MECHANIZING *PRINCIPIA LOGICO-METAPHYSICA* IN FUNCTIONAL TYPE THEORY

DANIEL KIRCHNER, CHRISTOPH BENZMÜLLER, AND EDWARD N. ZALTA

Abstract. *Principia Logico-Metaphysica* proposes a foundational logical theory for metaphysics, mathematics, and the sciences. It contains a canonical development of Abstract Object Theory [AOT], a metaphysical theory (inspired by ideas of Ernst Mally, formalized by Zalta) that differentiates between ordinary and abstract objects.

This article reports on recent work in which AOT has been successfully represented and partly automated in the proof assistant system Isabelle/HOL. Initial experiments within this framework reveal a crucial but overlooked fact: a deeply-rooted and known paradox is reintroduced in AOT when the logic of complex terms is simply adjoined to AOT's specially-formulated comprehension principle for relations. This result constitutes a new and important paradox, given how much expressive and analytic power is contributed by having the two kinds of complex terms in the system. Its discovery is the highlight of our joint project and provides strong evidence for a new kind of scientific practice in philosophy, namely, *computational metaphysics*.

Our results were made technically possible by a suitable adaptation of Benzmüller's metalogical approach to universal reasoning by semantically embedding theories in classical higher-order logic. This approach enables the fruitful reuse of state-of-the-art higher-order proof assistants, such as Isabelle/HOL, for mechanizing and experimentally exploring challenging logics and theories such as AOT. Our results also provide a fresh perspective on the question of whether relational type theory or functional type theory better serves as a foundation for logic and metaphysics.

§1. Abstract Summary. *Principia Logico-Metaphysica* (PLM) [15] is an online research monograph that contains a canonical presentation of Abstract Object Theory (AOT) [16, 17], along with motivation for, and commentary on, the theory. AOT is a foundational logical theory for metaphysics, mathematics and the sciences. It distinguishes between abstract and ordinary objects, by regimenting a distinction found in the work of the philosopher Ernst Mally [8] (though the distinction has appeared in other philosophical works).

AOT is outlined in §2. It systematizes two fundamental kinds of predication: classical exemplification for ordinary and abstract objects, and *encoding* for abstract objects. The latter is a new kind of predication that provides AOT with expressive power beyond that of quantified second-order modal logic, and this enables elegant formalizations of various metaphysical theories about different abstract objects, including the objects presupposed by mathematics and the sciences. More generally, the system offers a universal logical theory that may have a greater capability of accurately representing the contents of human thought than other foundational systems.

Independently, the use of *shallow semantical embeddings* (SSEs) of complex logical systems in classical higher-order logic (HOL) has shown great potential as a metalogical approach towards universal logical reasoning [1]. The SSE approach aims to unify logical reasoning by using HOL as a universal metalogic. Only the distinctive primitives of a target logic are defined in the metalogic in terms of their semantic interpretations, while the rest of the target system is captured by the existing infrastructure of HOL. This is why it is a *shallow* semantical embedding. For example, quantified modal logic can be encoded by representing propositions as sets of possible worlds and by representing the connectives, quantifiers, and modal operators as operations on those sets. In this way, the world-dependency of Kripke-style semantics can be elegantly modeled in HOL. Utilizing the powerful options for handling and hiding such definitions that are offered in modern proof assistants such as Isabelle/HOL [11], a human-friendly mechanization of even the most challenging target logics, including AOT, can thus be obtained.

AOT and the SSE approach are rather orthogonal. They have very different motivations and come with fundamentally different foundational assumptions. AOT uses a *hyperintensional second-order modal logic*, grounded on a *relational type theory*, as its foundation. It is in the tradition of Whitehead and Russell's *Principia Mathematica* [12, 9], which takes the notion of *relation* as primitive and defines the notion of *function* in terms of relations. The metalogic HOL in the SSE approach, by contrast, is fully extensional, and is defined on top of a functional type theory in the tradition of the work of Frege [7] and Church [5]. It takes the notion of (fully extensional) *function* as primitive and defines the notion of *relation* in terms of functions. These fundamentally different and, to some extent, antagonistic roots impose different requirements on the corresponding frameworks, in particular, with regard to the comprehension principles that assert the existence of relations and functions. Devising a mapping between the two formalisms is, unsurprisingly, a non-trivial, practical challenge [13].

The work reported here tackles this challenge. Further details can be found in Kirchner's M.A. thesis [10], where the SSE approach is utilized to mechanize and analyze AOT in HOL. Kirchner constructed a shallow semantical embedding of the second-order modal fragment of AOT in HOL, and this embedding was subsequently represented in the proof assistant system Isabelle/HOL (see §4). The proof assistant system enabled us to conduct experiments in the spirit of a *computational metaphysics*, with fruitful results that have helped to advance the ideas of AOT.

