A Patch to the Possibility Part of Gödel’s Ontological Proof

Johan E. Gustafsson

Kurt Gödel’s version of the Ontological Proof derives rather than assumes the crucial (yet controversial) Possibility Claim, that is, the claim that it is possible that something God-like exists. Gödel’s derivation starts off with a proof of the Possible Instantiation of the Positive, that is, the principle that, if a property is positive, it is possible that there exists something that has that property. I argue that Gödel’s proof of this principle relies on some implausible axiological assumptions. Nevertheless, I present a proof of the Possible Instantiation of the Positive which only relies on plausible axiological principles. Nonetheless, Gödel’s derivation of the Possibility Claim also needs a substantial axiological assumption, which is still open to doubt.

In its classic modal form, the Ontological Proof runs as follows:

The Possibility Claim
It is possible that something God-like exists.

The Necessity Claim
That something God-like exists strictly entails that it is necessary that something God-like exists.

Therefore, something God-like exists.

Given some standard principles of modal logic, the proof is valid. Yet the Possibility Claim is questionable. Maybe the properties needed to be God-like are inconsistent. And, if they are, it’s impossible that something God-like exists.

Notably, Kurt Gödel does not assume the Possibility Claim in his version of the Ontological Proof. Rather, his proof demonstrates the possibility of there being something God-like. He does this by first proving the following principle, which—in addition to any potential theistic implications—is interesting in its own right:

---

* I would be grateful for any thoughts or comments on this paper, which can be sent to me at johan.eric.gustafsson@gmail.com.


2 More precisely, the proof is valid given normal modal logic and the Brouwerian principle $q \supset \Box \Box q$.

3 In this respect, Gödel is following Leibniz (1969: 167–68).
The Possible Instantiation of the Positive

If a property is positive, then it is possible that there exists something that has that property.

A positive property, in Gödel’s terms, is a property that is ‘positive in the moral aesthetic sense’. Gödel (1995b: 435) also suggests that positive properties can be interpreted as perfective properties, that is, “purely good”-making properties. A positive property in this alternative sense is, presumably, a property such that it entails some basic good-making properties and no other properties except those entailed by the basic good-making properties.

Given the Possible Instantiation of the Positive and the further premise that the property of being God-like is positive, we have that it is possible that there exists something God-like.

There are two versions of Gödel’s Ontological Proof. One version is contained in two pages of handwritten notes by Gödel from February 1970, and a more elaborate version is contained in three pages written by Dana Scott, who discussed the proof with Gödel that same month (Adams 1995: 388).

In Scott’s version, the proof that it’s possible that there exists something God-like runs as follows, where \( P(\phi) \) means that property \( \phi \) is positive:

\[
(1) \quad P(\neg \phi) \equiv \neg P(\phi).
\]

\[
(2) \quad P(\phi) \land \Box \forall x (\phi(x) \supset \psi(x)) \supset P(\psi).
\]

If we adopt (1) and (2) as axioms, we can derive the Possible Instantiation of the Positive, which can be stated formally as follows (Scott 1987: 257 and Sobel 2004: 120):

---


5 It’s not obvious how the overall value-making of a non-basic property should be calculated based on the the properties it entails. See Carlson 1997 for a discussion of how to calculate the intrinsic value of non-basic states of affairs. The calculation for properties is probably analogous to the one for states of affairs.

6 Scott 1987: 257. We adopt the convention that, if \( \phi \) and \( \psi \) are properties, then \( \neg \phi \) is short for \( \lambda x (\neg (\phi(x))) \), \( \phi \lor \psi \) is short for \( \lambda x (\phi(x) \lor \psi(x)) \), and \( \phi \land \psi \) is short for \( \lambda x (\phi(x) \land \psi(x)) \), where \( \lambda x (f(x)) \) is a property an individual \( y \) has in virtue of being such that \( f(y) \). For example, \( \lambda x (x = x)(y) \) states that \( y \) has the property \( \lambda x (x = x) \), the property of being self identical, which \( y \) has in virtue of being such that \( y = y \). See Carnap 1947: 3.
Gödel's Ontological Proof then proceeds with the axiom that the property of being God-like is positive.\(^7\) That is, letting \(G\) be the property of being God-like,

\[(4) \quad P(G).\]

Finally, from (3) and (4), we have

\[(5) \quad \Diamond \exists x \; G(x).\]

Hence the idea is to derive the Possible Instantiation of the Positive from axiological principles such as (1) and (2). Then, given the Possible Instantiation of the Positive and the premise that the property of being God-like is positive, we have that it's possible that there exists something God-like.

