The Inconsistency in Gödel's Ontological Argument: A Success Story for AI in Metaphysics

Christoph Benzmüller*
Freie Universität Berlin & Stanford University
c.benzmueller@gmail.com

Bruno Woltzenlogel Paleo
Australian National University
bruno.wp@gmail.com

Abstract

This paper discusses the discovery of the inconsistency in Gödel's ontological argument as a success story for artificial intelligence. Despite the popularity of the argument since the appearance of Gödel's manuscript in the early 1970's, the inconsistency of the axioms used in the argument remained unnoticed until 2013, when it was detected automatically by the higher-order theorem prover Leo-II. Understanding and verifying the refutation generated by the prover turned out to be a time-consuming task. Its completion, as reported here, required the reconstruction of the refutation in the Isabelle proof assistant, and it also led to a novel and more efficient way of automating higher-order modal logic **S5** with a universal accessibility relation. Furthermore, the development of an improved syntactical hiding for the utilized logic embedding technique allows the refutation to be presented in a human-friendly way, suitable for non-experts in the technicalities of higher-order theorem proving. This brings us a step closer to wider adoption of logic-based artificial intelligence tools by philosophers.

1 Introduction

Without exaggeration Kurt Gödel's ontological argument for the existence of God [Gödel, 1970; Scott, 1972] is amongst the most discussed formal proofs in modern literature. A rich body of publications – including very recent ones – present, discuss, assess, criticize, modify and improve Gödel's original work (see e.g. Sobel [2004] and Oppy [2015] and the references therein). In philosophy lectures at universities the argument is regularly presented as a masterpiece argument in metaphysics. Since 2013, when Benzmüller and Woltzenlogel-Paleo [2013a; 2014] first reported their successful initial computer-assisted analysis of Gödel's proof and Scott's variant, their work has received a media repercussion on a global scale[1], and numerous bloggers commented on the proof [Fuhrmann, 2016].

The in-depth analysis presented here substantially extends previous computer-assisted studies of Gödel's ontological argument. Similarly to the related work [Benzmüller and Woltzenlogel-Paleo, 2013a; 2014] the analysis has been conducted with automated theorem provers for classical higher-order logic (HOL; cf. [Andrews, 2014] and the references therein), even though Gödel's proof is actually formulated in higher-order *modal* logic (HOML; cf. [Muskens, 2006] and the references therein). To bridge between the two logics we utilise and further improve the logic embedding approach [Benzmüller and Paulson, 2013; Benzmüller and Woltzenlogel-Paleo, 2014], which has already been employed successfully in preceding related work.

The main novel contribution reported in this paper is a detailed analysis (in various modal logics) of the inconsistency of Gödel's original version of the axioms used in his manuscript [1970]. The extraction, reconstruction and verification of an informal, human intuitive argument has been an open problem since the first detection of this inconsistency by Benzmüller and Woltzenlogel-Paleo [2014] with the Leo-II prover. The verified refutation (discussed in §4) displays a surprisingly accessible explanation of the inconsistency, which is philosophically profound and never presented in the literature. The detection of this inconsistency in combination with the work reported here thus demonstrates that artificial intelligence systems – particularly higher-order automated theorem provers – are capable of assisting in the discovery and elucidation of *new* and philosophically relevant knowledge.

On the technical side, the quest for constructing a compelling refutation, capable of convincing also human non-experts, led to an improvement of the syntax of the embedding of modal logics in Isabelle/HOL (as discussed in §3.2). With the new syntax, a (nearly) perfect match between the original pen and paper presentations and our encoding in Isabelle/HOL is feasible. A more user-friendly syntax, as reported here, is clearly an important prerequisite for promoting the theorem proving technology employed here to a wider community of philosophers, who are not necessarily experts in automated reasoning or HOL.

Another novel contribution reported here (in §3.1) is the implementation of an alternative embedding for the more effective modal logic $\mathbf{S5^U}$, which is based on a universal acces-

*This work was supported by the German National Research Foundation (DFG) under grants BE 2501/9-2 and BE 2501/11-1.

[1] A collection of news articles is available at https://github.com/FormalTheology/GoedelGod/blob/master/Press/LinksToNews.md

sibility relation. Our experiments have shown that the new embedding is more efficient, as the following two previously open problems can now be solved:

- Automatically proving the final theorem T3 (*Necessarily, there exists God*), directly from Scott's [1972] (consistent) axioms alone, without relying on the argument's intermediate argumentation steps (i.e., lemmata).
- Automatically verifying, in Isabelle/HOL, the proof of the modal collapse [Sobel, 1987], which is one of the most strongly criticized logical consequences of the argument's axioms.

1.1 Related Work

First successful applications of theorem proving technology in metaphysics were reported by Fitelson, Oppenheimer and Zalta [2007; 2011], who coined the term *Computational Metaphysics* for this new research area and employed the first-order Prover9 [McCune, 2010] in their experiments. Later on, Rushby [2013] used the proof assistant PVS [Owre *et al.*, 1992]. Common to both works is a significant amount of proof-hand-coding work as well as their focus on a non-modal formalization of St. Anselm's [1078] simpler and older ontological argument. In contrast, the greater complexity of Gödel's argument requires the formalization and automation of variants of *higher-order* and *modal* logics.