The inspiration for Kirchner's embedding comes from the model for AOT proposed by Peter Aczel.[1] Kirchner adapted techniques used in Benzmüller's initial attempts to embed AOT in Isabelle/HOL. An important goal of the research was to avoid *artifactual theorems*, i.e., theorems that (a) are derivable on the basis of special facts about the Aczel model that was used to embed AOT in

[1] An earlier model for AOT was proposed by Dana Scott. His model is equivalent to a special case of an Aczel model with only one *special urelement*. See below for a discussion of the Aczel model.

Isabelle/HOL, but (b) aren't theorems of AOT.[2] AOT is, in part, a body of theorems, and so care has been taken not to derive artifactual theorems about the Aczel model that are not theorems of AOT itself.

This explains why the embedding of AOT in Isabelle/HOL involves several layers of abstraction. In the Aczel model of AOT that serves as a starting point, abstract objects are modeled as sets of properties, where properties are themselves modeled as sets of urelements. Once the axioms of AOT are derived from the shallow semantical embedding of AOT in HOL, a controlled, and suitably restricted, logic layer is defined. By reconstructing the inference principles of AOT in the system that derives the axioms of AOT, only the theorems of AOT become derivable. By utilizing Isabelle/HOL's sophisticated support tools for interactive and automated proof development at this highest level of the embedding, it became straightforward to map the pen and paper proofs of PLM into corresponding, intuitive, and user-friendly proofs in Isabelle/HOL. In nearly all cases this mapping is roughly one-to-one, and in several cases the computer proofs are even shorter. In other words, the *de Bruijn factor* [14] of this work is close to 1. In addition, the layered construction of the embedding has yielded a detailed, experimental analysis in Isabelle/HOL of the underlying Aczel model and the semantic properties of AOT.

As an unexpected, but key result of this experimental study, it was discovered that if a classical logic for complex terms such as λ-expressions and definite descriptions is adjoined, without taking any special precautions, to AOT's specially-formulated comprehension principle for relations, a known paradox that had been successfully put to rest is reintroduced (see §5). Since the complex terms add significant expressive and analytic power to AOT, and play a role in many of its more interesting theorems and applications, the re-emergence of the known paradox has become a *new* paradox that has to be addressed. In the ongoing attempts to find an elegant formulation of AOT that avoids the new paradox, the computational representation in Isabelle/HOL now provides a testing infrastructure and serves as an invaluable aid for analyzing various conjectures and hypothetical solutions to the problem. This illustrates the very idea of *computational metaphysics*: humans and machines team up and split the tedious work in proportion to their cognitive and computational strengths and competencies. And, as intended, the results we achieved reconfirm the practical relevance of the SSE approach to universal logical reasoning.

Though the details of the embedding of AOT in Isabelle/HOL are developed in Kirchner [10], we discuss the core aspects of this work in the remainder of this article.

§2. The Theory of Abstract Objects. AOT draws two fundamental distinctions, one between *abstract* and *ordinary* objects, and one between two modes of predication, namely, classical *exemplification* (F^1x, or more generally, $F^nx_1 \dots x_n$) and *encoding* (xF^1). The variables $x, y, z, \dots$ range over both ordinary and abstract objects and we can distinguish claims about these two

[2] We have not yet investigated the question of whether the embedding of AOT in HOL is complete in the sense that if the representation of ϕ is provable in HOL, then ϕ is provable in AOT. However, this will be a topic of future research.

kinds of objects by using the exemplification predications $O!x$ or $A!x$ to assert, respectively, that x exemplifies *being ordinary* or x exemplifies *being abstract*. Whereas ordinary objects are characterized only by the properties they exemplify, abstract objects may be characterized by both the properties they exemplify and the properties they encode. But only the latter play a role in their identity conditions: $A!x \,\&\, A!y \rightarrow (x = y \equiv \Box\forall F(xF \equiv yF))$, i.e, abstract objects are identical if and only if they necessarily encode the same properties. The identity for ordinary objects on the other hand is classical: $O!x \,\&\, O!y \rightarrow (x = y \equiv \Box\forall F(Fx \equiv Fy))$, i.e., ordinary objects x and y are identical if and only if they necessarily exemplify the same properties. It is axiomatic that ordinary objects necessarily fail to encode properties ($O!x \rightarrow \Box\neg\exists F xF$), and so only abstract objects can be the subject of true encoding predications. For example, whereas Pinkerton (a real American detective) exemplifies being a detective and all his other properties (and doesn't encode any properties), Sherlock Holmes encodes *being a detective* (and all the other properties attributed to him in the novels), but doesn't exemplify *being a detective*. Holmes, on the other hand, intuitively exemplifies being a fictional character (but doesn't encode this property) and exemplifies any property necessarily implied by *being abstract* (e.g., he exemplifies *not having a mass*, *not having a shape*, etc.).[3]

The key axiom of AOT is the comprehension principle for abstract objects. It asserts, for every expressible condition on properties (i.e., for every expressible set of properties), that there exists an abstract object that encodes exactly the properties that satisfy the condition; formally:

$$\exists x(A!x \,\&\, \forall F(xF \equiv \phi)),$$

where ϕ is any condition on F in which x doesn't occur free. Therefore, abstract objects can be modeled as elements of the power set of properties: every abstract object uniquely corresponds to a specific set of properties.

