A Failure of Conservativity in Classical Logic

T. Parent (Nazarbayev University)
nontology@gmail.com

Abstract

In a system with identity, quotation, and an axiom predicate, a classical extension of the system yields a falsity. The result illustrates a novel form of instability in classical logic. Notably, the phenomenon arises without vocabulary such as 'true' or 'provable'.

Conservative extensions are safe expansions: They add expressive resources while proving the same theorems (or at most, terminological variants thereof). Conservative extensions are foundational for major developments, including the Löwenheim-Skolem theorems, precise comparisons of proof-theoretic strength (Simpson 2009), and the understanding of reflection principles in arithmetic and set theory (Feferman 1962). The purpose here is not to question these developments, but rather to advise caution for the future. Some extensions that appear quite conservative end up not being so. In a system with identity, quotation, and a metalinguistic singular term, a purely syntactic predicate for axioms can create instability under an innocent-looking extension.[1]

Let T be any consistent first-order theory that includes:

(i) The theorem $c =$ 'd', where 'c' is a constant that names 'd' (though these constants are chosen arbitrarily),

(ii) Among the non-logical axioms, any instance of the formula (α) below, where 'x' is replaced by the quotation of a letter-constant (' 'a' ', ' 'b' ', etc.):

(α) If y is an instance of (α) and an axiom, then the last term in y is not x.

For example, the non-logical axioms of T include:

(α1) If y is an instance of (α) and an axiom, then the last term in y is not 'c'.

(α2) If y is an instance of (α) and an axiom, then the last term in y is not 'd'.

Suppose further that (ii) indicates the only the instances of (α) which are axioms in T. Then, each of those axioms is correct. All the relevant y end in a quotation rather than the letter being quoted.

[1] Collorary 9.9 from Visser (2006) is a conceptually adjacent result; it says (roughly) that if a theory like Q is extended to include a definition of the axiom predicate, the extension fails to preserve features like completeness or finite axiomatizability. In contrast, the conservativity failure below occurs not by adding semantic or reflective strength, but just by relocating one theorem to the set of axioms.

However, let T^+ be a theory that is identical to T, except that it is extended per the following:

(iii) Any sentence derivable from an identity theorem and an axiom is also an axiom.

The extension T^+ appears conservative: it adds nothing that wasn't already a theorem in T. Yet (α1) becomes false in T^+. Note that the following is derivable from $c =$ 'd' and (α2):[2]

(α3) If y is an instance of (α) and an axiom, then the last term in y is not c.

Given (iii), (α3) is an axiom of T^+. It is also an instance of (α) and its last term is 'c'. But as such, it is a counter-example to (α1). So unlike in T, (α1) is false in T^+, even though T^+ adds only theorems from T. The extension is thus *not* conservative, even though its additions are proof-theoretically benign.

The argument shows that an axiom predicate used internally can introduce instabilities when combined with a metalinguistic singular term.[3] The instability arises even in systems free of semantic vocabulary.[4] The axiom predicate widens its applicability in T versus T^+, but this is not unexpected, and its applicability is constant within each system. The shift in applicability is similar to when, e.g., a first-order theory is extended by adding '=' and the correlative axioms. Such a thing is usually seen as innocuous.

In some extensions, a different predicate will count as "the" axiom predicate. But in a theory where 'axiom' is only partially defined, it is possible to assume that the predicate is the same relative to the intended model, when the old axioms are a subset of the new. Indeed, the partial definitions of 'axiom' at (ii) and (iii) would be compatible if no term like 'c' were present.

Regardless, the shift in application conditions for 'axiom' is reasonably seen as causing the issue, even though (again) such a shift is usually thought harmless. However, the aim is not to dispute established theorems, but to expose a subtle fragility in classical logic—one that arises even in the absence of vocabulary like 'true' or 'provable'.[5]

[2]N.B., The requisite substitution of co-referring terms does not occur within quotation marks.

[3]It should be clear that the issue also arises with singular terms besides constants, e.g., where $f(0) =$ '$f(1)$' is a theorem and the axioms include instances of (α) ending with ' '$f(0)$' ' and ' '$f(1)$' '.

[4]Thus, besides the system from Kripke (1973), the instability is relevant to Tarski (1933/1983) as well.

[5]To be clear, there are no anti-Gödelian consequences. Gödel's proof depends on a sentence that denotes the code for that very sentence; it does not exploit a constant for a constant such as 'c'. It also does not appeal to an extension of formal arithmetic.

References

[1] Feferman, S. (1962). Transfinite Recursive Progressions of Axiomatic Theories. *Journal of Symbolic Logic*, 27(3), 259–316.

[2] Kripke, S. (1975). Outline of a Theory of Truth. *Journal of Philosophy*, 72(19), 690–716.

[3] Simpson, S. G. (2009). *Subsystems of Second Order Arithmetic* (2nd ed.). Cambridge University Press.

[4] Tarski, A. (1983). The Concept of Truth in Formalized Languages. In J. Corcoran (Ed.), *Logic, Semantics, Metamathematics* (pp. 152–278). Hackett. (Originally published 1933.)

[5] Visser, A. (2006). Categories of Theories and Interpretations. In A. Enayat, I. Kalantari, & M. Moniri. (Eds.), *Logic in Tehran: Lecture Notes in Logic* 26 (pp. 284–341). ASL and A.K. Peters, Ltd.