Safety and Conservativity of Definitions in HOL and Isabelle/HOL

ONDŘEJ KUNČAR, Technische Universität München, Germany
ANDREI POPESCU, Middlesex University London, United Kingdom and Institute of Mathematics Simion Stoilow of the Romanian Academy, Romania

Definitions are traditionally considered to be a safe mechanism for introducing concepts on top of a logic known to be consistent. In contrast to arbitrary axioms, definitions should in principle be treatable as a form of abbreviation, and thus compiled away from the theory without losing provability. In particular, definitions should form a conservative extension of the pure logic. These properties are crucial for modern interactive theorem provers, since they ensure the consistency of the logic, as well as a valid environment for total/certified functional programming.

We prove these properties, namely, safety and conservativity, for Higher-Order Logic (HOL), a logic implemented in several mainstream theorem provers and relied upon by thousands of users. Some unique features of HOL, such as the requirement to give non-emptiness proofs when defining new types and the impossibility to unfold type definitions, make the proof of these properties, and also the very formulation of safety, nontrivial.

Our study also factors in the essential variation of HOL definitions featured by Isabelle/HOL, a popular member of the HOL-based provers family. The current work improves on recent results which showed a weaker property, consistency of Isabelle/HOL's definitions.

CCS Concepts: • **Theory of computation → Logic and verification; Higher order logic; Type structures; Interactive proof systems**;

Additional Key Words and Phrases: higher-order logic (HOL), proof theory, interactive theorem proving, type definitions, conservative extensions, Isabelle/HOL

ACM Reference Format:
Ondřej Kunčar and Andrei Popescu. 2018. Safety and Conservativity of Definitions in HOL and Isabelle/HOL. *Proc. ACM Program. Lang.* 2, POPL, Article 24 (January 2018), 31 pages. https://doi.org/10.1145/3158112

1 INTRODUCTION

Higher-Order Logic (HOL) [Pitts 1993] (recalled in Section 3 of this paper) is an important logic in the theorem proving community. It forms the basis of several interactive theorem provers (also known as proof assistants), including HOL4 [Gordon and Melham 1993; Slind and Norrish 2008], HOL Light [Harrison 1996], Isabelle/HOL [Nipkow and Klein 2014; Nipkow et al. 2002], ProofPower-HOL [Arthan 2004] and HOL Zero [Adams 2010].

In addition to supporting the development of formalized mathematics, most modern interactive theorems provers also include a functional programming language, supporting the paradigm of

Authors' addresses: Ondřej Kunčar, Fakultät für Informatik, Technische Universität München, Munich, Germany, kuncar@in.tum.de; Andrei Popescu, Department of Computer Science, Middlesex University London, London, United Kingdom , Institute of Mathematics Simion Stoilow of the Romanian Academy, Romania, uuomul@yahoo.com.

2475-1421/2018/1-ART24
https://doi.org/10.1145/3158112

total programming [Turner 2004]. For example, in provers based on type theory such as Agda [Bove et al. 2009], Coq [Bertot and Casteran 2004] and Matita [Asperti et al. 2011], totality is ensured by a global *strong normalization property*. There is a tight relationship between this property, allowing functions/programs to be reduced to a normal form by recursively unfolding all definitions and reducing all redexes, and the logical consistency of these systems.

In HOL-based provers, programming is supported by a different mechanism: All recursive datatype specifications and all recursive specifications of functions on these datatypes are translated into *nonrecursive* HOL primitives, i.e., constant and type definitions; then the recursive specifications are proved automatically as *theorems* in the logic. This scheme involves a massive background compilation and proof process (supported by tools consisting of tens of thousands of lines of code, e.g., [Blanchette et al. 2014; Krauss 2009; Melham 1989]). It ensures a high degree of trustworthiness—given that all constructions must pass through the "firewall" of HOL's minimalistic kernel. In particular, a potential bug in the compilation tools could cause correct user specifications to fail, but will not introduce logical inconsistencies unless the kernel has a bug.