Given this basic theory of abstract objects, AOT can elegantly define a wide variety of objects that have been postulated in philosophy or presupposed in the sciences, including Leibnizian concepts, Platonic forms, possible worlds, natural numbers, logically-defined sets, etc.

Another interesting aspect of the theory is its hyperintensionality. Relation identity is defined in terms of encoding rather than in terms of exemplification. Two properties F and G are stipulated to be identical if they are necessarily *encoded* by the same abstract objects ($F = G \equiv \Box\forall x(xF \equiv xG)$). However, the theory does not impose any restrictions on the properties encoded by a particular abstract object. For example, the fact that an abstract object encodes the property $[\lambda x\, Fx \,\&\, Gx]$ does not imply that it also encodes either the property F, or G or even $[\lambda x\, Gx \,\&\, Fx]$ (which, although extensionally equivalent to $[\lambda x\, Fx \,\&\, Gx]$, is a distinct intensional entity).

Therefore, without additional axioms, pairs of materially equivalent properties (in the exemplification sense), and even necessarily equivalent properties, are not forced to be identical. This is a key aspect of the theory that makes it possible

[3]He encodes *having a mass*, *having a shape*, etc., since these are properties attributed to him, at least implicitly, in the story. As an abstract object, however, he does *not* exemplify these properties, and so exemplifies their negations.

to represent the contents of human thought much more accurately than classical exemplification logic would allow. For instance, the properties *being a creature with a heart* and *being a creature with a kidney* may be regarded as distinct properties despite the fact that they are extensionally equivalent. And *being a barber who shaves all and only those persons who don't shave themselves* and *being a set of all those sets that aren't members of themselves* may be regarded as distinct properties, although they are necessarily equivalent (both necessarily fail to be exemplified).

A full description of the theory goes beyond the scope of this paper, but detailed descriptions are available in two books [16, 17] and various papers by Zalta. A regularly updated, online monograph titled *Principia Logico-Metaphysica* ([15]) contains the latest formulation of the theory and serves to compile, in one location, both new theorems and theorems from many of the published books and papers. The mechanization described below follows the presentation of AOT in PLM.

The complexity and versatility of AOT, as well as its philosophical ambitions, make it an ideal candidate to test the universality of the SSE approach. However, recent work [13] has posed a challenge for any embedding of AOT in functional type theory. In the next section, we briefly discuss this challenge.

§3. AOT in Functional Logic. Russell's well-known paradox in naive set theory arises by (a) considering the set of all sets that don't contain themselves, and (b) noting that this set contains itself if and only if it doesn't. A similar construction ('the Clark-Boolos paradox') is possible in naive versions of AOT (cf. [4] for details about the paradox first described by Clark [6] and reconstructed independently by Boolos [3]): assume that the term $[\lambda x\ \exists F(xF\,\&\,\neg Fx)]$ denotes a property (i.e., being an x that encodes a property that x does not exemplify); call it K. The comprehension axiom for abstract objects then ensures that there is an abstract object that encodes K and no other properties. This abstract object then exemplifies K if and only if it does not, and so involves one in a paradox.[4]

AOT undermines the paradox by restricting the matrix of λ-expressions to so-called *propositional formulas*, that is, to formulas without encoding subformulas. This way, the term $[\lambda x\ \exists F(xF\,\&\,\neg Fx)]$ is no longer well-formed and the construction of the paradox fails. Thus, AOT contains formulas, e.g., $\exists F(xF\,\&\,\neg Fx)$, that may *not* be placed within a λ-expression or otherwise converted to a term.

Whereas relational type theory allows one to have formulas that cannot be converted to terms, functional type theory does not; in functional type theory, it is assumed that every formula can be converted to a term. That is crucial to the analysis of the universal quantifier. The binding operator $\forall x$ in a formula of the form $\forall x\phi$ is represented, in functional type theory, as a function that maps the

[4] Let a be the abstract object guaranteed by object comprehension, so that we know:

$(\vartheta)\ \ \forall F(aF \equiv F = K)$

Now suppose, for reductio, Ka. Then by β-conversion, there is a property, say P, such that $aP\,\&\,\neg Pa$. Since aP, it follows by (ϑ) that $P = K$. So from $\neg Pa$ it follows that $\neg Ka$, which contradicts our reductio hypothesis. So suppose $\neg Ka$. Then by β-conversion and predicate logic, $\forall F(aF \to Fa)$. Now since $K = K$, it follows from (ϑ) that aK. Hence Ka. Contradiction.

