The Collapse of the Hilbert Program: Why a System Cannot Prove its Own 1-Consistency

Saul A. Kripke

This is a published abstract from Ewald, W. (2009), 2008 Winter Meeting of the Association for Symbolic Logic. *Bulletin of Symbolic Logic*, 15(2): 229-236, which can be obtained from the publisher at https://doi.org/10.2178/bsl/1243948488.

2008 Winter Meeting of the Association for Symbolic Logic

William Ewald

Bulletin of Symbolic Logic / Volume 15 / Issue 02 / June 2009, pp 229 - 236
DOI: 10.2178/bsl/1243948488, Published online: 15 January 2014

THE BULLETIN OF SYMBOLIC LOGIC
Volume 15, Number 2, June 2009

2008 WINTER MEETING OF THE ASSOCIATION FOR SYMBOLIC LOGIC

The Marriott Hotel
Philadelphia, Pennsylvania
December 27–30, 2008

A meeting of the Association of Symbolic Logic was held December 27–30, 2008, at the Marriott Hotel, in conjunction with the annual meeting of the Eastern Division of the American Philosophical Association. The Program Committee consisted of William Ewald (Chair), Juliet Floyd, and Michael Hallett. The ASL hosted a reception on the evening of December 28th.

The program included two symposia and a special session with an invited talk by Saul Kripke:

Symposium on Historical Ideals of Rigor in Mathematics
Janet Folina (Macalaster College)
Douglas Jesseph (University of South Florida)
Dirk Schlimm (McGill University)

Symposium on Diagrammatic Reasoning in Mathematics
Emily Grosholz (Pennsylvania State University)
Kenneth Manders (University of Pittsburgh)
Sun-Joo Shin (Yale University)

Special Session: "The Collapse of the Hilbert Program"
Saul Kripke (CUNY-Graduate Center)

The program also included one session of contributed papers in which three talks were presented by logicians from the US and abroad.

Abstracts of the invited talks and contributed talks given (in person or by title) by members of the Association for Symbolic Logic follow.

For the Program Committee
WILLIAM EWALD

Abstract of invited Special Talk

► SAUL A. KRIPKE, *The collapse of the Hilbert program: why a system cannot prove its own 1-consistency*.
The City University of New York, Graduate Center, 365 Fifth Avenue, New York, NY 10016, USA.

1079-8986/09/1502-0005/$1.80

Consider a standard system S of number theory—the usual 'Peano Arithmetic' is much more than sufficient. Hilbert's program of the 1920's, had it succeeded, would have proved not only the mere consistency of S, but that every Π^0_2 statement provable in S is true. Gödel's famous paper showed that if the program is to be carried out by means formalizable in S itself, it must fail. His argument, however, used something of a *deus ex machina* as a corollary of an ingeniously constructed undecidable self-referential statement. Here we show that the program can be seen more directly to imply its own collapse, even though it convinced a generation of logicians to believe that it obviously must succeed and that only technical work remained to finish it. In fact, it contains an implicitly self-referential element in and of itself. Moreover, a minimalization argument already used in set theory shows the failure of the program.

The basic ideas of the program are two:

(1) Instead of writing $(\exists x)A(x)$, write $A(\varepsilon xAx)$, where (εxAx) denotes any true instance of $A(x)$, and is arbitrary if there is none. When all quantifiers have been eliminated in this way, one needs only the axiom scheme if $A(t) \supset A(\varepsilon xAx)$, were t is any (ordinary) term. One then needs only propositional logic. Moreover, when particular values are assigned to each ε term, it is always decidable whether a given formula is true. (One could have arbitrary primitive recursive predicates as primitive; one could not bother eliminating bounded quantifiers on the inside.)

(2) By a systematic process of trial and error (somewhat anticipating Friedberg–Mucnik priorities, because of the complexities involved when some ε terms involve others), one tries to find values for the ε terms that would make all the lines of a proof true. Rather than model-theoretically interpreting the axioms, one need only interpret proofs. (If one believes, though not 'officially', that the axioms are true, it might seem obvious that in particular proofs appropriate values for the ε terms can be found.)