In this paper, we turn our attention to the HOL kernel itself, which is the guarantor of logical consistency and certified programming in the above scheme. In spite of extensive foundational studies and the relative simplicity of the logic, the normalization process underlying the HOL kernel, i.e., the process of unfolding the HOL definitions, remains less understood than the corresponding "normalization" process in type theory, and occasionally leads to controversial design decisions and heated debates—as we are about to show, after recalling some background information.

While its ideas go back a long way (to the work of Alonzo Church [Church 1940] and beyond), HOL contains a unique blend of features proposed by Mike Gordon at the end of the eighties, inspired by practical verification needs: Its type system is the rank-one polymorphic extension of simple types, generated using the function-space constructor from two base types, bool and ind; its terms have built-in equality (from which all the usual connectives and quantifiers are derived); deduction, operating on terms of type bool called formulas, is regulated by the built-in axioms of Equality, (Hilbert) Choice and Infinity (for the type ind). In addition to this purely logical layer, which we shall refer to as *initial HOL*, users can perform constant and type declarations and definitions. Type definitions proceed by indicating a predicate on an existing type and carving out the new type from the subset satisfying the predicate. For accepting a type definition, the system requires a proof that the subset is nonempty (the predicate has a witness). This is because *HOL types are required to be nonempty*—a major design decision, with practical and theoretical ramifications [Gordon and Melham 1993; Paulson 1990]. No new axioms are accepted (more precisely, they are strongly discouraged), besides the aforementioned definitions. This minimalistic, *definitional approach* offers good protection against the accidental introduction of *inconsistency* (the possibility to prove False).

Isabelle/HOL is a notable member of the HOL family, and a maverick to some extent. It implements an essential variation of HOL, where constant definitions can be overloaded in an ad hoc manner, for different instances of their types. This flexibility forms the basis of Haskell-style type classes [Nipkow and Snelting 1991],[1] a feature that allows for lighter, suppler formalizations and should probably be credited, together with the high-level structured proof language [Wenzel 1999], the powerful automation [Paulson 2010] and the convenient user interface [Wenzel 2014], for Isabelle/HOL's wide popularity and prolificness: thousands of users in both academia and industry, a large library of formalized results [Isabelle 2016; Klein et al. 2016], major verification success stories [Esparza et al. 2013; Klein et al. 2010; Lochbihler 2010].

[1] Type classes do not require any additional extension of the logic, but are completely reduced (including at the level of proofs) to HOL with type definitions and ad hoc overloaded constants [Wenzel 1997, Section 5].

The founders of HOL have paid special attention to consistency and related properties. Andrew Pitts designed a custom notion of *standard model* [Pitts 1993], aimed at smoothly accommodating both polymorphism and type definitions. He proved that constant and type definitions are *model-theoretically conservative w.r.t. standard models*: Any standard model of a theory can be expanded to a standard model of the theory plus the definitions. This of course implies consistency of HOL with definitions. Surprisingly, the HOL founders have not looked into the more customary notion of *proof-theoretic conservativity*, which we shall simply call *conservativity*. It states that, by adding new constants and types and their definitions, nothing new can be proved in the old language. This does not follow from the model-theoretic version (because of the restriction to *standard* models, for which deduction is not complete). In fact, as we discuss below, it does not even hold in general.

In Isabelle/HOL, the foundational problem is more challenging. Here, even the *consistency* of definitions has not been fully understood until very recently (Section 2.2). The culprit is precisely the feature that contributes to Isabelle/HOL's popularity—ad hoc overloading—which has a delicate interaction with type definitions [Kunčar and Popescu 2015, Section 1].

Motivated by the desire to settle the Isabelle foundations, in early work Wenzel formulates criteria for safety of definitions in HOL-like logics [Wenzel 1997]. For a theory extension $\Theta_1 \subseteq \Theta_2$, he considers (proof-theoretic) conservativity, a property much stronger than preservation of consistency, to be a minimum requirement for deeming a theory extension truly definitional [Wenzel 1997, p.7]. In fact, he argues for an even stronger notion, *meta-safety*. Let Σ_1 and Σ_2 be the languages (signatures) of Θ_1 and Θ_2, respectively. (Thus, $\Sigma_1 \subseteq \Sigma_2$.) Meta-safety requires that, whenever a Σ_2-formula φ is deducible from Θ_2, there exists a Σ_1-formula $\varphi[\ldots, t/c, \ldots]$, obtained by replacing all the items $c \in \Sigma_2 \setminus \Sigma_1$ with some suitable Σ_1-terms t, which is deducible from Θ_1. This way, the items c can be considered to be "defined" because they can always be compiled away without losing provability. He also shows that, under appropriate well-formedness restrictions, a set of constant definitions forms a meta-safe extension.