2 A Brief History of the Argument

St. Anselm's ontological argument [Anselm, 1078] can be regarded as the ancestor of modern ontological arguments such as Gödel's. In the millenium between Anselm and Gödel, many philosophers modified and arguably improved Anselm's argument. Of particular importance to Gödel was the work of Leibniz [Adams, 1995]. Although Gödel's notion of positive property is not exactly the same as Leibniz's notion of perfection, Gödel's manuscript (Fig. 6a) can be considered a translation of Leibniz's presentation of the argument into modern modal logic. Gödel discussed his manuscript with Scott, who shared a slightly different version with a larger public. Scott's version of the axioms and definitions, formalized in Isabelle, is shown in Fig. 1. The main difference to Gödel's version is an extra conjunct in the definition of *essence* (*ess*). Gödel's different definition of essence can be seen either in his manuscript (Fig. 6a) or, in more modern notation, in the Isabelle formalization shown in Fig. 5. For Scott, an essential property of an individual must be possessed by him/her. For Gödel, this is not required.

Gödel's omission has been considered inessential and merely an oversight by many. For instance, Hazen [1998, p.365] states that "Gödel left this clause out [...] but this appears to have been an oversight". For more than four decades, its serious consequences remained unnoticed, despite numerous analysis and criticisms of the argument. Especially since the discovery by Sobel [1987] that modal collapse (MC)[2] is entailed by Gödel's (or also Scott's) axioms,

[2]The modal collapse, $\phi \rightarrow \Box\phi$, states that contingent truth implies necessary truth; it can be interpreted as *everything is predetermined* or even *there is no free will*.

Scott_S5U.thy

```
theory Scott_S5U imports QML_S5U
begin
 consts P :: "(μ⇒σ)⇒σ"
 axiomatization where
  A1a: "⌊∀Φ. P(¬Φ) → ¬P(Φ)⌋" and
  A1b: "⌊∀Φ. ¬P(Φ) → P(¬Φ)⌋" and
  A2:  "⌊∀Φ Ψ. P(Φ) ∧ □(∀x. Φ(x) → Ψ(x)) → P(Ψ)⌋"
 definition G where
  "G(x) = (∀Φ. P(Φ) → Φ(x))"
 axiomatization where
  A3: "⌊P(G)⌋" and
  A4: "⌊∀Φ. P(Φ) → □(P(Φ))⌋"
 definition ess (infixr "ess" 85) where
  "Φ ess x = Φ(x) ∧ (∀Ψ. Ψ(x) → □(∀y. Φ(y) → Ψ(y)))"
 definition NE where
  "NE(x) = (∀Φ. Φ ess x → □(∃ Φ))"
 axiomatization where
  A5:  "⌊P(NE)⌋"

 theorem T3: "⌊□ (∃ G)⌋" -- {* LEO-II proves T3 in 2,5sec *}
  sledgehammer [provers = remote_leo2]
  by (metis (lifting, full_types)
     A1a A1b A2 A3 A4 A5 G_def NE_def ess_def)

 lemma True nitpick [satisfy,user_axioms,expect=genuine] oops
 -- {* Consistency is confirmed by Nitpick *}

 theorem T2: "⌊∀x. G(x) → G ess x⌋"
  sledgehammer [provers = remote_leo2]
  by (metis A1b A4 G_def ess_def)

 lemma MC: "⌊∀Φ. Φ → (□ Φ)⌋"  -- {* Modal Collapse *}
  sledgehammer [provers = remote_satallax, timeout=600]
  by (meson T2 T3 ess_def)
end
```

Figure 1: Full Automation of T3 in $\mathbf{S5^U}$; Consistency of Scott's Axioms; Automatic Verification of Modal Collapse

several variants have been proposed [Anderson, 1990; Anderson and Gettings, 1996; Hájek, 1996; 2001; Hájek, 2002; Bjørdal, 1999] attempting to avoid the modal collapse. Many of these variants omit the crucial conjunct in the definition of essence as well.[3] Opponents of the argument (e.g. Oppy [1996, p.226–227; 2000, p.364; 2008, p.1068]) have also proposed parodies and other criticisms, referring to variants where the conjunct is omitted.

However, as explained here, the extra conjunct is in fact crucial. Without it, Gödel's original axioms are inconsistent. With it, Scott's axioms are consistent (cf. Fig. 1 where the model finder Nitpick [Blanchette and Nipkow, 2010] confirms consistency).[4]

3 Automating HOML in HOL

Logic textbooks commonly utilize higher-order logic in an informal/semi-formal way as a meta-language to introduce the syntax and the semantics of object logics of interest, in which reasoning problems in concrete application domains can be modeled and solved with pen and paper. In fact, this

[3]As these variants also change other axioms, on which the inconsistency of Gödel's axioms depends, it is not necessarily the case that these variants are also inconsistent; they must be analyzed separately.

[4]In personal communication, Dana Scott confirmed that he was unaware at the time that Gödel's axioms were inconsistent.

