The Premise Paradox
T. Parent (Nazarbayev University)
nontology@gmail.com

1. Introduction

This paper argues that unrestricted talk of “premises” renders a classical system of logic unsound. More precisely: If the predicate ‘is a premise’ is not constrained in some way, one can derive “S is a premise” and “S is not a premise,” for at least one sentence S, in violation of the law of noncontradiction. This is so, even though a “premise” is here defined by a purely formal or syntactic feature: It is a sentence that is underived rather than derived in the context of a proof. More broadly, a classical proof system can be shown unsound even if our object language is semantically open in Tarksi’s (1944) sense.¹

If a paradox arises in a semantically open language, this gives it noteworthy significance. Many of the standard paradoxes (the Liar, Grelling’s paradox, Berry’s paradox, etc.) exploit a semantic term such as ‘is not true’, ‘does not satisfy itself’, ‘is not describable…’, where such terms are defined on expressions using those self-same terms. But the presence of any such term in the language means that the language is semantically closed. Not all paradoxes must arise in semantically closed languages, most notably Curry’s paradox. But the premise paradox below is also one of these exception cases, for it does not depend on any semantic terms, much less semantic terms defined on expressions of its own language.²

¹ Tarski tells us that a language L is “semantically open” iff: L does not contain its own semantic terms, or more precisely, L does not contain semantic terms that are defined on expressions of L.
² Stove (1986, p. 125) and Read (1979) also identify certain paradoxes from problematic premises. However, these other “premise paradoxes” exploit the resources of a semantically closed language. The classic premise paradox is, of course, Carroll’s (1895) fable about Achillies and the Tortoise. But unlike in the present paper, the issue in Carrol’s fable is a vicious regress of premises, not the unsoundness of a proof system.
2. Preparatory remarks

A preliminary matter: In a classical formal system, expressions must be individuated as linguistic *types*. For it is quite easy to show that our derivational system is unsound if an “expression” is individuated as a non-repeatable occurrence.³ Most basically, if an “expression” is a one-time affair, then the Law of Identity will have false instances:

(i) this very expression = this very expression ⁴

(ii) ‘dogs’ = ‘dogs’

(iii) ‘Socrates is mortal’ = ‘Socrates is mortal’

Such identity-statements would be false—for the terms flanking ‘=’ in each case would denote different linguistic occurrences. (In the latter two examples, this holds assuming that a quotation denotes the “expression” inside the quotes.)⁵

One response (raised by William Lycan, in correspondence) is that this shows merely that the terms flanking ‘=’ in these examples would be equivocal. So no sentence would violate the Law of Identity, since the Law requires univocal terms on either side of ‘=’. Semantically speaking, this is quite correct. However, I am here referring to the Law as a *formal* rule, i.e. a generalization whose instances are defined by their shape or syntactic form. Thus construed, the Law says that whenever ‘=’ is flanked by occurrent expressions of the same formal type, the resulting sentence is true. So the language must avoid any form whose occurrences shift denotation. But if expressions are individuated as occurrences, this requirement is violated. E.g., different occurrences of ‘this very expression’ will be of the same formal type yet fail to co-refer.

---

³ An “occurrence” of a type is similar to a token of a type. However, tokens are concrete particulars, whereas occurrences are the abstract parts of a sentence-type or proof-type. (Here I follow Wetzel 2008.)

⁴ Or more precisely: this very demonstrative phrase = this very demonstrative phrase (where demonstrative phrases are a subset of “expressions”).

⁵ Reichenbach (1947) offers some closely related observations; see p. 286 *passim*. (But n.b., Reichenbach speaks of tokens rather than occurrences; cf. n. 3 above)
Further, it is no help to use subscripts to distinguish ‘this very expression₁’ from ‘this very expression₂’. Granted, the Law would not formally entail ‘this very expression₁ = this very expression₂’. Yet it would still formally derive ‘this very expression₁ = this very expression₁’.

Here too, if an “expression” is a single occurrence of a type, it is a false instance of the Law.⁶

So to repeat, at least in a formal system for classical logic, it appears expressions cannot be identified with individual occurrences; they must instead be identified with types.