However, as formulated, meta-safety does not apply to type definitions, because in HOL it is impossible to replace a defined type with its defining expression. In fact, Wenzel makes the following observation: *In general, type definitions in HOL are not even consistency-preserving, let alone conservative (let alone meta-safe in any reasonable way),* as witnessed by the following example. Consider the HOL theory consisting of a single formula φ stating that no type has *precisely* three elements (i.e, for all types α, if α has at most three elements x, y, z then two of them must be equal):

$$\forall x, y, z : \alpha.\ (\forall v : \alpha.\ v = x \ \vee\ v = y \ \vee\ v = z) \longrightarrow x = y \ \vee\ x = z \ \vee\ y = z$$

The theory $\{\varphi\}$ is consistent since there exists a model that satisfies it—the full-frame model of initial HOL, where all finite types are function-space combinations over bool, hence their cardinality is a power of two, in particular, no type has cardinality three. On the other hand, the extension of $\{\varphi\}$ with the definition of a type having three elements, $\tau = \{0, \text{Suc } 0, \text{Suc}(\text{Suc } 0)\}$, is clearly inconsistent—which exhibits a type definition that does not preserve consistency. This analysis has led Wenzel, who is Isabelle's long-standing lead developer and release manager, to deem type definitions *axiomatic* (i.e., having *zero* consistency or conservativity guarantees attached) rather than definitional. This departure from a well-established HOL tradition has generated confusion and misunderstanding amongst Isabelle/HOL's users and developers [Wolff 2015].

But the above counterexample involves a non-definitional theory—φ is not a definition, but merely an axiom that happens to be consistent. Thus, the counterexample only shows that, unlike constant definitions, type definitions do not preserve consistency, a fortiori, are not conservative, *over an arbitrary (axiomatic) theory*. Nonetheless, it is still legitimate to ask:

> *Are arbitrary combinations of constant and type definitions conservative over initial HOL? And are they even meta-safe (again, over initial HOL) in a suitable sense?*

	Over Initial HOL	Over Arbitrary HOL Theories	Over Initial HOL	Over Arbitrary HOL Theories
Constant Definitions	Yes (from right)	Yes [Wenzel 1997]	Yes (from right)	Yes (from below)
Constant Definitions Mixed with Type Definitions	**Yes (this paper)**	No [Wenzel 1997]	Yes (from right)	Yes [Pitts 1993]
Isabelle-HOL Constant Definitions	Yes [Wenzel 1997] [Obua 2006]	No (easy)	Yes (from second left)	No (easy)
Isabelle-HOL Constant Definitions Mixed with Type Definitions	**Yes (this paper)**	No (from above)		No (from above)
	(Proof-Theoretic) Conservativity		Model-Theoretic Conservativity w.r.t. Standard Models	

Fig. 1. Conservativity of Definitions in HOL and Isabelle/HOL

We believe these are important questions for deepening our understanding of the nature of HOL and Isabelle/HOL definitions. Conservativity also provides the most compelling way of witnessing consistency: Any proof of False using definitions can be traced down to a proof of False in initial HOL (the latter being manifestly consistent thanks to its standard set-theoretic semantics). This is especially relevant for the brittle foundational terrain of Isabelle/HOL, where it should help rehabilitating type definitions as genuine, safe definitions.