Notice that the Hilbert program would imply that every provable Σ^0_1 statement must be true, and hence provable, since it is of the form $A(\varepsilon xAx)$, and a true numerical value for the ε term must be available. Provable Π^0_2 statements must also be true, and instances of the universal quantifier must have provable instantiations. So given any proof p of a Π^0_2 statement $(x)(\exists y)A(x,y)$ where $A(x,y)$ is PR (or bounded quantifier, etc.), and number m, then there is a proof p_1 of an instance $A(0^{(m)}, 0^{(n)})$.

$$(p)\ (m)\ ((\text{p proves a } \Pi^0_2 \text{ statement} \ll (x)(\exists y)A(x,y) \gg) \supset (\exists p_1)\ (p_1 \text{ proves an instance} \ll A(0^{(m)}, 0^{(n)}) \gg)) \qquad (^*)$$

In fact, we can show (assuming the Σ_1 correctness of the system) that not only is this general statement unprovable, but even the instance where we assume p = m must be unprovable:

$$(p)\ ((\text{p proves a } \Pi^0_2 \text{ statement} \ll (x)(\exists y)A(x,y) \gg) \supset (\exists p_1)\ (p_1 \text{ proves an instance} \ll A(0^{(p)}, 0^{(n)}) \gg)) \qquad (^{**})$$

[Since $(^*)$ is the stronger generalization where we don't insist that m = p, and is *a fortiori* unprovable. $(^*)$ is easily seen to be equivalent to the conditional 'if S is consistent, it is 1-consistent'. Hence if we show that $(^{**})$ is unprovable, we have shown that the conditional in question cannot be provable in S itself.

Now $(^*)$, and hence, $(^{**})$ is itself a Π^0_2 statement. (So in fact is the statement that values can be found for the ε terms of every proof. This is the self-referential element implicit in the program: it makes a sweeping Π^0_2 statement about all provable Π^0_2 statements.) Inspection of what it means shows that if it had a proof p, there would have to be a proof p_1 of an instance $A(0^{(p)}, 0^{(n)})$, and n would have to be itself a Gödel number of a proof of the very same instance, which is impossible if we assume that p_1 is the shortest such proof.

Analogously to the Gödel case, $(^{**})$ cannot be provable if S is 1-consistent, but each numerical instance must be. So $(^{**})$ must be undecidable if S is 2-consistent.

Note that we can prove all this about (**) directly, no knowledge of ε terms is required.

A variant argument would use a different statement than (*) claiming the truth, rather than the provability of the appropriate Σ_1 instance.

Yet another variant would argue for the unprovability of (**) by the simpler minimalization argument based on the equivalence of $(\exists w)A$ with $(\exists z)(\exists w < z)A$. I prefer the original argument as showing conceptually what it is involved, and as applicable even if we abandoned arithmetization in favor of a direct formulation of syntax.

The original argument is also somewhat analogous to the oldest argument against the proof of a sort of consistency statement; that one cannot prove in ZF that ZF has an $R(\alpha)$ model (in fact, as observed later, that it has a transitive model). Though one might think this ought to be possible, in fact it is not, since there would be no model of smallest ordinal. The present writer thought of a model theoretic approach to Gödel's theorem (to be presented elsewhere), and finally came up with the present argument as a proof-theoretic analog applicable to the Hilbert program, which purported to replace models by interpreted proofs.

Abstracts of Invited Papers for the invited panel on Diagrammatic Reasoning in Mathematics

► EMILY R. GROSHOLZ, *Combining abstract and concrete representations in mathematics: a case study in textbook exposition*.
Department of Philosophy, 240 Sparks Building, The Pennsylvania State University, University Park, PA 16802, USA.
E-mail: erg2@psu.edu.

New mathematical methods often propose a more abstract way of representing problematic items to be used alongside more familiar and concrete ways of representing them. One well-known example is Descartes' use of the algebra of arithmetic (sporting a perspicuous new notation thanks to him, Vieta and Fermat) in order to classify and investigate geometric curves; another is Poincaré's invention and introduction of the Fundamental Group (first homotopy group) to classify 2-dimensional smooth manifolds, a central problem for the nascent domain of topology. In these applications of novel methods we typically find that the more abstract representation only applies if one of the more concrete representations also applies. Nancy Cartwright in ***The Dappled World*** makes a similar point about reasoning in physics, and uses the suggestive image of dressing up for an occasion: thus we may say here that in novel applications of method a more concrete representation must 'fit out' the abstract representation in a given context of use, a given problem-solving situation. In mathematics textbooks, we often find that diagrammatic representations do the work of more concretely indicating the item that is being talked about, while algebraic notation does the work of more abstractly analyzing and exhibiting the item's 'conditions of intelligibility' and thus the problem's 'conditions of solvability,' phrases that I borrow from Leibniz. These two kinds of representation, juxtaposed and superimposed, are set in rational relation by natural language, which explains how they are to be understood in tandem. To support my claim, I will use two late twentieth century textbook expositions of important results. One immediate consequence of my case studies is that re-writing such expositions in predicate logic will erase the irreducible and productive heterogeneity among the idioms of the two kinds of representation and natural language, and thus fail to capture important aspects of the reasoning.

► KENNETH MANDERS, *Geometrical diagram-inference: philosophical and logical opportunities*.
Philosophy, University of Pittsburgh, 1001 CL, 4200 Fifth Av., Pittsburgh, PA 15260, USA.
E-mail: mandersk@pitt.edu.