In this paper, I shall argue that the assumptions in Gödel's proof of the Possible Instantiation of the Positive—that is, assumptions (1) and (2)—are implausible and, moreover, that the alternative proofs in the literature also rely on implausible assumptions. Nevertheless, I shall present a new proof, which only relies on plausible axiological principles.\(^8\) Finally, I argue that there is still room for doubt about the substantial axiological assumption in Gödel's proof of the Possibility Claim. It is not clear why we should accept premise (4).

1. Some objections, revisions and further objections

C. Anthony Anderson objects to the right-to-left direction of (1). Let \(M\) be the property of being male. This property seems indifferent in the sense that we seem to have both \(\neg P(M)\) and \(\neg P(\neg M)\), which violates (1).\(^9\)

Anderson (1990: 295) shows that, for the purposes of the proof of Possible Instantiation of the Positive, (1) can be weakened as follows:

\[G(x) =_{df} \forall \phi (P(\phi) \supset \phi(x)).\]

See note 18 for some alternative definitions. Still, the definition of the property of being God-like won't matter for our discussion of the Possible Instantiation of the Positive.

\(^7\) The plausibility of this axiom depends, of course, on how the property of being God-like is defined. Gödel (1987: 256, 1995a: 403) defines it as follows:

\[G(x) =_{df} \forall \phi (P(\phi) \supset \phi(x)).\]

\(^8\) By plausible axiological principles, I mean axiological principles that are compelling after sustained reflection and cohere with other compelling principles in the logic of value.

\(^9\) Anderson's (1990: 295) example of an indifferent property is being such that there are stones. That example, however, could be blocked if we require that all properties have to be intrinsic, which Sobel (2004: 561 n. 20), Kovač (2003: 569) and Koons (2006: 239–40) suggest in order to avoid a modal collapse from Gödel's assumptions. Yet Sobel (2004: 562 n. 26) points out that replacing (1) with (6) is sufficient for avoiding the collapse.
This revision allows \( \neg P(M) \) & \( \neg P(\neg M) \), since it drops the problematic right-to-left direction of (1). Moreover, (6) mirrors a plausible principle in the logic of intrinsic value, namely, the principle that, if a state of affairs is intrinsically good, its negation is not intrinsically good (Chisholm and Sosa 1966: 246 and Åqvist 1968: 260). And the combination of (2) and (6) entails (3).\(^{10}\)

Petr Hájek (2002: 150), however, objects to (2). Let \( D \) be the property of being Devil-like. From (2) and \( P(G) \), we then have that \( P(G \lor D) \). While \( G \) is (arguably) positive, \( G \lor D \) is not. There seems to be no reason to regard a disjunction of a positive property and a negative property as positive rather than negative.\(^{11}\) Moreover, it seems that tautological disjunctive properties, like \( G \lor \neg G \), are not positive even if one of their disjuncts is positive.\(^{12}\) Furthermore, it seems that the logic of negativity

\[
(6) \quad P(\phi) \Rightarrow \neg P(\neg \phi).\]

Yet (1) does not seem to fare better than (1). With (1), we get much the same problems as we get with (1). Being necessarily male doesn’t seem positive, and being not necessarily male doesn’t seem positive. And then we have that both \( \neg P(M) \) and \( \neg P(\neg M) \), which contradicts (1).

\(^{10}\) Anderson 1990: 292, 296. In a later joint work, however, Anderson and Gettings (1996: 169) instead suggest replacing (1) and (6) with the following axiom:

\[
(1) \quad \neg \left(P(\square \phi) \equiv P(\neg \square \phi)\right).\]

\(^{11}\) Sobel 2004: 122. Kovač (2003: 581) tries to rebut this objection, pointing out that this implication of Gödel's system can be justified in the system itself, since, in that system, positivity is logically (and also ontologically) stronger than negativity in the following sense: Positivity, unlike negativity, is necessarily exemplified (since, in Gödel’s system, there is a possible world where only a God-like being exist and no possible world where no God-like being exists). But to rely on the necessary existence of something God-like in the defence of a premise in an argument for the existence of something God-like is to assume the point at issue. Kovač (2003: 581) also points out that, if being Devil-like is to not have any positive qualities, then a Devil-like being couldn’t possibly exist if, as Gödel (1987: 257, 1995a: 403) maintains, self-identity is positive. This reply misses the mark: The objection that \( G \lor D \) does not seem positive even if \( G \) is positive works with any bad property in the place of \( D \). So, even if there couldn’t possibly be something Devil-like, we could still raise the objection as long as there is some negative property that something could possibly have, for example, being evil or being irrational. And having those properties is compatible with having the property of being self-identical. It may be objected that, if being self-identical is positive, then every negative property entails at least one positive property; and, plausibly, it need not be the case that every positive property entails a negative property. But, first, note that (even if we grant Gödel's claim that self-identity is positive) this asymmetry seems to have little relevance to how disjunctive properties should be evaluated given the value of their disjuncts. Second, Gödel's (1987: 257, 1995a: 403) only argument that self-identity is positive is that it follows from (2). Hence Kovač's second rebuttal also assumes the point at issue, namely, whether we should accept (2).