In this paper, we provide a positive answer to both questions. Figure 1 shows in bold our new conservativity results in the context of similar known facts. For Isabelle/HOL constant definitions, ad hoc overloading immediately causes both (proof-theoretic) conservativity and model-theoretic conservativity over arbitrary base theories to fail. On the other hand, Wenzel [Wenzel 1997] argues by a proof sketch that any set of Isabelle/HOL constant definitions *is* conservative over any base theory provided the latter's signature does not contain these constants—in particular, this covers the case of initial HOL, which later Obua settles by a rigorous proof [Obua 2006]. Moreover, for (HOL and Isabelle/HOL) constant definitions over initial HOL, it is known that we can infer model-theoretic conservativity from conservativity by replacing the defined constants with existentially quantified variables. However, this trick no longer works when we consider combinations of constant *and type* definitions—hence the empty slot in the figure's table, meaning we don't know whether model-theoretic conservativity holds in this case. (This is an open problem only for the case of Isabelle/HOL, since for standard HOL the fact even holds for *arbitrary* base theories, as shown by Pitts's well-known model-theoretic argument.) At the end of Section 5, we briefly come back to these aspects concerning model-theoretic conservativity, and suggest a possible positive answer to fill the figure's empty slot in the light of our techniques. Until then, we will focus entirely on conservativity in the proof-theoretic sense.

Here is an overview of the rest of this paper. First, we focus on traditional HOL, where we formulate meta-safety by defining translation operators for types and terms that unfold the definitions (Section 4). Unfolding a type definition has to be done in an indirect fashion, since HOL does not support comprehension/refinement types (of the form $\{x : \sigma \mid t\ x\}$). Namely, a formula operating on defined types will be relativized to a formula on the original, built-in types that hosted the type definitions; so the "unfolding" of a defined type will be a predicate on its host type. Since type definitions are paired with nonemptiness proofs (in the current contexts, having available all the previously introduced definitions), we are forced to proceed gradually, one definition at a time. Consequently, the proof of meta-safety (also leading to conservativity) is itself gradual, in a feedback loop between preservation of deduction, commutation with substitution, and nonemptiness of the relativization predicates.

We organized the proof development for traditional HOL modularly, separating lemmas about termination of the definitional dependency relation. This allows a smooth upgrade to the more complex case of Isabelle/HOL (Section 5), where termination is no longer ensured by the historic order of definitions, but via a more global approach. Due to ad hoc overloading, here the translations no longer commute with type substitution. We recover from this "anomaly" by mining the proofs and weakening the commutation lemma—leading to an Isabelle/HOL version of the results.

Our constructions have a logical-relation flavor [Reynolds 1983], but with some non-standard (and non-parametric) aspects due to the need to ensure non-emptiness of the representation predicates and, for Isabelle/HOL, to cope with ad hoc polymorphism.

The appendix gives more details on the HOL logic concepts and shows some omitted proofs. We implemented for Isabelle/HOL the unfolding and relativization functions presented in this paper, and used them to check the paper's examples. The documented implementation is available from [Kunčar and Popescu 2017b].

2 MORE RELATED WORK

There is a vast literature on the logical foundations of theorem provers, which we will not attempt to survey here. We focus on work that is directly relevant to our present contribution, from the point of view of either the object logic or the techniques used.

2.1 HOL Foundations

Wiedijk [2009] defines *stateless HOL*, a version of HOL where terms and types carry *in their syntax* information about the defined constants and type constructors. Kumar et al. [2014] define a set-theoretic (Pitts-style) model for stateless HOL and a translation from standard (stateful) HOL with definitions to stateless HOL, thus proving the consistency of both. Their stateful to stateless HOL translation is similar to our translation, in that they both internalize the definitions (which are part of "the state") into "stateless" formulas; however, for conservativity, we need to appeal to pure HOL entities, not to syntactically enriched ones. In a subsequent paper [Kumar et al. 2016], the same authors renounce the stateless HOL detour and prove model-theoretic conservativity directly on initial HOL.

Kumar et al.'s work, which has been mechanized in HOL4, is based on pioneering self-verification work by Harrison [Harrison 2006], who uses HOL Light to give semantic proofs of soundness of the HOL logic without definitional mechanisms, in two flavors: either after removing the infinity axiom from the object HOL logic, or after adding a "universe" axiom to the meta-logic.