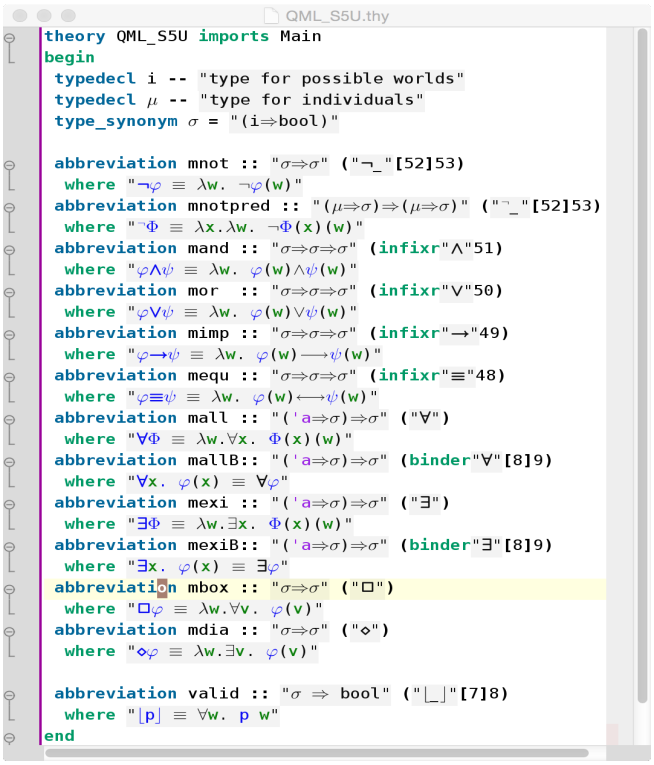


Figure 2: Improved Embedding of $\mathbf{S5^U}$

approach can also be followed on the computer (using HOL as a formal meta-language) for even very challenging object logics (such as HOML) to enable interactive and automated theorem proving with existing theorem provers for HOL.

For a computational analysis of Gödel's ontological argument, the embedding of HOMLs such as **K**, **KB** and **S5** with various domain conditions (possibilist and actualist quantification) is required. This idea has been successfully employed in related work [Benzmüller and Woltzenlogel-Paleo, 2014]. The embedding of HOML is in fact straightforward. Formulas in HOML are *lifted*, i.e., converted into predicates over worlds, which are themselves explicitly represented as terms. The logical constants of HOML are translated to HOL terms in such a way that, for instance, $\Box\varphi$ and $\Diamond\varphi$ (relative to a current world w_o) are mapped, respectively, to the HOL formulas $\forall w.(rw_0w)\rightarrow(\varphi w)$ and $\exists w.(rw_0w) \wedge (\varphi w)$. This form of embedding is precisely the well-known standard translation [Ohlbach, 1991], which is here intra-logically realized — and extended for quantifiers — in HOL by stating a set of equations defining the logical constants (Fig. 2). The resulting object logic is the HOML **K** with rigid terms and constant domains (possibilist quantifiers). Other logics (e.g. **KB**, **S5**) are embedded by adding axioms that restrict the accessibility relation r. Varying domains and actualist quantifiers can be simulated by using an existence predicate to guard the quantifiers. The embedding approach is, therefore, very flexible.

3.1 Improved Embedding

The modal logic **S5** requires that the accessibility relation be reflexive, symmetric and transitive. The usual approach to embed **S5** would be to use the standard translation for **K** described above and to state that r is an equivalence relation, e.g., by postulating the following axioms:

- Reflexivity: $\forall x.(r\ x\ x)$
- Symmetry: $\forall x.\forall y.(r\ x\ y) \rightarrow (r\ y\ x)$
- Transitivity: $\forall x.\forall y.\forall z.(r\ x\ y) \wedge (r\ y\ z) \rightarrow (r\ x\ z)$

Instead, we consider here an alternative description, that we call $\mathbf{S5^U}$, based on the following condition on r:

- Universality: $\forall x.\forall y.(r\ x\ y)$

It is important to note that $\vDash_{\mathbf{S5}} \varphi$ iff $\vDash_{\mathbf{S5^U}} \varphi$ [Blackburn *et al.*, 2001], and therefore **S5** and $\mathbf{S5^U}$ are traditionally considered to be two different descriptions of the same modal logic. Nevertheless, **S5** and $\mathbf{S5^U}$ differ in the shapes of frames they admit: $\mathbf{S5^U}$ only admits complete[5] frames, whereas **S5** admits non-complete frames as long as all their components are complete. In other words, in $\mathbf{S5^U}$ we face one single equivalence class of possible worlds, while in **S5** we may face several disconnected equivalence classes. In fact, for this reason, $\mathbf{S5^U}$ is considered as metaphysically more appropriate by some philosophers; cf. [Williamson, 2013, p. 127].

Furthermore, for $\mathbf{S5^U}$ an improved embedding is possible. Universality implies that the guarding predicates in the definitions of $\Box$ and $\Diamond$ always hold. Therefore, they can be omitted and the accessibility relation can be dispensed altogether. The modal operators can then be defined merely as:

$$\Box\varphi \equiv \lambda w.\forall v.\varphi(v) \qquad \text{and} \qquad \Diamond\varphi \equiv \lambda w.\exists v.\varphi(v)$$

The new embedding of $\mathbf{S5^U}$ in Isabelle/HOL is shown in Fig. 2. With this improved embedding, the final theorem T3 (*Necessarily, there exists God*) can be derived from Scott's consistent version of the axioms fully automatically. The fully automatic proof has been generated (in about 2.5 seconds) by the theorem prover Leo-II [Benzmüller *et al.*, 2015] and subsequently verified in the proof assistant Isabelle/HOL [Nipkow *et al.*, 2002], as shown in Fig. 1. The collaboration between the two systems has been orchestrated by Isabelle's Sledgehammer tool [Blanchette *et al.*, 2013].

With the embedding used by Benzmüller and Woltzenlogel-Paleo [2014], the provers still had to be given the intermediate theorem T2 and the corollary C in order to manage to prove T3.