3. The paradox

The paradox given here consists in proving a falsity from apparently true premises. In what follows, let ‘MP’ name modus ponens (seen as a purely syntactic transformation rule), and let “=Elim” be the formal rule corresponding to the indiscernability of identicals, i.e.,

\[ \forall x \forall y (x = y \supset [F x \equiv F y]) \].

Then, consider the following:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Socrates is mortal</td>
<td>[Premise]</td>
</tr>
<tr>
<td>(2)</td>
<td>(5) is not a premise in this proof</td>
<td>[Premise]</td>
</tr>
<tr>
<td>(3)</td>
<td>(1) = (5)</td>
<td>[Premise]</td>
</tr>
<tr>
<td>(4)</td>
<td>Socrates is mortal ( \supset ) Socrates is mortal</td>
<td>[Trivial]</td>
</tr>
<tr>
<td>(5)</td>
<td>Socrates is mortal</td>
<td>[(4), (1), MP]</td>
</tr>
<tr>
<td>(6)</td>
<td>(1) is not a premise in this proof</td>
<td>[(2), (3), =Elim]</td>
</tr>
</tbody>
</table>

Above, the premises seem true while (6) is false. So the derivational system looks unsound.⁷

In brief, (1) and (5) have contrasting roles in the proof, and that leads to contradiction if (1) = (5). But here is an objection: If we take seriously the type-identity between (1) and (5), then

---

⁶ This problem does not arise when occurrences are referred to descriptively, rather than by means of quote-names or demonstrative phrases. Thus, an unproblematic instance of the (formal) Law of Identity is:

(iv) the left-hand occurrence of a DP in (i) = the left-hand occurrence of a DP in (i).

But the point remains that (i) itself is a false instance of the (formal) Law. It alone renders the formal rule unsound.

⁷ A shorter variant of the proof exists if the reiteration rule is allowed. But I have encountered “purists” who look down on the reiteration rule, and the paradox by reiteration may seem just to confirm their scruples. I thus prefer the modus ponens version above.
since (1) is a premise in the proof, we might insist that (5) is a premise too. In which case, (2) is false. The problem, however, is that (5) is derived by *modus ponens*—yet a premise by our definition is an underived sentence in a proof. Thus, the derived status of (5) means that (2) is true.

This is not sleight of hand. We can make the paradox vivid by asking: Is (5) a premise? Since (5) is a derived sentence, the answer is “no;” however, since (1) = (5), then it is equally true that (1) is a not premise. Yet of course, (1) is a premise.

In the end, this should prompt us to identify (1) and (5) not as numerically one type, but rather as distinct *occurrences* of the type. Then, we can say that (1) is a premise in the proof and (5) is not. (Premise (3) on this interpretation is therefore false.) But although this is the correct thing to say, *it is not something we can say in a classical formal system*. Again, if ‘Socrates is mortal’ is a one-time occurrence, it is then false that ‘Socrates is mortal’ = ‘Socrates is mortal’, contra the (formal) Law of Identity.

Summing up, then, the problem is:

➢ If ‘Socrates is mortal’ is an occurrence, then it falsifies the formal Law of Identity.

➢ If ‘Socrates is mortal’ is a type, then it is both a premise and not a premise in the above proof.

In one way or another, then, ‘Socrates is mortal’ could be used to show that such a derivational system is unsound.\(^8\)

---

\(^8\) It may be wondered how this is possible, given the soundness proof for classical logic. Well, in order to be a proof, the argument for soundness must start with true premises. Yet the soundness argument for classical logic assumes that one cannot derive a falsity from true literals. So if our language contains literals like (2), then not all assumptions of the soundness argument will be true.
4. A non-paradox

One colleague (who prefers anonymity) protests that my argument cannot be correct. For if it were, then the sentence ‘If Socrates is mortal, then Socrates is mortal’ would be paradoxical. Yet it is obviously not.

I myself am unsure about these judgments. There is indeed a parallel argument one could give about ‘If Socrates is mortal, then Socrates is mortal’. But I am not confident that the paradox is real in that instance. Yet let me first present the pertinent reasoning.