property $[\lambda x\, \phi]$ to a truth value, namely, the function that maps $[\lambda x\, \phi]$ to The True just in case every object y in the domain is such that $[\lambda x\, \phi](y)$ holds. So in order to represent quantified AOT formulas that contain encoding subformulas, such as $\forall x \exists F(xF \,\&\, \neg Fx)$, their matrices have to be convertible to terms.[5] But, as we've seen, if $[\lambda x\, \exists F(xF \,\&\, \neg Fx)]$ were a term subject to β-conversion, AOT would yield a contradiction.[6]

Thus, it is not trivial to devise a semantical embedding that supports AOT's distinction between formulas and propositional formulas, but at the same time preserves a general theory of quantification. Another challenge has been to accurately represent the hyperintensionality of AOT: while relations in AOT are hyperintensional (i.e., necessarily equivalent relations may be distinct), functions (and relations) in HOL are fully extensional, and can not be used to represent the relations of AOT directly.

§4. Embedding AOT in Isabelle/HOL. The embedding of AOT in Isabelle/HOL overcomes these issues by constructing a modal, hyperintensional variant of the Aczel-model of AOT. Modality is represented by introducing a dependency on primitive possible worlds in the manner of Kripke semantics of modal logic. Hyperintensionality is achieved by an additional dependency on a separate domain of primitive *states*. Consequently, propositions are represented as Boolean-valued functions acting on states and possible worlds. The model also includes a domain partitioned into ordinary and *special* urelements. Properties are represented as functions mapping urelements to propositions. Whereas the ordinary objects of AOT can be represented by ordinary urelements, the abstract objects of AOT are represented as sets of properties and these sets are assigned a

[5]Note that $\forall x \exists F(xF \,\&\, \neg Fx)$ is a well-formed formula of the system, but in fact false. For example, it fails when x is ordinary, and when x is the abstract object that encodes no properties. However, the negation of this formula is true, and our question is how to *interpret* the embedded quantifier in functional type theory.

[6]Readers familiar with Isabelle/HOL might find this notation confusing. In AOT, the symbol ϕ is a metavariable that ranges over formulas which may contain free occurrences of x that can be bound by a binding operator. In Isabelle/HOL, however, a formula with a free variable would be represented as a function from individuals to truth-values, and the quantified formula would be written as $\forall x.\, \phi\, x$. Such a formula is true, if $\phi\ x$, i.e., the function application of ϕ to x holds for all x in the domain. In this scenario it is true that $\phi = (\lambda x.\phi x)$. Consequently the primitive, functional λ-expressions of Isabelle/HOL cannot be used to represent the λ-expressions of AOT, since the λ-expressions of Isabelle/HOL cannot simultaneously exclude non-propositional formulas while allowing quantified formulas with encoding subformulas.

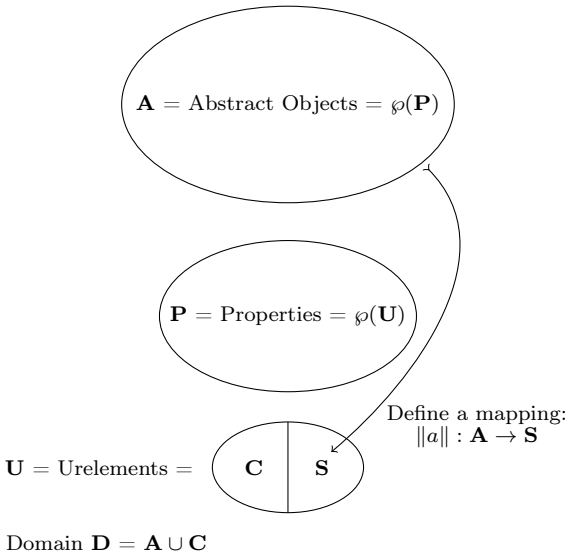

FIGURE 1. Extensional, non-modal Aczel model of AOT.

proxy among the *special* urelements (and given that there are more sets of properties than urelements, some abstract objects will be assigned the same proxy).[7] See Figure 1 for a graphical view of these domains and relationships.[8]

From this description, it becomes clear that if x is an ordinary object, then the truth conditions of an exemplification formula Px are captured by the proposition that is the result of applying the function representing the property P to the ordinary urelement representing x. If x is an abstract object, then the truth conditions of an exemplification formula Px are captured by the proposition that is the result of applying the function representing the property P to the *special* urelement that serves as the proxy of x. An encoding formula xP, by contrast, is true just in case x is an abstract object and the property P is an element of the set of properties representing x. This latter feature of the model validates the comprehension axiom for abstract objects: for every set of properties there exists a unique abstract object that encodes exactly those properties. Figure 2 shows the representation of 1-place exemplification and encoding in Isabelle/HOL.

