

Gödel's mathematics of philosophy

Piergiorgio Odifreddi

February 1995

Mathematicians with an interest in philosophy, such as the present author, find an interest in the latter when they see it as a metaphor or, even better, an inspiration for the former. Gödel provides here a study case, since he is well known to have declared that (part of) his mathematical work was a direct consequence of his philosophical assumptions. If one takes these to mean Gödel's own philosophy in an academic sense, for example as crystallized in the *Nachlass*, then it is difficult to make precise sense of his remark, as some scholars have experienced. The following observations intend to show that the difficulties disappear if one interprets Gödel's 'philosophical assumptions' in a more popular sense, as meaning assumptions of philosophers whose thought he happened to know and find interesting.

We claim that some of Gödel's main results can be interpreted as being mathematically precise formulations of intuitions of Aristotle, Leibniz and Kant. We know of a direct cause-effect connection only in some of the cases, but this is not the point. What we really care about, as non professional readers, is to show that (part of) philosophy can be reinterpreted as asking questions and suggesting answers that mathematics makes precise. Or, to put it in a more general slogan, that in intellectual history everything happens twice: first as philosophy, and then as mathematics.

The arithmetization method (1931)

According to Gerald Sacks,¹ Gödel told him that he got the idea of arithmetization from Leibniz. Sacks seems to be skeptical about the causal effect, and suggests that Gödel may have only thought of it after facts.

¹Personal communication.

Be that as it may, it cannot be denied that in the *Dissertatio de arte combinatoria* (1666) one indeed finds a detailed exposition of a method to associate numbers to linguistic notions. But one also finds a quite surprising naïveté: to associate composite numbers to composite notions, Leibniz proposes (in §69) to use multiplication, thus making decomposition uncertain (since a number may have multiple decompositions).

Gödel's improvement on Leibniz was the use of prime exponentiation in place of multiplication, to make decomposition unique by the Prime Factorization Theorem.

The incompleteness theorem (1931)

Kant's main point in the *Critique of pure reason* (1781 and 1787) and the *Prolegomena to any future metaphysics* (1783) is to show limitations of pure reasons. A quick rendition of the core of his arguments in modern logical language will help us make our point.

Kant uses a system of twelve concepts of the understanding, to which correspond twelve types of judgements, constituting the basis for a system of first-order modal logic. More precisely, the list of judgements in Kant's wording (and their, hopefully self-explanatory, logical translations) is the following:

Quantity	Quality
Universal (\forall)	Affirmative (propositional)
Particular (\exists)	Negative (\neg)
Singular ($\exists!$)	Infinite (first-order)
Relation	Modality
Categorical (atomic)	Problematic (\diamond)
Hypothetical (\rightarrow)	Assertoric (\vdash)
Disjunctive (\vee)	Apodeictic (\square)

Kant claims that the list of categories is complete, and attempts a proof called the transcendental deduction of the pure concepts of the understanding. The word completeness here means 'functional completeness', in the sense in which the usual connectives of classical logic are sufficient to generate any Boolean function.

Kant then claims that if reason is complete, in the different sense of being

able to deal freely with the concepts of understanding, then it is inconsistent. His proof proceed in two steps.

He first concentrates on the three concepts belonging to the Relation group, and claims that completeness allows one to consider a limit version of each of them, called a transcendental idea. More precisely, the three transcendental ideas are *soul*, *first cause*, and *God*: they are obtained, respectively, by pushing to the limit the categories of atomic predicate, implication and disjunction².

The second step of Kant's proof is to show that the transcendental ideas lead to contradiction, which he does by means of the four antinomies of pure reason.

Gödel's work on the incompleteness theorem can be seen as a formal analogue of Kant's own. Indeed, the statement of the incompleteness theorem is a contrapositive of Kant's formulation: a sufficiently powerful and sound formal system that is not inconsistent must be incomplete. And Gödel's proof neatly separates and axiomatizes various aspects of Kant's proof, as follows.

The request that the formal system be sufficiently powerful expresses a weak form of completeness, and it can be rephrased as saying that it is possible to consider a formal version of a particular transcendental idea, associated to (that is, obtained by pushing to the limit) the category of 'non provability', and expressing the statement 'I am not provable'.

The role of antinomies is played here by the contradiction according to which, if 'I am not provable' were provable in a sound system, it would be true, and hence also not provable.

The double negation translation (1933)

In the *Metaphysics* (Γ , especially 1006a), Aristotle isolates two basic axioms of Being:

- the non-contradiction principle, i.e. $\neg(A \wedge \neg A)$ (denied by Heraclitus)
- the excluded middle, i.e. $A \vee \neg A$ (denied by Anaxagoras).

²Thus God is identified with an omnicomprehensive disjunction.

He asserts that a proof of these principles should not be attempted: not everything can be proved, and it is a sign of good education to know when to stop.

One would expect that at this point nothing much would remain to be said, but Aristotle makes an unexpected move: he claims that those principles, although impossible to *prove*, can however be shown to be *impossible to disprove*, in the sense that the assumption of their negation leads to contradiction.

The details of Aristotle's attempted proof, by a method known as the *élenchos*, are of course of no interest to us. But his statement is, being just an assertion of the provability of the double negation of the axioms. This is precisely the core of the double negation interpretation of classical propositional logic into the intuitionistic one, proved by (Kolmogorov in 1925, Glivenko in 1929, and) Gödel in 1933.