Another evidence that the new embedding provides a significant performance boost is the successful automatic verification in Isabelle/HOL (with its automatic tactic Meson) of the modal collapse [Sobel, 1987], which is one of the most strongly criticized 'side-effects' of Gödel's and Scott's variants of the proof. In previous work [2014] the modal collapse has been proven by the higher-order provers Satallax [Brown, 2012] and Leo-II, but a fully automatic verification in the highly trusted Isabelle/HOL system still failed [Benzmüller and Woltzenlogel-Paleo, 2013b]. The success with the new embedding can be seen in Fig. 1.

[5] A graph is *complete* iff there is a directed edge connecting every ordered pair of vertices.

```
definition ess :: "(μ ⇒ σ) ⇒ μ ⇒ σ" (infixr "ess" 85) where
  "Φ ess x = Φ x m∧ ∀(λΨ. Ψ x m→ □ (∀(λy. Φ y m→ Ψ y)))"
```

Figure 3: Definition of Essence using Old Syntax

3.2 Improved Syntax in Isabelle

Wider adoption of HOL theorem proving technology for reasoning about and within embedded object logics, especially among non-expert users, is still hindered by the gap between the syntax used by people, when they write logical formulas with pen and paper, and the syntax used by HOL theorem provers. Even when the syntax of the underlying higher-order system is elegant (as is the case in Isabelle/HOL), the embedding of HOML into HOL may easily expose details of HOL that may be uncommon to the user, disturbing his/her experience while using the system. To illustrate this point, Fig. 3 shows how the definition of essence looked like in previous work [Benzmüller and Woltzenlogel-Paleo, 2013b], where advanced syntax-sugaring features were not used. It looks notably higher-order, and its style differs significantly from the common style seen in works on modal logics and the ontological argument. The following specific issues can be enumerated:

1. λ-abstractions, which are typically a HOL feature, appear explicitly in places where they did not need to in a pure HOML formulation (cf. Gödel's manuscript, Fig. 6a).
2. Quantifiers appear as higher-order defined constants, and not as binders. This forces the user to read (and write) formulas of the form $\forall(\lambda x.A(x))$ instead of the more common $\forall x.A(x)$.
3. The lifted modal connectives are represented by prefixing the letter "m" (e.g. $m\wedge$ and $m\rightarrow$). The prefix disturbs the user, as it constantly reminds him/her that there is something unusual about the modal connectives.
4. Higher-order parenthesis conventions for the application of a predicate to a term are used. Instead of reading $\psi(y)$, as he/she would expect, he/she has to read $(\psi\ y)$. Outside niche areas in computer science, the former syntax is more widely known than the latter.

In the embedding presented here, in Fig. 2, advanced syntax-sugaring effects provided by Isabelle were used to prevent issues as those enumerated above. The possibility to define boldface connectives allows us to drop the prefix; "`binder`" annotations enable modal quantifiers to be used in the standard binding way and reduce the need for explicit lambda abstractions; and a careful choice of priorities for infix connectives gives the parenthesis conventions that are more familiar to the user. As desired, the definition of essence in Fig. 1 is undeniably more immediately recognizable and comprehensible than the definition in Fig. 3. The embedding technique is now completely transparent to the user.

The syntax improvements described here render the computer-assisted analysis of ontological arguments accessible to a wider audience and ease the adoption of logic-based artificial intelligence tools by philosophers interested in topics where modal logic reasoning is required.

4 Intuitive Inconsistency Argument

In the typical workflow during an attempt to prove a conjecture with a theorem prover, it is customary to check the consistency of the axioms first. For if the axioms are inconsistent, anything (including the conjecture) would be trivially derivable in classical logic (*ex falso quodlibet*). Surprisingly, when this routine check was performed on Gödel's axioms [Benzmüller and Woltzenlogel-Paleo, 2014], the Leo-II prover claimed that the axioms were inconsistent. Unfortunately, the refutation generated by Leo-II was barely human-readable. The text file was 153 lines[6] long and used machine-oriented calculus (higher-order resolution [Sultana and Benzmüller, 2013]) and syntax (TPTP THF [Sutcliffe and Benzmüller, 2010]). Part of the file is displayed in Fig. 4.

Although Leo-II's resolution refutation is not easy to read for humans, it did contain relevant hints to the importance of the empty property $\lambda x.\bot$ (also denoted $\emptyset$, as in HOL it is customary to think of unary predicates as sets). Note that the terms for the empty property[7] ($\lambda x.\bot$) and for the property of self-difference ($\lambda x.x \neq x$) have identical denotations in a logic setting with full functional and Boolean extensionality as given here. Nevertheless, some philosophers[8] may actually prefer the use of self-difference over the empty property in the analysis below. However, for the proof to go through it is irrelevant which notion we use and the reader may simply replace the empty property by self-difference.

4.1 Informal Argument

Based on the hints found in Leo-II's refutation, we conceived the following informal explanation for the inconsistency of Gödel's axioms:

1. From Gödel's definition of essence ($\phi\ ess\ x \leftrightarrow \forall\psi(\psi(x)\rightarrow\Box\forall y(\phi(y)\rightarrow\psi(y))))$ it follows that the empty property (or self-difference) is an essence of every individual (**Empty Essence Lemma**):
$$\forall x\ (\emptyset\ ess\ x)$$
2. From theorem T1 (*Positive properties are possibly exemplified*: $\forall\phi[P(\phi)\rightarrow\Diamond\exists x\phi(x)]$) and axiom A5 ("necessary existence" is a positive property: $P(NE)$), it follows that NE is possibly exemplified:
$$\Diamond\exists x[NE(x)]$$
3. Expanding the definition of "necessary existence" ($NE(x) \equiv \forall\phi[\phi\ ess\ x\rightarrow\Box\exists y\phi(y)]$), the following is obtained:
$$\Diamond\exists x[\forall\varphi[\varphi\ ess\ x\rightarrow\Box\exists y[\varphi(y)]]]$$
4. The sentence above holds for all φ and thus, in particular, for the empty property (or self-difference):
$$\Diamond\exists x[\emptyset\ ess\ x\rightarrow\Box\exists y[\emptyset(y)]]$$

[6] Long lines with an average of 184 characters per line.