[7]The problem that Aczel solves in the model is this: if abstract objects are represented as sets of properties, then how are we to understand the fact that in object theory, there is an object x and property F such that both xF and Fx? The encoding claim, xF, is easy: in the model, this is true if $F \in x$. However, how can a set of properties *exemplify* a property that is an element of it? We cannot model Fx as $x \in F$ without a violation of the foundation axiom. Interestingly, Aczel chose *not* to use nonwellfounded sets for his model. Instead he mapped abstract objects, modeled as sets of properties, to proxies in the domain of special urelements and set the truth conditions for Fx to the following disjunctive condition: $x \in F$ if x is ordinary and $\|x\| \in F$ (i.e., the proxy of x is an element of F), if x is abstract. We have extended Aczel's model with modality and hyperintensionality.

[8]For simplicity the graphical representation does not include the additional domains of possible worlds and intensional states, so the model shown in the graphic is extensional and non-modal.

```
lift_definition exe1::"Π₁⇒κ⇒o" ("⦇_,_⦈") is
  "λ F x s w . (proper x) ∧ F (νυ (rep x)) s w" .
lift_definition enc :: "κ⇒Π₁⇒o" ("⦃_,_⦄") is
  "λ x F s w . (proper x) ∧ case_ν (λ ω . False) (λ α . F ∈ α) (rep x)" .
```

FIGURE 2. Definition of 1-place exemplification and encoding in Isabelle/HOL.

In the definition of 1-place exemplification, which starts on the first line, exe1 is defined as a function of type $\Pi_1 \Rightarrow \kappa \Rightarrow o$ that maps 1-place relation terms (of type Π_1) and individual terms (of type κ) to propositions (type o). This function then becomes represented as a 4-argument function that maps properties, individuals, states, and worlds to a Boolean (this function, in turn, is defined by means of proper, rep and the $\nu\upsilon$ mappings).[9] In the definition of encoding, enc is defined as a function of type $\kappa \Rightarrow \Pi_1 \Rightarrow o$ that maps individual terms and 1-place relation terms to propositions. This function then becomes represented as a 4-argument function that maps individuals, properties, states, and worlds to a Boolean (this function, in turn, is defined by means of proper, rep and a case distinction on types case_ν).[10]

Since well-formed λ-expressions in AOT are required to have a propositional matrix, they correspond to functions on urelements. Given that encoding subformulas are excluded from these expressions in AOT, the only formulas that can occur in the matrix of a λ-expression are those built up from exemplification formulas. The truth conditions of these formulas are determined solely by the properties and relations of the urelements in the model.[11]

Consequently, the λ-expressions of AOT are not represented using the unrestricted primitive λ-expressions of HOL, but have a more complex semantic representation which is captured by the definition of a new class of λ-expressions in Isabelle/HOL that will represent AOT λ-expressions; the definition for the 1-place case is given in Figure 3.

```
lift_definition lambdabinder1 :: "(ν⇒o)⇒Π₁" (binder "λ" [8] 9) is
    "λ φ . λ u s w . ∃ x . νυ x = u ∧ φ x s w" .
```

FIGURE 3. Definition of AOT's λ-expressions in Isabelle/HOL.

[9]Here proper x is true, if x denotes an individual (x can also be a non-denoting definite description), rep x is the individual denoted by x (given that x denotes), and $\nu\upsilon$ is the mapping from individuals to urelements. As a result, the exemplification function maps a property term and an individual term to a proposition that is true in a given intensional state and possible world if and only if the individual term denotes and the property denoted by the property term maps the triple consisting of the urelement corresponding to the denoted individual, the given intensional state, and the given possible world, to The True. For a full description of all types, symbols and concepts involved, refer to [10].

[10]The second conjunct in the definition evaluates to False, if rep x is an ordinary object; if rep x is an abstract object, the conjunct is true if and only if the property F is contained in the abstract object.

[11]It turns out that in the October 28, 2016 version of *PLM*, there was an exception to this rule that led to the reintroduction of the Clark-Boolos paradox. We'll discuss this in a subsequent section.

It may help to explain the definition of lambdabinder1 in Figure 3. lambdabinder1 is defined as a function of type $(\nu \Rightarrow o) \Rightarrow \Pi_1$, which maps functions from individuals to propositions to 1-place relation terms. The function φ is mapped to a 1-place relation term, which is in turn represented as a Boolean-valued ternary function on urelements u, states s, and worlds w. This function evaluates to true if there exists an individual x such that both x is mapped to the urelement u under the mapping $\nu\upsilon$ and the function φ evaluates, for x, to a proposition true in s and w.

Thus, non-well-formed λ-expressions of AOT, which can't be syntactically excluded from the SSE representation, are given a non-standard semantics and this avoids, modulo the discussion below, the Clark-Boolos paradox. As a result, β-conversion for the *defined* λ-expressions holds in general for terms that are syntactically well-formed in AOT, whereas for terms that are not syntactically well-formed in AOT (but which are still part of the SSE), β-conversion is not derivable.