\(^{12}\) Sobel 2004: 120, 2006a: 406–7, 2006b: 286. Gödel (1987: 257, 1995a: 403) seems to have welcomed this result, noting in his proof that being self-identical is positive. Sobel (2006a: 398 n. 1) points out that, given (1) and (2), all tautological properties are positive.
should be analogous to that of positivity. That is, if $N(\phi)$ means that property $\phi$ is negative in the ‘moral aesthetic’ sense, then, if (2) holds, the following principle should hold too:

(7) $N(\phi) \& \Box \forall x(\phi(x) \supset \psi(x)) \supset N(\psi)$.

Being God-like seems positive, and being Devil-like seems negative. But—from $P(G)$, $N(D)$, (2) and (7)—we have both $P(G \lor D)$ and $N(G \lor D)$, which violates the principle that a property cannot be both positive and negative:

(8) $\neg(P(\phi) \& N(\phi))$.

This is a plausible principle. It mirrors a standard principle in the logic of intrinsic value, namely, the principle that no state of affairs is both intrinsically good and intrinsically bad (Chisholm and Sosa 1966: 248).

Hájek (2002: 156) puts forwards the following replacement for (1), (2) and (6):

(9) $P(\phi) \& \Box \forall x(\phi(x) \supset \psi(x)) \supset \neg P(\neg \psi)$.

With (9) rather than (2), we no longer have $P(G \lor D)$ from $P(G)$. So (9) avoids the above problems. Moreover, (9) entails (3) by itself (Hájek 2002: 156). Thus (9) is sufficient to derive (3) and avoids the previous problems.

But (9) is implausible. It faces a variation of the previous counterexamples—or, at least, it does so on the good-making interpretation of positivity. Let $K$ be the property of being all knowing. And consider the conjunctive properties $K \& M$ and $K \& \neg M$, that is, the property of being all knowing and male and the property of being all knowing and not male. Since $K$ seems positive and each one of $M$ and $\neg M$ seems neutral, each of the conjunctive properties should be overall positive. That is,

(10) $P(K \& M)$

and

(11) $P(K \& \neg M)$.

Since $K \& M$ and $K \& \neg M$ are mutually exclusive, we have

(12) $\Box \forall x(\lambda y(K(y) \& M(y))(x) \supset \neg \lambda y(K(y) \& \neg M(y))(x))$.

Then—from (9), (10) and (12)—we have

(13) $\neg P(\neg \neg(K \& \neg M))$.

---

Plausibly, if a property is positive, then the complement of the complement of that property is also positive—that is,

(14) \( P(\phi) \supset P(\neg \neg \phi) \).

Finally, from (13) and (14), we have

(15) \( \neg P(K \& \neg M) \),

which contradicts (11).

It may be objected that it’s not obvious why we should accept both (10) and (11)—especially on the “purely good”-making interpretation of positivity. Perhaps we shouldn’t grant that conjunctive properties with a positive and a neutral conjunct are positive. Moreover, note that, if we accept both (10) and (11), it follows directly that there is no God-like being with every positive property.

Yet there is a related, more general worry about (9). From (9), we have that any two properties are jointly consistent. So, if the joint consistency of some of the positive properties needed in order to be God-like is in doubt, then (9) assumes the point at issue to some extent. If we have no reason to accept that any two positive properties are jointly consistent, we have no reason to accept (9).

2. A cogent proof of the Possible Instantiation of the Positive

As we have seen, earlier proofs of the Possible Instantiation of the Positive rely on implausible assumptions. There is, however, a way to derive the Possible Instantiation of the Positive from plausible axiological principles.

First, consider the axiom

(16) \( \neg P(\lambda y(y \neq y)) \).

That is, the contradictory property of being self-different is not positive. This is a compelling axiom. It mirrors a plausible principle in the logic of intrinsic value, namely, that contradictory states of affairs are not intrinsically good.\(^4\) Gödel (1987: 257, 1995a: 403–4) mentions (16) in his proof but does not adopt it as an axiom, deriving it instead from (2).