[7] An additional lambda abstraction occurs in the empty property in Leo-II's proof (and also in the reconstruction in Isabelle) because the embedding approach lifts the boolean type o to $\iota\rightarrow o$.

[8] Private communication with André Fuhrmann.

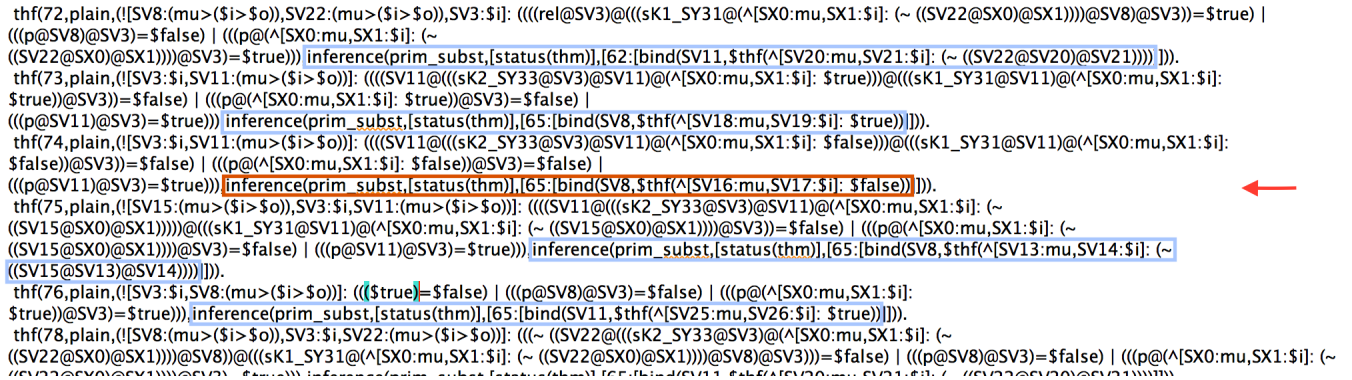


Figure 4: Lines 115–120 of Leo-II's refutation. Primitive substitutions (e.g. with the empty property) are highlighted. In the red part (see ←), property variable SV8 has been instantiated with the $\lambda SV16_{\mu}.\lambda SV17_{\iota}.\bot$, i.e., the (lifted) empty property.

5. By the Empty Essence Lemma, the antecedent of the implication above is valid. Therefore, the sentence above entails:
$$\Diamond\exists x[\Box\exists y[\emptyset(y)]]$$
6. By definition of $\emptyset$:
$$\Diamond\exists x[\Box\bot]$$
7. As the existential quantifier is binding no variable within its scope, the sentence is equi-valid with:
$$\Diamond\Box\bot$$
8. To see that the sentence above is contradictory, we may reason semantically, thinking of possible worlds. If w_0 is the arbitrary current world, the $\Diamond$ operator forces the existence of a world w accessible from w_0 such that $\Box\bot$ is true in w. But $\Box\bot$ can only be true in w, if there is no world w' accessible from w. In logics with a reflexive or symmetric accessibility relation (e.g. **KB**), it is easy to see that there must be a world w' accessible from w: either w' itself, in case of a reflexive relation, or w_0, in case of a symmetric relation. In fact, even in **K**, with no accessibility condition, there must be a world w' accessible from w. The reason is that $\Diamond\Box\bot$ should be *valid* (true in all worlds). Therefore, it is true in w as well, where the existence of an accessible world w' is forced by the $\Diamond$ operator. As a model for $\Diamond\Box\bot$ (which is a consequence of Gödel's axioms) cannot be built, Gödel's axioms are inconsistent.

Interestingly, the refutation automatically generated by Leo-II uses a symmetric accessibility relation, and thus requires the modal logic **KB**. The informal, human-constructed refutation described above, on the other hand, requires only the weaker modal logic **K**. In our experiments Leo-II (like all other HOL provers) was still too weak to automatically prove the inconsistency already in logic **K**. Hence, this remains an open problem for automated theorem provers.

4.2 Argument Reconstruction in Isabelle

To verify the correctness of the informal argument explained above, it was reconstructed in Isabelle/HOL, using Metis[9] to automate the inessential parts (cf. Fig. 5). The essential use of the Empty Essence Lemma, on the other hand, is explicitly stated, to ensure that Isabelle is reconstructing the same argument. In fact, without the help of this lemma, Metis is still not strong enough to refute Gödel's axioms.

[9]Metis, unlike external provers such as Leo-II or Satallax, constructs proofs in Isabelle's highly trusted kernel calculus.