The model structure we've just described can represent all the terms of the target logic and can preserve hyperintensionality. Moreover, the axiom system and inference rules of AOT become derivable. Thus, the embedding makes it possible to introduce additional layers of abstraction. Given the model structure, the first layer of abstraction is the representation of the formal semantics of PLM. On the basis of that representation, the axiom system and the fundamental inference rules of PLM are derived and constitute the basis for a second layer of abstraction. Initially, this second abstraction layer solely consists of the axioms and rules of PLM itself and this makes it possible to reason directly in the target logic but independently of the underlying model structure. Thus, the second layer of abstraction avoids the derivation of artifactual theorems; the fact that the model structure validates formulas that aren't theorems of AOT is of no further consequence. And, just as importantly, the model guarantees that the system of AOT is sound.

These results are illustrated in the following figures. In Figure 4, we show the derivation of some axioms; in Figure 5, we show the derivation of the $\Diamond$ version of the Barcan formula; and in Figure 6, the derivation of the artifactual theorem $xF \leftrightarrow F \in x$ requires one to unfold the semantic definitions (this is revealed by the second line).

```
lemma pl_1[axiom]:
  "[[φ → (ψ → φ)]]"
  by axiom_meta_solver
lemma pl_2[axiom]:
  "[[(φ → (ψ → χ)) → ((φ → ψ) → (φ → χ))]]"
  by axiom_meta_solver
lemma pl_3[axiom]:
  "[[(¬φ → ¬ψ) → ((¬φ → ψ) → φ)]]"
  by axiom_meta_solver
```

FIGURE 4. Some of the axioms of AOT, derived automatically in Isabelle/HOL.

```
lemma BFs_3[PLM]:
  "[◇(∃ α. φ α) → (∃ α . ◇(φ α)) in v]"
  proof -
    have "[(∀α. □(¬(φ α))) → □(∀α. ¬(φ α)) in v]"
      using BF by metis
    hence 1: "[(¬(□(∀α. ¬(φ α)))) → (¬(∀α. □(¬(φ α)))) in v]"
      using contraposition_1 by simp
    have 2: "[◇(¬(∀α. ¬(φ α))) → (¬(∀α. □(¬(φ α)))) in v]"
      apply (PLM_subst_method "¬□(∀α . ¬(φ α))" "◇(¬(∀α. ¬(φ α)))")
      using KBasic2_2 1 by simp+
    have "[◇(¬(∀α. ¬(φ α))) → (∃ α . ¬(□(¬(φ α)))) in v]"
      apply (PLM_subst_method "¬(∀α. □(¬(φ α)))" "∃ α. ¬(□(¬(φ α)))")
       using cqt_further_2 apply metis
      using 2 by metis
    thus ?thesis
      unfolding exists_def diamond_def by auto
  qed
lemmas "BF◇" = BFs_3
```

FIGURE 5. Reasoning in the abstract layer in Isabelle/HOL. Only theorems and rules of AOT are used in derivations.

```
lemma "[{|(αν x)^P,F|} in v] ⟷ F ∈ x"
  by (simp add: meta_defs meta_aux)
```

FIGURE 6. Artifactual theorems are only provable by expanding the metalogical definitions.

Furthermore, there are other advantages to our methodology. For one thing, it is straightforward to convert statements derived within Isabelle/HOL into traditional pen and paper proofs for AOT. Thus, our approach facilitates experimental studies within the computational implementation and informs discussions about them. Moreover, the approach is suitable for conducting a deeper analysis of AOT and its model structure. The analysis led to the discovery of how a previously known paradox could easily resurface if care isn't taken in the formulation of PLM. This paradox will be sketched in the next section.

§5. Reintroduction of a Paradox. As explained in the previous section, our goal was to ensure that all of the λ-expressions of the embedding that conform to AOT's syntactic restrictions have a standard semantics. We wanted β-conversion to govern all λ-expressions with a propositional matrix.

However, as we analyzed our work, it became apparent that in the working version of PLM that served as the basis of our investigations, certain λ-expressions involving definite descriptions did not exhibit the desired behavior in the embedding. These were λ-expression in which a variable that occurred free within a definite description embedded in the λ-expression became bound by the λ. We could not verify that β-conversion was derivable for those expressions. Using the sophisticated infrastructure provided by Isabelle/HOL, it became possible to

show that β-conversion for these terms does not hold generally in an Aczel model and this suggested that there might be some problem with these expressions.