\(^4\) von Wright 1972: 163–64 and Hansson 2001: 119. Halldén (1957: 41) maintains that the value of tautologies and contradictions should not be determined by logical principles. His idea is that logic should be impartial with regard to the evaluation of specific states of affairs or properties. (See also Åqvist 1968: 268 for a similar view.) But, even if (16) were a substantial axiological claim, rather than a formal axiological principle, it would still be implausible that contradictions are intrinsically good or positive. And it is implausible for formal reasons: contradictions entail everything; so they do not rate the world a plus any more than they rate it a minus. For every good or bad thing that contradictions entail, they also entail the complement. Contradictions are symmetrical in their relation to the good and the bad, hence they cannot plausibly be intrinsically good or positive.
Second, consider the axiom

\( \Box \forall x (\phi(x) \equiv \psi(x)) \supset (P(\phi) \equiv P(\psi)). \)

The idea, following Gödel (1995b: 433), is that, if properties mutually entail each other, then they are alike in positivity. Like (16), this axiom mirrors a plausible principle in the logic of intrinsic value, namely, that equivalent states of affairs have the same intrinsic value (Rescher 1966: 58 and Åqvist 1968: 259). If two properties mutually entail each other, any goodness and badness that follows necessarily from one of them also follows necessarily from the other. So the properties should have the same necessary advantages and disadvantages and should be alike in intrinsic positivity.

To see that we can derive (3) from (16) and (17), assume, for proof by contradiction, the negation of (3):

\( \neg \forall \phi (P(\phi) \supset \Diamond \exists x \phi(x)). \)

From (18), we have

\( \exists \phi \neg (P(\phi) \supset \Diamond \exists x \phi(x)). \)

From (19), we have, by existential instantiation,

\( \neg (P(\phi') \supset \Diamond \exists x \phi'(x)). \)

Then, from (20), we have

\( P(\phi') \)

and

\( \neg \Diamond \exists x \phi'(x). \)

From (22), we have, by the standard duality definitions and normal modal logic,

\( \Box \forall x \neg \phi'(x). \)

From (23), we have, by normal modal logic,

\( \Box \forall x (\phi'(x) \supset \lambda y (y = y)(x)). \)

And, since self-difference is logically impossible, we have

\( \Box \forall x \neg \lambda y (y = y)(x). \)

From (25), we have, by normal modal logic,

\( \Box \forall x (\lambda y (y = y)(x) \supset \phi'(x)). \)
From (24) and (26), we have, by normal modal logic,

\[ \Box \forall x (\phi'(x) \equiv \lambda y (y \neq y)(x)) \].

From (17) and (27), we have

\[ P(\phi') \equiv P(\lambda y (y \neq y)). \]

And, from (16) and (28), we have

\[ \neg P(\phi'). \]

Finally, from (21) and (29), we have

\[ P(\phi') \& \neg P(\phi'). \]

We have derived a contradiction from assumption (18), that is, the negation of (3). So we can conclude

\[ P(\phi) \supset \Diamond \exists x \phi(x). \]

We have that (16) and (17) together entail the Possible Instantiation of the Positive.\(^{15}\) Hence the Possible Instantiation of the Positive can be derived from plausible axiological principles.\(^{16}\)

\(^{15}\) We can revise the proof so that it allows for a variable domain of individuals. For simplicity, we can keep the fixed-domain quantifiers and deal with variable domains with an existence predicate. Let \( E \) be a predicate applied to individuals, with \( E(x) \) read as 'x actually exists'. Following Fitting and Mendelsohn (1998: 106) and Fitting (2002: 90), we adopt the following definitions:

\[ \forall^E x \phi =_{df} \forall x (E(x) \supset \phi). \]
\[ \exists^E x \phi =_{df} \exists x (E(x) \& \phi). \]

The proof works equally well if we replace ‘\( \forall x \)' with ‘\( \forall^E x \)' and ‘\( \exists x \)' with ‘\( \exists^E x \)' throughout. That is, we can derive

\[ (II) \ P(\phi) \supset \Box \exists^E x \phi(x) \]

from (16) and

\[ (III) \ \Box \forall^E x (\phi(x) \equiv \psi(x)) \supset (P(\phi) \equiv P(\psi)) \]

by the same line of argument.