2.2 Isabelle/HOL Foundations

Wenzel's work cited in the introduction [Wenzel 1997] sketched proofs of meta-safety and conservativity of constant definitions but left type definitions aside. In spite of Wenzel's theoretical observation that orthogonality and termination are required to ensure meta-safety, overloading of constants remained unchecked in Isabelle/HOL for many years—until Obua looked into the problem and proposed a way to implement Wenzel's observation with an external termination checker [Obua 2006]. Obua also aimed to extend the scope of consistency by factoring in type definitions. But his syntactic proof missed out possible inconsistencies through delayed overloading intertwined with type definitions. Soon after, Wenzel designed and implemented a more structural solution based on work of Haftmann, Obua and Urban (parts of which are reported in [Haftmann and Wenzel 2006]).

The foundational work on Isabelle/HOL was resumed by us in 2014, after the aforementioned inconsistencies caused by delayed overloading and type definitions were discovered. To address the problem, we defined a new dependency relation [Kunčar and Popescu 2015], operating on constants

and types (which became part of the system starting from Isabelle2016). Employing a nonstandard semantics, we proved that, after these modifications, any definitional theory is consistent. In more recent work, we gave an alternative syntactic proof, based on translating HOL to a richer logic, HOLC, having comprehension types as first-class citizens [Kunčar and Popescu 2017a]. The current paper improves on these results, by proving properties much stronger than consistency.

2.3 Other Work

The general-purpose interactive theorem proving community is largely dominated by two successful camps: provers based on type theory (Agda, Coq, Matita, etc.) and provers based on HOL.[2] For the former, the notion of normalizing terms is fairly well studied and well understood [Abel et al. 2007; Altenkirch 1993; Barras 2010; Coquand et al. 1990; Coquand and Spiwack 2006; Geuvers 1993]. Our notion of meta-safety can be seen as the HOL counterpart of type-theoretic normalization, hence as a foundation for HOL-based programming. Of course, the technical challenges we face in HOL are quite different—here, it is not the expressiveness of the logic or of its underlying type system (e.g., fancy dependent types or polymorphism) that complicates the argument, but to a large extent its *lack of expressiveness*: The logic disallows unfolding type definitions, which forces us into a labyrinth of relativization techniques. Another difference is that HOL is an inherently classical logic: Type definitions require possibly non-constructive proofs of nonemptiness, and the Hilbert Choice is paramount. This makes our proof translations less clean than in type theory.

Other foundational work for theorem provers includes Myreen and Davis's mechanized proof of consistency for Milawa [Myreen and Davis 2014], a prover based on first-order logic in the style of ACL2, and Owre and Shankar's set-theoretic semantics of PVS [Owre and Shankar 1999]—featuring a logic similar to HOL, but with dependent types.

Outside the world of theorem proving, conservative extensions are widely employed in mathematical logic, e.g., in the very popular Henkin technique for proving completeness [Henkin 1949]. They are also employed in algebraic specifications to achieve desirable modularity properties [Sannella and Tarlecki 2012]. However, in these fields, *definitional* extensions are often trivially conservative, thanks to their simple equational structure and freshness conditions.

3 HOL PRELIMINARIES

By HOL, we mean classical higher-order logic with Infinity, Choice and rank-one polymorphism, and mechanisms for constant and type definitions and declarations. This section explains all these concepts and features in detail.

3.1 Syntax

All throughout this paper, we fix the following:

- an infinite set TVar, of *type variables*, ranged by α, β
- an infinite set VarN, of *(term) variable names*, ranged by x, y, z

A *type structure* is a pair (K, arOf) where:

- K is a set of symbols, ranged by k, called *type constructors*, containing three special symbols: "bool", "ind" and "$\Rightarrow$" (aimed at representing the type of booleans, an infinite type of individuals and the function type constructor, respectively)
- $\mathsf{arOf} : \mathsf{K} \Rightarrow \mathbb{N}$ is a function associating arities to the type constructors, such that $\mathsf{arOf}(\mathsf{bool}) = \mathsf{arOf}(\mathsf{ind}) = 0$ and $\mathsf{arOf}(\Rightarrow) = 2$.

[2] There are of course successful provers outside these two camps, but they are usually focused on more specialized tasks, and on automation more than on interaction. They include ACL2 [Kaufmann et al. 2000], Dafny [Leino 2010] and Key [Ahrendt et al. 2016].