Consequently we focused our attention on the assumption that β-conversion holds for such terms in Isabelle/HOL. This assumption turned out to be inconsistent; the layered structure of the embedding made it possible to construct a proof of the inconsistency using object level reasoning at the highest level of abstraction. This way, a human-friendly proof of the paradox was reconstructed and quickly confirmed. The logic of λ-expressions and definite descriptions combines to circumvent the restriction that encoding subformulas not be allowed in λ-expressions. Indeed, the paradox turned out to be one that was previously known (the Clark-Boolos paradox mentioned earlier), but which had re-emerged through the back door (see the discussion below). This new route to a previously known paradox constituted a new paradox. The new paradox is due in part to the precise definition of *subformula*. The matrix of a λ-expression in AOT is allowed to contain encoding formulas as long as they are *nested within a definite description*. Encoding formulas so nested are not considered subformulas of the matrix and so such matrices are still considered propositional formulas. Therefore, the term $[\lambda x\, G\imath y\psi]$ is considered well-formed even if ψ contains encoding subformulas. By choosing G to be a property that is universally true (e.g. $[\lambda z\, \forall p(p \rightarrow p)]$) and letting ψ be $y = x \,\&\, \exists F\, (xF \,\&\, \neg Fx)$, one could construct a property that is extensionally equivalent to the property K described above in Section 3. This is sufficient to reconstruct the Clark-Boolos paradox.

More specifically, to see how the Clark-Boolos paradox gets reintroduced, suppose that the following λ-expression denotes a property, for any choice of G:

$$[\lambda x\, G\imath y(y{=}x \,\&\, \exists F(xF \,\&\, \neg Fx))]$$

Then if G is a universal property such that $\forall x Gx$, it can be shown that

$$\begin{array}{rcl} G\imath y(y{=}x \,\&\, \exists F(xF \,\&\, \neg Fx)) & \equiv & \exists! y(y{=}x \,\&\, \exists F(xF \,\&\, \neg Fx)) \\ & \equiv & \exists F(xF \,\&\, \neg Fx) \end{array}$$

We leave the proof as an exercise. So the matrix $G\imath y(y{=}x \,\&\, \exists F(xF \,\&\, \neg Fx))$ is equivalent to $\exists F(xF \,\&\, \neg Fx)$, when G is a universal property. Although the λ-expression built from the latter matrix was banished from AOT, a λ-expression built from the former matrix would just as easily lead to the Clark-Boolos paradox.

The discovery of the reemergence of the Clark-Boolos paradox has led not only to a modest revision of the axioms of AOT that avoids the paradox (without sacrificing any important theorems) but also to a deeper definition of *logical existence* for terms. Logical existence, i.e., the fact that a term has a denotation and is thus significant, is now defined using predication instead of using identity (typically, in free logic, the logical existence of a term is defined using identity).

§6. Final Considerations. The complexity of the target system and the multiple abstraction layers presents a challenge for the development and use of automated reasoning tools. One option for automating proofs is to use Isabelle/HOL's inbuilt reasoning tools (e.g., Sledgehammer and Nitpick) to unfold the semantical definitions used to represent AOT in HOL and to reason with the resulting statements about the model. A better option is to directly automate

the proof theory of PLM at an abstract layer, i.e., without unfolding the semantical definitions. We had two reasons for adopting this latter option: it easily avoids the problem of generating artifactual theorems and it allows for the interactive construction of complex, but human-friendly, proofs for PLM. To simplify the implementation of this option, we used the *Eisbach package* of Isabelle to define powerful proof methods for the system PLM, including a resolution prover that can automatically derive the classical propositional tautologies directly in AOT.

One interesting problem that has not yet been resolved is the one identified in Oppenheimer & Zalta [13]. As noted in Section 3, AOT has formulas that can't be converted to terms and this makes it difficult to give a general representation of AOT in functional type theory. Oppenheimer & Zalta concluded from this that relational type theory is more fundamental than functional type theory. But though the SSE embedding of AOT in Isabelle/HOL doesn't challenge this conclusion directly, it does show that the functional setting of HOL can offer a reasonably accurate representation of the reasoning that can be done in AOT. This approach addresses, at least in part, Oppenheimer & Zalta's claim, though we haven't yet addressed whether functional type theory, in the absence of abstraction layers, can generally represent systems of relational type theory in which not every formula can be converted to a term.

We've discovered that the key to the development of a sound axiomatization of the complex relation terms of AOT is to be found in the study of, and solution to, the representation of λ-expressions. With a paradox-free emendation of AOT, future research should be directed to giving an extended analysis of the faithfulness of the embedding approach we used; this would shed further light on the debate about relational and functional type theory. This study should be complemented by an analysis of the reverse direction, i.e., an embedding of the fundamental logic of HOL in the (relational) type-theoretic version of AOT. Both studies should then be carefully assessed.