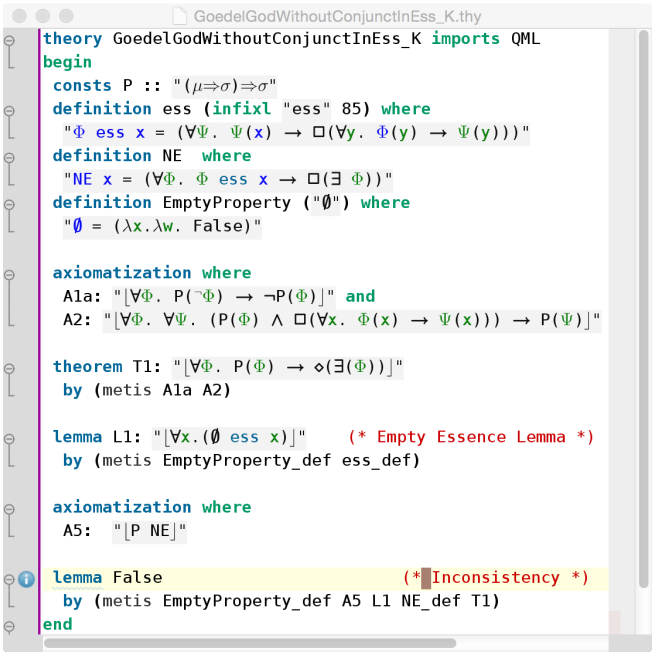
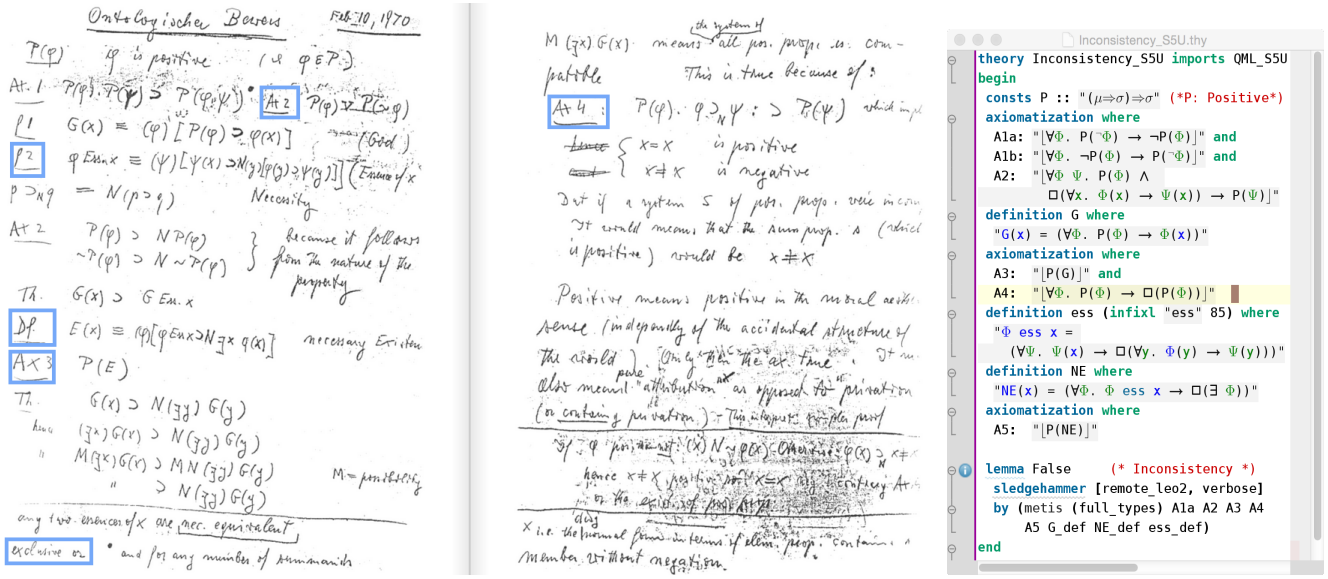


Figure 5: Inconsistency of Gödel's Axioms in HOML **K** verified in Isabelle/HOL

4.3 Mapping the Inconsistency to Gödel

The inconsistency verified in Fig. 5 follows from the definition of essence (ess), the definition of necessary existence (NE), the axioms A1a and A2 (which entail theorem T1), and axiom A5. It remains to show that these ingredients are actually present in Gödel's manuscript in Fig. 6a.

This can be easily seen: Axiom A1a in Fig. 5 is implied by Axiom Ax2 and the highlighted footnote remark in Fig. 6a. Axioms A2 and A5 in Fig. 5 correspond to Ax4 and Ax3 in



```
theory Inconsistency_S5U imports QML_S5U
begin
 consts P :: "(μ⇒σ)⇒σ" (*P: Positive*)
 axiomatization where
  A1a: "⌊∀Φ. P(¬Φ) → ¬P(Φ)⌋" and
  A1b: "⌊∀Φ. ¬P(Φ) → P(¬Φ)⌋" and
  A2:  "⌊∀Φ Ψ. P(Φ) ∧
          □(∀x. Φ(x) → Ψ(x)) → P(Ψ)⌋"
 definition G where
  "G(x) = (∀Φ. P(Φ) → Φ(x))"
 axiomatization where
  A3:  "⌊P(G)⌋" and
  A4:  "⌊∀Φ. P(Φ) → □(P(Φ))⌋"
 definition ess (infixl "ess" 85) where
  "Φ ess x =
    (∀Ψ. Ψ(x) → □(∀y. Φ(y) → Ψ(y)))"
 definition NE where
  "NE(x) = (∀Φ. Φ ess x → □(∃ Φ))"
 axiomatization where
  A5:  "⌊P(NE)⌋"

 lemma False     (* Inconsistency *)
  sledgehammer [remote_leo2, verbose]
  by (metis (full_types) A1a A2 A3 A4
      A5 G_def NE_def ess_def)
end
```

(a) Gödel's manuscript, with mutually inconsistent axioms and definitions highlighted (with permission from the Kurt Gödel Papers, Shelby White and Leon Levy Archives Center, Princeton, NJ, USA, on deposit at Princeton University)

(b) Inconsistency in HOML $\mathbf{S5^U}$

Figure 6: The inconsistency in Gödel's manuscript has been detected and verified by HOL ATPs

Fig. 6a. The definitions of essence and necessary existence are easy to identify. Therefore, the verified inconsistency from Fig. 5 does apply to Gödel's original manuscript.