Again, assume that expressions in a classical formal system are individuated as types (not occurrences). Then, the following seems to be a proof of a falsity from apparently true premises:

- (1) Socrates is mortal [Premise]
- (2) The consequent of (4) = the antecedent of (4) [Premise]
- (3) The antecedent of (4) is not derived in this proof [Premise]
- (4) Socrates is mortal $\Rightarrow$ Socrates is mortal [Trivial]
- (5) Socrates is mortal [(4), (1), MP]
- (6) The consequent of (4) is not derived in this proof [(3), (2), =Elim]

Uncontroversially, (6) is false. But it is natural to ask is whether premise (3) is true. After all, if we take seriously the type-identity at (2), then we should aver that the antecedent of (4) is derived in the proof. In which case, (3) is false.

The worry, however, is that this suggests that the proof commits the fallacy of affirming the consequent. After all, given the type-identity at (2), (1) is the consequent of (4) whereas (5) is its antecedent. In which case, the derivation of (5) affirms the consequent.

One might reply that the proof utilizes MP rather than affirming the consequent. But while it is true that MP is applied, it would be dicey to suggest that affirming the consequent is not being committed. Again, if we take seriously the type-identity at (2), then it seems that the
proof applies MP and affirms the consequent. Since the latter is sufficient for committing a
fallacy, we would still be left with the implication that the proof commits a fallacy.

Thus, distinguishing MP from affirming the consequent requires distinguishing the
antecedent of (4) from the consequent of (4). Such a distinction would mean that (2) is false. But
again, that distinction could not be made within a classical formal system. ‘Socrates is mortal’
must be individuated as a single linguistic type within the system, and not as an occurrence of a
type. But if it is a single type, it seems that our formal system sometimes allows affirming the
consequent, as per the above proof.

As noted, however, I am not certain that this a genuine problem. The reason is that
affirming the consequent is not always an invalid inferential move. Indeed, the proof above
might be a case in point. But there are less controversial examples as well. E.g.,

(1) If dogs are pets, then canines are pets. [Premise]
(2) Canines are pets.                     [Premise]
(3) Dogs are pets.

Here, the conclusion at (3) is entailed by (2). The argument therefore is valid, even though it has
the form of affirming the consequent. (Granted, the argument is not formally valid, but clearly
there are valid instances of “affirming the consequent” where ‘Socrates = Socrates’ is the
conclusion.)

9 In such examples, at least one premise is not relevant to deducing the conclusion—and as Sanford (1981) has
already shown, arguments with irrelevant premises lead to various puzzles. However, Sanford’s puzzles are
epistemic in nature, concerning what kinds of arguments confirm the explanatory and/or evidential worth of a
premise. But above, the issue is not about whether the premises count as epistemically relevant for the conclusion;
the point is rather that certain arguments with an allegedly fallacious form are nonetheless truth-preserving.

However, Sorensen (2019) discusses Sanford-cases as puzzles for soundness rather than epistemic worth.
Consider argument A: (1) A has a superfluous premise; (2) So, A has a superfluous premise. The premise is false,
since (1) is obviously not superfluous to deducing the conclusion. But A is patently valid. Yet while this is an
interesting case, it presents no real problem. As Sorensen says, we can just regard A as valid but not sound.
This means that, potentially, it is defensible that the conditional version of the paradox displays the fallacy of affirming the consequent—and yet the fallacy is not “fallacious” in this instance. Accordingly, it may be argued that it is not worrisome if we cannot separate MP from affirming the consequent in the relevant proof. Again, there are quite general reasons for thinking that some arguments “affirm the consequent” yet are nonetheless valid. And the validity of the reasoning is what really matters. Hence, it would be hasty to label the argument of this appendix a *bona fide* “paradox.” One might instead just deny (3) and be done with it.¹⁰

The upshot is that ‘If Socrates is mortal, then Socrates is mortal’ need not be seen as paradoxical. That is so, even while admitting the paradoxical import of the proof from section 3.

5. *A proof-theoretic objection*

A different objection to the section 3 proof has been independently raised (in conversation) by Graham Priest and Peter Woodruff. The basic thought is that we should not speak of a sentence type as a premise, but rather say this of a sentence type *relative to a line*—or what amounts to the same, *relative to its position in the sequence* of sentences comprising the proof. In the background is the idea that a proof is an ordered *n*-tuple of sentence types, and a single type can have more than one position in the sequence—just in the way that the number 1 has more than one position in the Fibonacci sequence \(1, 1, 2, 3, 5, 8, 13\ldots\). If so, then in the section 3 proof, we should speak of the type ‘Socrates is mortal’ as indexed to line 1 of the sequence—where it is a premise—and as indexed to line 5 of the sequence—where it is not a premise. In this manner, we may avoid contradiction while avoiding talk of occurrences.