The *types* associated to (K, arOf), ranged by σ, τ, are defined as follows:

$$\sigma \ ::= \ \alpha \ \mid \ (\sigma_1, \ldots, \sigma_{\mathsf{arOf}(k)})\, k$$

Thus, a type is either a type variable or an n-ary type constructor k postfix-applied to a number of types corresponding to its arity. We write $\mathsf{Type}_{(\mathsf{K},\mathsf{arOf})}$ for the set of types associated to (K, arOf).

A *signature* is a tuple Σ = (K, arOf, Const, tpOf), where:

- (K, arOf) is a type structure
- Const, ranged over by c, is a set of symbols called *constants*, containing four special symbols: "=", "ε", "zero" and "suc" (aimed at representing equality, Hilbert choice of some element from a type, zero and successor, respectively)
- $\mathsf{tpOf} : \mathsf{Const} \Rightarrow \mathsf{Type}$ is a function associating a type to every constant, such that:

$$\mathsf{tpOf}(=) = \alpha \Rightarrow \alpha \Rightarrow \mathsf{bool} \qquad \mathsf{tpOf}(\varepsilon) = (\alpha \Rightarrow \mathsf{bool}) \Rightarrow \alpha$$
$$\mathsf{tpOf}(\mathsf{zero}) = \mathsf{ind} \qquad \mathsf{tpOf}(\mathsf{suc}) = \mathsf{ind} \Rightarrow \mathsf{ind}$$

For the rest of this section, we fix a signature Σ = (K, arOf, Const, tpOf). We usually write Type_Σ, or simply Type, instead of $\mathsf{Type}_{(\mathsf{K},\mathsf{arOf})}$.

$\mathsf{TV}(\sigma)$ is the set of type variables of a type σ. A *type substitution* is a function $\rho : \mathsf{TVar} \Rightarrow \mathsf{Type}$. We let TSubst denote the set of type substitutions. The application of ρ to a type σ, written $\sigma[\rho]$, is defined recursively by $\alpha[\rho] = \rho(\alpha)$ and $((\sigma_1, \ldots, \sigma_m)\, k)[\rho] = (\sigma_1[\rho], \ldots, \sigma_m[\rho])\, k$. If $\alpha_1, \ldots, \alpha_m$ are all different, we write $\tau_1/\alpha, \ldots, \tau_n/\alpha_m$ for the type substitution that sends α_i to τ_i and each $\beta \notin \{\alpha_1, \ldots, \alpha_m\}$ to β. Thus, $\sigma[\tau_1/\alpha, \ldots, \tau_n/\alpha_m]$ is obtained from σ by substituting, for each i, τ_i for all occurrences of α_i.

We say that σ is an *instance* of τ *via* ρ, written $\sigma \leq_\rho \tau$, if $\tau[\rho] = \sigma$. We say that σ is an *instance* of τ, written $\sigma \leq \tau$, if there exists $\rho \in \mathsf{TSubst}$ such that $\sigma \leq_\rho \tau$. Two types σ_1 and σ_2 are called *orthogonal*, written $\sigma_1 \# \sigma_2$, if they have no common instance; i.e., for all τ, $\tau \not\leq \sigma_1$ or $\tau \not\leq \sigma_2$.

Given $\rho_1, \rho_2 \in \mathsf{TSubst}$, we write $\rho_1 \cdot \rho_2$ for their *composition*, defined as $(\rho_1 \cdot \rho_2)(\alpha) = (\rho_1(\alpha))[\rho_2]$. It is easy to see that, for all types σ, it holds that $\sigma[\rho_1 \cdot \rho_2] = \sigma[\rho_1][\rho_2]$.