More precisely, one way of proving the double negation interpretation (whose statement asserts that if a propositional formula is classically provable, then its double negation is intuitionistically provable) is by showing that the double negation of any tautology is an intuitionistic consequence of instances of double negations of the excluded middle, and then to prove the latter outright in intuitionistic logic (this final step being the formal analogue of the *élenchos*).

Of course, one should always take Aristotle's (or anybody else's) intuitions with a grain of salt: after all, he also claimed (in 1012a) that if the excluded middle failed for some formula, it should fail for all of them; and that if there were more than two truth values, there should be infinitely many. Two assertions proved incorrect by later developments of logic.

Sets and classes (1940)

In his discussion of the ontological proof, Kant rephrases his main point about limitations of reason in a way that ties it more to set theory than to the incompleteness theorem.

More precisely, he observes (A 390–391) that one could distinguish God from beings by saying that that the latter are defined by *sets* of atomic predicates, while the former is 'only' defined by a (proper) *class*: the natural illusion of reason here thus takes the form that it is possible to apply to classes the same properties of sets.

The difference between classes and sets is exploited and made precise in the so-called Von Neumann-Bernays-Gödel system, in which classes are taken as primitive and seen as extensions of arbitrary predicates, while sets are defined as classes belonging to some (other) class. One can then prove that proper classes (i.e. classes that are not sets, of which God is an example according to Kant) are indistinguishable from the whole universe, in the sense that they can all be put into one-one onto correspondence with it (thus uncovering an unexpected pantheistic flavour in Kant's theology).

Cosmological models (1949)

A basic assumption of Kant's *Critique of pure reason* is that space and time do not exist, and are only illusions: although the objects of the external world do appear to us as having spatial and temporal extension, these are not properties of the objects, and are instead a consequence of the structure of our sensorial and mental apparatus (Kant's terminology is that they are *a priori* constituting the form of our perception). In particular, we cannot have any objective knowledge of the world, and are bound to subjective pictures determined by our human nature.

Gödel explicitly states³ that his work on cosmological solutions of the field equations of general relativity was prompted by the question of whether the latter could be shown to be in agreement (or, at least, not in disagreement) with Kant's assumption on time.

On the positive side, he showed that it is indeed consistent with general relativity that there is no notion of absolute and objective time. More precisely, he first constructed a model with rotation of the major mass-points, a condition which is sufficient (and necessary) to show that there is no absolute time in the model. He then provided stronger solutions in which there are even closed time-like lines: in the latter models one could travel in the past by going sufficiently far in the future, and thus even the notion of an objective time for individual observers is ruled out.

On the negative side, Gödel argued that Kant's thesis of the unknowability of things in themselves was however a subjectivistic exaggeration, and not a logical consequence of his assumptions on the *a priori* character of the form of human perception. In particular, relativity theory itself shows that

³*Collected works*, volume III, Oxford University Press, 1995, pp. 274.

properties of space and time (such as the relativity of simultaneousness, the non-Euclidicity of space, the Lorentz contraction, etc.) that flatly contradict our *a priori* intuition can nevertheless be discovered: science thus has the possibility of “going beyond the appearances, and approach the world of things”.

The ontological proof (1970)

According to Hao Wang,⁴ Gödel told him that he got the idea for his version of the ontological proof from Leibniz.

It is clear from the following analysis that Gödel refers to the short paper *Quod Ens perfectissimum existit* (1676), in which Leibniz defines God as a being having all perfections, and proves that God exists because existence is a perfection. The argument is of course Descartes’ version of Anselmus’ own, and Leibniz’s improvement consisted in attempting to show that the definition was not contradictory.

Despite the fact that, according to Leibniz, his argument satisfied Spinoza, it could certainly not satisfy anybody aware of compactness (such as Gödel, who first discovered it). Leibniz indeed makes the following hair-raising step: he only ‘shows’ that perfections are compatible *two by two*, and then deduces from this that they are *all* compatible.

To avoid such an embarrassing mistake, Gödel reformulated the argument as follows (where we have dropped any reference to modalities). He defines God as a being having all positive properties, and assumes that the positive properties form an ultrafilter on the universe: thus closure under intersection (expressing the fact that the positive properties are compatible two by two) is simply postulated. Since any ultrafilter of subsets of a finite set is principal, the existence of God (i.e. the intersection of the ultrafilter) follows immediately from the assumption of finiteness of the universe. To avoid this unsatisfying hypothesis, it is enough to notice that it is used in the argument only to show that the ultrafilter contains its own intersection: if the latter is assumed directly, the result follows without any finiteness assumption. But the intersection of the ultrafilter is God itself, and to assume that it is in the ultrafilter means to assume that ‘being God’ is a positive property, which is what Gödel does.

⁴*Reflections on Kurt Gödel*, MIT Press, 1987, p. 195.

According to this discussion, Gödel simply saw how to turn a faulty argument into a logically correct one. There is no need to see this as a revival of rational theology in the XX century, a windmill against which Burton Dreben battled at the Boston Colloquium on ‘Gödel’s general philosophical significance’ (February 6–7, 1995).

Conclusion

We have shown that certain passages of Gödel’s favorite philosophers can be seen as an inspiration for, or reinterpreted in the light of, some formal developments of his mathematical work. Historical evidence showing factual causal connections, which indeed exists in some of the cases, is not relevant to our main point: namely, that both mathematicians and philosophers may profit from a non academic reading of philosophy.