---

¹⁰ Even so, there may be a related *bona fide* paradox where ‘Faa’ and ‘a = b’ are premises. If we infer ‘Fba’ from these, then one can prove that the single linguistic type ‘a’ has and has not been replaced. (That is so, assuming we can make no distinctions between different occurrences, although an appeal to indices may again hold promise.) I cannot discuss this version of the paradox here, but I mention it simply for the reader’s consideration.
There is more than one way to construe the objection, and on one construal, it indeed has potential. (See the next section.) However, on another construal, it changes only the formulation and not the substance of the issue. The claim would be that ‘Socrates is mortal’ should just be understood as a linguistic type which can be assigned multiple indexes in a proof. But even granting this, the following proof would still appear unsound:

(1) Socrates is mortal [Premise]
(2) The sentence at index 5 is not a premise in this proof [Premise]
(3) The sentence at index 5 = the sentence at index 1 [Premise]
(4) Socrates is mortal \( \Rightarrow \) Socrates is mortal [Trivial]
(5) Socrates is mortal [(4), (1), MP]
(6) The sentence at index 1 is not a premise in this proof [(2), (3), =Elim]

In this latest proof, (6) is false. Yet the objector would seem committed to the truth of (2), along with (1) and (3). The objection thus fails, for ‘the sentence at index 1’ and ‘the sentence at index 5’ still refer to numerically the same sentence type. So if the sentence at index 5 is not a premise, then so too with the sentence at index 1.

And just to be clear: If it is instead claimed that the sentence at index 5 is a premise, then by our definition of “premise,” the sentence at index 5 is an underived sentence in the proof. But of course, the sentence at index 5 is derived by MP.

5. Closing remarks

How should the paradox be resolved? The promising suggestion would be to make our talk of a “premise” relativized to an index. Then, ‘Socrates as mortal’ would never count as a premise absolutely. It would instead be a “premise-at-index-1,” and not a “premise-at-index-5.” But refusing (1) the status of a “premise” tout court is a revisionary proposal and seems to have
no precedent in standard discussions of proof calculi (natural deduction, truth trees/tableau, etc.).

To say the proposal is revisionary is not to say that it is revolutionary…its implications may end up being quite minimal. Be that as it may, I am not prepared to defend it here.\footnote{One might try to avoid the index-relativity of a “premise” by building the relativity into “being underived.” Thus, one could claim that ‘Socrates is mortal’ is a “premise tout court” in virtue of being underived-at-some-index. Yet hiding the index-relativity this way seems to obscure rather than clarify the status of ‘Socrates is mortal’ in the relevant proof. (Besides the basic point of the paper would still stand: Without such measures, true premises about premises can lead to absurdity.)}

It is not unprecedented, of course, to recognize that a premise may occur later in a proof. But when it does, the custom has been to say that it is still a premise tout court. Yet when custom speaks of a premise occurring later in the proof, it apparently regards one occurrence of the sentence as a premise, and a distinct occurrence as a non-premise. This view seems quite accurate—but as we saw, a classical formal system must always individuate a sentence as a type, never as an occurrence.

My main purpose here was negative only; it was just to show that under present assumptions, if the predicate ‘is a premise’ has no restrictions on its use, this enables the derivation of an absurdity.\footnote{Tim Button suggests (in conversation) a different solution for the paradox. We might restrict the predicate ‘$x$ is a premise in $y$’ so that values for ‘$y$’ do not include the very proof in which the predicate is used. But this is a constraint on self-reference—the predicate cannot be composed (at the $y$-position) with a term for its own containing proof. Such a constraint would also require further elaboration and defense before it could be affirmed.}

Acknowledgments

Thanks to Axel Barceló, Tim Button, Mihnea Capraru, Ben Jantzen, Cory Johnson, Gary Kemp, Dan Linford, William G. Lycan, Michaelis Michael, Yves Peraire, Martin Pleitz, Graham Priest, Peter Woodruff, and Min Xu for feedback on earlier drafts.
References