4.4 Inconsistency of Gödel's Axioms in $\mathbf{S5^U}$

Isabelle/HOL's Sledgehammer tool, which orchestrates calls to external provers such as Leo-II, still fails to detect the inconsistency of Gödel's axioms in the standard embedding of **S5**, while a direct modeling of the problem in TPTP THF syntax in combination with a direct call of Leo-II succeeded. In other words, without independent experiments with no mediation through Sledgehammer, the inconsistency would not have been detected.

On the other hand (and further confirming the claims from §3.1), the reconstruction in Isabelle/HOL with the improved embedding for $\mathbf{S5^U}$ was more efficient: the inconsistency could be detected by Leo-II also when called via Sledgehammer. Moreover, the result could subsequently be verified with Metis even without the Empty Essence Lemma (cf. Fig. 6b).

5 Conclusion

The axioms and definitions in Gödel's manuscript are inconsistent; this was detected automatically by the prover Leo-II. Here we presented a rational reconstruction and verification of the inconsistency argument in Isabelle/HOL. This argument is valid in all normal HOMLs including base logic **K**.

We have also presented several technical improvements regarding the semantic embedding approach. In particular, we have achieved a nearly perfect match between pen and paper presentations in HOML and the syntax in Isabelle/HOL. As a result, the embedding of HOML in HOL is now fully transparent, more user-friendly and ready for wider adoption.

On the other hand, there is still room for many pragmatical improvements in Isabelle; just one example: in default setting, Sledgehammer does not immediately inform the user when a proof has been found and instead silently first executes a series of time-consuming proof analysis processes (e.g. its dependency minimization), before it eventually reports success. For Gödel's theorem T3 (*Necessarily, there exists God*), for example, this phase of silence takes several minutes — during which the user might actually give up on the proof attempt — even though Leo-II already reported success to Sledgehammer after 2.5 seconds.

More importantly, our work reveals a challenge for automated reasoning: the (so far partially manual) extraction of an informal argument from a formal proof. Without accompanying human-understandable explanations, the proofs generated by provers such as Leo-II or Metis, will presumably be only of limited value for philosophers, for whom intuitive arguments remain crucial for the acceptance of novel results.

Another open problem that we solved in this paper is a fully automatic proof of T3 directly from Scott's axioms. Again, this proof was contributed by Leo-II. This has become possible only after we provided a more efficient embedding for HOML $\mathbf{S5^U}$ (instead of **S5**) in HOL.

Both the automated detection of the inconsistency in Gödel's axioms and the fully automatic proof of T3 from Scott's axioms demonstrate the potential of our AI technology for philosophy: this technology is, in its current state of development, already capable of contributing novel results to metaphysics and to conduct reasoning steps at granularity-levels beyond common human capabilities.

Acknowledgments: We thank Chad Brown, who contributed to the rational reconstruction of the inconsistency argument.

References

[Adams, 1995] R.M. Adams. Introductory note to *1970. In *Kurt Gödel: Collected Works Vol. 3: Unpubl. Essays and Letters*. Oxford Univ. Press, 1995.

[Anderson and Gettings, 1996] A.C. Anderson and M. Gettings. Gödel ontological proof revisited. In *Gödel'96: Logical Foundations of Mathematics, Computer Science, and Physics: Lecture Notes in Logic 6*, pages 167–172. Springer, 1996.

[Anderson, 1990] C.A. Anderson. Some emendations of Gödel's ontological proof. *Faith and Philosophy*, 7(3), 1990.

[Andrews, 2014] P.B. Andrews. Church's type theory. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.

[Anselm, 1078] St. Anselm. Proslogion. In M. Charlesworth, editor, *St. Anselm's Proslogion*. Oxford:OUP, 1078. Republished in 1965.

[Benzmüller and Paulson, 2013] C. Benzmüller and L.C. Paulson. Quantified multimodal logics in simple type theory. *Logica Universalis*, 7(1):7–20, 2013.

[Benzmüller and Woltzenlogel-Paleo, 2013a] C. Benzmüller and B. Woltzenlogel-Paleo. Formalization, mechanization and automation of Gödel's proof of God's existence. *arXiv:1308.4526*, 2013. Preprint available as arXiv:1308.4526.

[Benzmüller and Woltzenlogel-Paleo, 2013b] C. Benzmüller and B. Woltzenlogel-Paleo. Gödel's God in Isabelle/HOL. *Archive of Formal Proofs*, 2013, 2013.

[Benzmüller and Woltzenlogel-Paleo, 2014] C. Benzmüller and B. Woltzenlogel-Paleo. Automating Gödel's ontological proof of God's existence with higher-order automated theorem provers. In *ECAI 2014*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 93 – 98. IOS Press, 2014.

