent.* If a sentence has an elementary model, such as *a*, then its negation is the model: ¬ *a*; and if a sentence has the model, ¬ *a*, then its negation is: *a*, where lower-case letters denote elementary models such as: △. Suppose instead that a compound sentence, D, has the following conjunction of two models of default possibilities:

a b

¬ a ¬ b

The *partition* of one or more sentences is the set of all possible cases based on the affirmation and negation of the simple clauses in the sentences, e.g., the partition for D, it is:

a b

a ¬ b

¬ a b

¬ a ¬ b

Negation eliminates D's models from the partition and yields the remainder:

a ¬ b

¬ a b

In other words, the negation of D is the complement of D's set of models. The negation of nil, the model of a self-contradiction, is T, the model of a tautology, and vice versa. In one further detail, some compound sentences have models representing presuppositions, which also hold for the negation of the sentences, e.g., for a conditional, *If A then B*, the second and third of its models in Table 1 are presuppositions. Negation adds presuppositions to the models of a negation, and so the negation of *If A then B* has the following models:

A ¬ B

¬ A ¬ B

¬ A B

They are equivalent to the sentence: *If A then not B*. The process of negating a set of models yields the complement of a set of models, with the addition of any presupposed models. It returns nil only for the negation of a tautology. Hence, negation is consistent.

Are individuals aware of all these processes? No, of course not. And they have difficulties in formulating both the possibilities for a denial and its correct description, but the nature of these difficulties corroborates the theory's account of system 1's intuitive process (see Khemlani, Orenes, & Johnson-Laird, 2014).

Lemma 4: *Conjunction is consistent.* Conjunction is needed to deal with compound premises, because it is part of the meaning of each connective (see Table 1) and it is also needed to conjoin an existing set of models with those for a new sentence. It takes two sets of models and forms each pair-wise conjunction of them (aka the 'Cartesian' product of two sets). So, it begins with two sets of models, such as:

a b

¬ a ¬ b

and:

b ¬ c

¬ b c

It forms their pair-wise conjunctions, bearing in mind that if a model from one set contains an element, such as *b*, and a model from the other set contains its negation, ¬ *b*, no need exists to form their conjunction, because it would be a self-contradiction. The conjunction of the models above proceeds as follows:

a b and b ¬ c yields a b ¬ c

a b and ¬ b c do not conjoin because b contradicts ¬ b

¬ a ¬ b and b ¬ c do not conjoin because ¬ b contradicts b

¬ a ¬ b and ¬ b c yields ¬ a ¬ b c

The result of this process of conjunction is two models of possibilities:

a b ¬ c

¬ a ¬ b c

If conjunction never conjoins any pair of models from its two input sets then it yields nil, that is, the two sets contradict one another. If the conjunction yields every case in the partition, then it yields T, that is, the two sets form a tautology. And, if one model in a pair to be conjoined is T, then its conjunction with the other model yields the other model; and if one model in a pair to be conjoined is nil, then its conjunction with the other model yields nil. In sum, conjunction yields a set of models if the two sets are consistent, T if their conjunction yields all possibilities in the partition, and nil if the two sets are inconsistent. Hence, conjunction

is consistent.

Lemma 5: *The parser, lexicon, and the grammar's compositional principles, which construct models from a sentence, are consistent.* The previous lemmas show (1) that the composition of models is consistent if the meanings of connectives and negation are consistent; (2) that the meanings of connectives are consistent if the operations of negation and conjunction are consistent; (3) that negation is consistent, and (4) that conjunction is consistent. So, lemmas (3) and (4) prove lemma (2). And lemmas (2) and (3) prove lemma (1). So, the construction of models from a sentence is consistent. As an additional check, the theory postulates that the deliberative process can test the consistency of each model in a set for a premise. It searches for any element, such as *b*, in a model that co-occurs with ¬*b* in the same model. If it discovers such a conjunction, it rejects the model.

Lemma 6: *The conjunction of the models of premises is consistent.* Given that the models of each premise are consistent (lemma 5), the conjunction of the models of premises into a single set is consistent provided that conjoining one set of existing models (based on earlier premises) with another set (for the current premise) is consistent. This process is consistent if conjunction is consistent. Lemma 4 shows that conjunction is consistent. Hence, the conjunction of models of premises is consistent too.

Lemma 7: *The process in which deliberation constructs models is consistent.* The construction of models for each premise is consistent (lemma 5), and so is the conjunction of its models to those of any previous premises (lemma 6). So, deliberation's construction of models is consistent, and it reflects any inconsistency in premises.

Theorem: *The deliberative evaluation of inferences from premises to given conclusions is consistent.* The process evaluates the alethic status of conclusions as *necessary*, *possible*, or

impossible. It discounts conclusions that have no elementary clauses in common with those of the premises: such conclusions are independent of the premises. A consistent method to establish the three sorts of alethic evaluations of conclusions is to make a conjunction of the models for the premises with the models of the conclusion, where the process of conjunction is consistent (lemma 4):

• If the conjunction yields the same models as those for the conclusion then the premises show that all the models of the conclusion hold: given the premises, the conclusion is *necessary*. An example is:

A and, B or else C.

Therefore, B or else C.

A special case is that the premises yield T, a tautology, in which case a conclusion with no elements outside the premises' partition is necessary. For example:

A and B or else not both A and B;

Therefore, A or B.

• If the conjunction yields the null model then the premises and conclusion contradict one another, and so given the premises, the conclusion is *impossible*. A special case is that the premises yield nil, and in this case, nothing follows from them.

• In any other case, the conjunction yields at least one model that is not nil, and so it is *possible*. For example:

A.

Therefore, A or B.

This example establishes an important point: inferences that are valid in the classical sentential calculus may not be necessary in the model theory.

Any inference falls into one of the three cases, and the process guarantees that the alethic status of the conclusion given the models of the premises and the models of the conclusion. Because the construction of models is consistent (lemma 7) and their conjunction is consistent (lemma 6), the alethic evaluation of conclusions given premises is also consistent, and so theorem holds. The model theory of sentential reasoning is therefore consistent.

The preceding proof shows that the model theory's deliberative system of sentential reasoning is consistent. The system also generates conclusions, but they are a proper subset of those that are necessary according to their evaluation. Hence, in principle, the generation of conclusions is consistent. The proof ignores the consistency of the process of modulation in which knowledge modulates the interpretation of sentences. It operates by forming a conjunction of the models of each premise with fully explicit models based on relevant knowledge. It may therefore eliminate one or more models, but since conjunction is consistent, the process is consistent too. In principle, modulation can also add information, such as temporal relations between events, but such additions take the theory outside sentential reasoning, and so they are not considered here. Finally, sentential reasoning whether based on logic or models or models of default possibilities is computationally intractable—it is NP-complete (Cook, 1971), which means that with an increasing number of independent premises, there comes a point where it is impossible to evaluate an inference, because it calls for more memory and more time than is available to human reasoners or eventually to machines.

References

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.

Khemlani, S.S., Byrne, R.M.J., & Johnson-Laird, P.N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, 1-38. DOI: 10.1111/cogs.12634

Khemlani, S., Orenes, I., & Johnson-Laird, P.N. (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica*, *151*, 1-7.

Tarski, A. (1956). The concept of truth in formalized languages. In Tarski, A. *Logic, semantics, metamathematics: Papers from 1923 to 1938*. (Originally published in 1936.). Oxford: Oxford University Press.