\(^{16}\) Consider the following negative analogues of (16) and (17):

\[ (IV) \ \neg N(\lambda y (y \neq y)) \].
\[ (V) \ \Box \forall^E x (\phi(x) \equiv \psi(x)) \supset (N(\phi) \equiv N(\psi)). \]

These principles also seem plausible. Note that, here, 'negative' should be understood as negative in 'the moral aesthetic sense'. I am not denying that \( \lambda y (y \neq y) \) is negative in the sense that it is a negation of the property being self-identical. Perhaps (IV) still seems less compelling than (16); see, however, the argument in note 14. From (IV) and (V), we can derive the following in the same way as we derived (3) from (16) and (17), changing what needs to be changed:
3. The necessity part

After the possibility part, Gödel’s Ontological Proof proceeds with a derivation of the Necessity Claim, that is,

\[(31) \ \Box(\exists x G(x) \supset \Box \exists x G(x)).\]

His derivation relies on one further axiological principle, namely,

\[(32) \ \mathcal{P}(\phi) \supset \Box \mathcal{P}(\phi).\]

This principle is plausible. As Gödel notes, (32) seems to ‘follow from the nature of the property’, that is, the property of positivity. At least, (32) seems to follow if the positivity is intrinsic positivity (Moore 1922: 260–61 and Sobel 1987: 244). Which is what Gödel (1987: 257, 1995a: 404) seems to have in mind in his proof, adding the note that ‘[p]ositive means positive in the moral aesthetic sense (independently of the accidental structure of the world).’

Yet Gödel’s derivation of the Necessity Claim also relies on some of the implausible axiological principles we have rejected. Hájek, however, proposes some emended proofs of the Necessity Claim, which do not rely on any axiological principles. Hence, given the proposed patch to the

\[(VI) \ \mathcal{N}(\phi) \supset \Box \exists x \phi(x).\]

Hence we also have

**The Possible Instantiation of the Negative**

If a property is negative, then it is possible that there exists something that has that property.

The neutrality analogue of (16) looks less plausible, however. So we cannot plausibly prove in this manner that, for any neutral property, it is possible that there is something that has that property.

17 Gödel (1987: 256, 1995a: 403) also proposes

\[(VII) \ \sim \mathcal{P}(\phi) \supset \Box \sim \mathcal{P}(\phi),\]

which is, likewise, plausible given that the relevant positivity is intrinsic.

18 Hájek 1996: 128, 2002: 156. Hájek (2002: 159) also provides a proof that allows for variable domains. Although he doesn’t highlight this fact, Hájek only relies on (9) in his proof of the Possibility Claim and not in his proof of the Necessity Claim. Note that, although Hájek’s proofs of the Necessity Claim do not rely on any axiological principles, they do assume a substantial axiological premise, namely, premise (4)—that is, \(\mathcal{P}(G)\). Hájek makes this assumption under different definitions of \(G\) in the different proofs—in Hájek 1996: 128:

\[G(x) =_{df} \forall \phi(\mathcal{P}(\phi) \equiv \Box \phi(x)),\]

in Hájek 2002: 156:

\[G(x) =_{df} \forall \phi(\Box \phi(x) \equiv \exists \psi(\mathcal{P}(\psi) \& \Box \forall y(\psi(y) \supset \phi(y))))\]
possibility part, Gödel’s Ontological Proof does not need any implausible axiological principles. So the soundness of the proof depends on its substantial axiological premise, that is, premise (4): that being God-like (on the relevant definition) is positive. Unless one finds this premise independently plausible, however, one might worry that the plausibility of (4) depends (in the light of the Possible Instantiation of the Positive) on the consistency of the positive properties needed in order to be God-like.¹⁹ And, if it does, it seems that Gödel’s derivation of the Possibility Claim achieves very little.²⁰

References


and, in Hájek 2002: 159:

\[ G(x) =_{df} \forall \phi \left( \square \phi(x) \equiv \exists \psi \left( P(\psi) \& \exists y \left( \psi(y) \supset \phi(y) \right) \right) \right). \]

¹⁹ In the version in Gödel’s (1987: 256, 1995a: 403) hand, the proof’s substantial axiological premise is not (4) but the premise that having necessary existence as an essence is positive. In that version, the Possibility Claim is derived with the help of the following principle:

(VIII) \[ P(\phi) \& P(\psi) \supset P(\phi \& \psi). \]

But (VIII) is only plausible—and only consistent with (16) and (17)—as long as all positive properties are consistent. Since (VIII) is only plausible if all positive properties are consistent, it seems to assume the point at issue in an argument for the possibility of a God-like being. Moreover, note that, in the version of the proof in Scott’s hand, (4) is motivated by (VIII).

²⁰ I wish to thank Christoph Benzmüller, Erik Carlson, David Efird, Krister Bykvist, Ariel Gonzalez, Barry Lee, Daniel Molto, Paul Noordhof, Jonas Olson, Włodek Rabinowicz, Tom Stoneham and Rob Trueman for valuable comments.