In conclusion, the semantical embedding approach has been fruitfully employed to encode the logic of AOT in Isabelle/HOL. By devising and utilizing a multi-layered approach (which at the most abstract level directly mechanizes the proof-theoretic system of AOT), the issues arising for an embedding in classical higher-order logic are not too difficult to overcome. A highly complex target system based on a fundamentally different tradition of logical reasoning (relational instead of functional logic) has been represented and analyzed using the approach of shallow semantical embeddings. The power of this approach has been demonstrated by the discovery of a previously unnoticed paradox that was latent in AOT. Furthermore, the work contributes to the philosophical debate about the tension between functional type theory and relational type theory and their inter-representability, and it clearly demonstrates the merits of *shallow semantical embeddings* as a means towards universal logical reasoning.

REFERENCES

[1] Christoph Benzmüller, *Universal Reasoning, Rational Argumentation and Human-Machine Interaction*, ***CoRR***, vol. abs/1703.09620 (2017).

[2] Christoph Benzmüller and Bruno Woltzenlogel Paleo, *Automating Gödel's Ontological Proof of God's Existence with Higher-order Automated Theorem Provers*, ***ECAI 2014*** (Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors), Frontiers in Artificial Intelligence and Applications, vol. 263, IOS Press, 2014, pp. 93–98.

[3] George Boolos, *The Consistency of FregeâĂŹs Foundations of Arithmetic*, ***On being and saying*** (J. Thomson, editor), Cambridge, MA: MIT Press, spring 2017 ed., 1987.

[4] Otávio Bueno, Christopher Menzel, and Edward N. Zalta, *Worlds and Propositions Set Free*, ***Erkenntnis***, vol. 79 (2014), pp. 797–820.

[5] Alonzo Church, *A Formulation of the Simple Theory of Types*, ***The Journal of Symbolic Logic***, vol. 5 (1940), no. 2, pp. 56–68.

[6] Romane Clark, *Not Every Object of Thought has Being: A Paradox in Naive Predication Theory*, ***Noûs***, vol. 12 (1978), no. 2, pp. 181–188.

[7] Gottlob Frege, ***Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens***, Verlag von Louis Nebert, Halle, 1879.

[8] Alexander Hieke and Gerhard Zecha, *Ernst Mally*, ***The stanford encyclopedia of philosophy*** (Edward N. Zalta, editor), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[9] Andrew David Irvine, *Principia Mathematica*, ***The stanford encyclopedia of philosophy*** (Edward N. Zalta, editor), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[10] Daniel Kirchner, *Representation and Partial Automation of the Principia Logico-Metaphysica in Isabelle/HOL*, ***Archive of Formal Proofs***, (2017), `http://isa-afp.org/entries/PLM.html`, Formal proof development.

[11] Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel, ***Isabelle/HOL — a proof assistant for higher-order logic***, LNCS, vol. 2283, Springer, 2002.

[12] Alfred North Whitehead and Bertrand Russell, ***Principia Mathematica***, 2 ed., vol. 3, Cambridge University Press, Cambridge, 1913.

[13] Paul E. Oppenheimer and Edward N. Zalta, *Relations Versus Functions at the Foundations of Logic: Type-Theoretic Considerations*, ***Journal of Logic and Computation***, vol. 21 (2011), no. 2, pp. 351–374.

[14] Freek Wiedijk, *The de Bruijn factor*, `http://www.cs.ru.nl/~freek/factor/`.

[15] Edward N. Zalta, *Principia Logico-Metaphysica*, `http://mally.stanford.edu/principia.pdf`, [Draft/Excerpt; accessed: April 01, 2017].

[16] ———, ***Abstract Objects: An Introduction to Axiomatic Metaphysics***, Synthese Library, Springer, 1983.

[17] ———, ***Intensional Logic and the Metaphysics of Intentionality***, A Bradford book, MIT Press, 1988.

D. KIRCHNER
FACHBEREICH MATHEMATIK UND INFORMATIK
FREIE UNIVERSITÄT BERLIN, ARNIMALLEE 14, 14195 BERLIN, GERMANY
E-MAIL: DANIEL@EKPYRON.ORG

C. BENZMÜLLER
COMPUTER SCIENCE AND COMMUNICATIONS
UNIVERSITY OF LUXEMBOURG
2, AVENUE DE L'UNIVERSITÉ, L-4365 ESCH-SUR-ALZETTE, LUXEMBOURG
E-MAIL: CHRISTOPH.BENZMUELLER@UNI.LU
and
FACHBEREICH MATHEMATIK UND INFORMATIK
FREIE UNIVERSITÄT BERLIN
ARNIMALLEE 14, 14195 BERLIN, GERMANY
E-MAIL: C.BENZMUELLER@FU-BERLIN.DE

E. N. ZALTA
CENTER FOR THE STUDY OF LANGUAGE AND INFORMATION
STANFORD UNIVERSITY
CORDURA HALL, 210 PANAMA STREET, STANFORD, CA 94305-4115, USA
E-MAIL: ZALTA@STANFORD.EDU