[Benzmüller *et al.*, 2015] Christoph Benzmüller, Lawrence C. Paulson, Nik Sultana, and Frank Theiß. The higher-order prover LEO-II. *J. of Automated Reasoning*, 55(4):389–404, 2015.

[Bjørdal, 1999] F. Bjørdal. Understanding Gödels ontological argument. In T. Childers, editor, *The Logica Yearbook 1998*. Filosofia, 1999.

[Blackburn *et al.*, 2001] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[Blanchette and Nipkow, 2010] J.C. Blanchette and T. Nipkow. Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In *ITP 2010*, number 6172 in LNCS, pages 131–146. Springer, 2010.

[Blanchette *et al.*, 2013] J.C. Blanchette, S. Böhme, and L.C. Paulson. Extending Sledgehammer with SMT solvers. *J. of Automated Reasoning*, 51(1):109–128, 2013.

[Brown, 2012] C.E. Brown. Satallax: An automated higher-order prover. In *IJCAR 2012*, number 7364 in LNAI, pages 111 – 117. Springer, 2012.

[Fitelson and Zalta, 2007] Branden Fitelson and Edward N. Zalta. Steps toward a computational metaphysics. *J. Philosophical Logic*, 36(2):227–247, 2007.

[Fuhrmann, 2016] A. Fuhrmann. Blogging Gödel: His ontological argument in the public eye. In K. Świetorzecka, editor, *Forthcoming in Gödels Ontological Argument - History, Modifications, and Controversies*. 2016.

[Gödel, 1970] K. Gödel. *Appx. A: Notes in Kurt Gödel's Hand*, pages 144–145. In Sobel [2004], 1970.

[Hájek, 1996] P. Hájek. Magari and others on Gödels ontological proof. In A. Ursini and P. Agliano, editors, *Logic and algebra*, page 125135. Dekker, New York etc., 1996.

[Hájek, 2001] P. Hájek. Der Mathematiker und die Frage der Existenz Gottes. In B. Buldt et al., editor, *Kurt Gödel. Wahrheit und Beweisbarkeit*, pages 325–336. bv & hpt, Wien, 2001. ISBN 3-209-03835-X.

[Hájek, 2002] P. Hájek. A new small emendation of Gödel's ontological proof. *Studia Logica*, 71(2):149–164, 2002.

[Hazen, 1998] A.P. Hazen. On Gödel's ontological proof. *Australasian Journal of Philosophy*, 76:361–377, 1998.

[McCune, 2010] W. McCune. Prover9 and Mace4 (2005–2010). `http://www.cs.unm.edu/~mccune/prover9/`, 2010.

[Muskens, 2006] R. Muskens. Higher Order Modal Logic. In P. Blackburn et al., editor, *Handbook of Modal Logic*, Studies in Logic and Practical Reasoning, pages 621–653. Elsevier, Dordrecht, 2006.

[Nipkow *et al.*, 2002] T. Nipkow, L. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. Number 2283 in LNCS. Springer, 2002.

[Ohlbach, 1991] H.J. Ohlbach. Semantics-based translation methods for modal logics. *J. Log. Comput.*, 1(5):691–746, 1991.

[Oppenheimer and Zalta, 2011] P.E. Oppenheimer and E.N. Zalta. A computationally-discovered simplification of the ontological argument. *Australasian J. of Philosophy*, 89(2):333–349, 2011.

[Oppy, 1996] G. Oppy. Gödelian ontological arguments. *Analysis*, 56(4):226–230, 1996.

[Oppy, 2000] G. Oppy. Response to Gettings. *Analysis*, 60(4):363–367, 2000.

[Oppy, 2008] G. Oppy. Higher-order ontological arguments. *Philosophy Compass*, 3(5):1066–1078, 2008.

[Oppy, 2015] G. Oppy. Ontological arguments. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition, 2015.

[Owre *et al.*, 1992] S. Owre, J. M. Rushby, and N. Shankar. PVS: A prototype verification system. In Deepak Kapur, editor, *CADE*, volume 607 of *LNAI*, pages 748–752, Saratoga, NY, jun 1992. Springer.

[Rushby, 2013] J. Rushby. The ontological argument in PVS. In *Proc. of CAV Workshop "Fun With Formal Methods"*, St. Petersburg, Russia, 2013.

[Scott, 1972] D. Scott. *Appx. B: Notes in Dana Scott's Hand*, pages 145–146. In Sobel [2004], 1972.

[Sobel, 1987] J.H. Sobel. Gödel's ontological proof. In *On Being and Saying. Essays for Richard Cartwright*, pages 241–261. MIT Press, 1987.

[Sobel, 2004] J.H. Sobel. *Logic and Theism: Arguments for and Against Beliefs in God*. Cambridge U. Press, 2004.

[Sultana and Benzmüller, 2013] N. Sultana and C. Benzmüller. Understanding LEO-II's proofs. In Konstantin Korovin, Stephan Schulz, and Eugenia Ternovska, editors, *IWIL 2012*, volume 22 of *EPiC Series*, pages 33–52, Merida, Venezuela, 2013. EasyChair.

[Sutcliffe and Benzmüller, 2010] G. Sutcliffe and C. Benzmüller. Automated reasoning in higher-order logic using the TPTP THF infrastructure. *J. of Formalized Reasoning*, 3(1):1–27, 2010.

[Williamson, 2013] T. Williamson. *Modal Logic as Metaphysics*. Oxford:OUP, 2013.