A *(typed) variable* is a pair of a variable name x and a type σ, written x_σ. We let Var denote the set of variables. A *constant instance* is a pair of a constant and a type, written c_σ, such that $\sigma \leq \mathsf{tpOf}(c)$. We let CInst denote the set of constant instances. We extend the notions of being an instance ($\leq$) and being orthogonal (#) from types to constant instances:

$$c_\tau \leq d_\sigma \ \text{ iff } \ c = d \text{ and } \tau \leq \sigma \qquad\qquad c_\tau \,\#\, d_\sigma \ \text{ iff } \ c \neq d \text{ or } \tau \# \sigma$$

The signature's *terms*, ranged over by s, t, are defined by the grammar:

$$t \ ::= \ x_\sigma \mid c_\sigma \mid t_1\, t_2 \mid \lambda x_\sigma .\, t$$

Thus, a term is either a variable, or a constant instance, or an application, or an abstraction. As usual, we identify terms modulo alpha-equivalence. We let Term_Σ, or simply Term, ranged by s and t, denote the set of terms. Typing is defined as a binary relation between terms and types, written $t : \sigma$, inductively as follows:

$$\frac{x_\sigma \in \mathsf{Var}}{x_\sigma : \sigma} \qquad \frac{c_\sigma \in \mathsf{CInst}}{c_\sigma : \sigma} \qquad \frac{t_1 : \sigma \Rightarrow \tau \qquad t_2 : \sigma}{t_1\, t_2 : \tau} \qquad \frac{t : \tau}{\lambda x_\sigma .\, t : \sigma \Rightarrow \tau}$$

We can apply a type substitution ρ to a term t, written $t[\rho]$, by applying it to the types of all variables and constant instances occurring in t with the usual renaming of bound variables if they get captured. $\mathsf{FV}(t)$ is the set of t's free variables. The term t is called *closed* if it has no free variables: $\mathsf{FV}(t) = \emptyset$. We write $t[s/x_\sigma]$ for the term obtained from t by capture-free substituting the term s for all free occurrences of x_σ.

A *formula* is a term of type bool. We let Fmla_Σ, or simply Fmla, ranged by φ and χ, denote the set of formulas. The formula connectives (e.g., $\wedge$ and $\longrightarrow$) and quantifiers ($\forall$ and $\exists$) are defined in the usual way, starting from the equality primitive. For example, for any type σ, we write $\forall x_\sigma.\ t$ for $\mathsf{all}_\sigma\ (\lambda x_\sigma.\ t)$, where all_σ is the term $\lambda p_{\sigma\Rightarrow\mathsf{bool}}.\ p = (\lambda x_\sigma.\ \mathsf{true})$. The appendix gives more details. Given terms $b : \mathsf{bool}$, $t_1 : \sigma$ and $t_2 : \sigma$, their if-then-else expression, written $\mathsf{if\_t\_e}\ b\ t_1\ t_2$, is the term $\varepsilon\ (\lambda x_\sigma.\ (b \longrightarrow x_\sigma = t_1)\ \wedge\ (\neg\ b \longrightarrow x_\sigma = t_2))$. Its behavior is the expected one: It equals t_1 if b is true and equals t_2 if b is false.

To avoid confusion with the object-logic definitions discussed later, we treat the logical connectives and quantifiers and the if-then-else operator as mere abbreviations (i.e., meta-level definitions of certain HOL terms). When writing terms, we sometimes omit the types of variables if they can be inferred—e.g, we write $\lambda x_\sigma.\ x$ instead of $\lambda x_\sigma.\ x_\sigma$. A *theory* (over Σ) is a set of closed (Σ-)formulas.

3.2 Axioms and Deduction

The HOL axioms, forming the set Ax, are the usual Equality axioms, the Infinity axioms (stating that suc is different from 0 and is injective, which makes the type ind infinite), the classical Excluded Middle and the Choice axiom, which states that the Hilbert choice operator returns an element satisfying its argument predicate (if nonempty): $p_{\alpha\Rightarrow\mathsf{bool}}\ x \longrightarrow p\ (\varepsilon\ p)$.

A *context* Γ is a finite set of formulas. We write $\alpha \notin \Gamma$ to indicate that the type variable α does not appear in any formula in Γ; similarly, $x_\sigma \notin \Gamma$ will indicate that x_σ does not appear *free* in any formula in Γ. We define *deduction* as a ternary relation $\vdash$ between theories D, contexts Γ and formulas φ, written $D; \Gamma \vdash \varphi$.

$$\frac{}{D; \Gamma \vdash \varphi}[\varphi \in \mathsf{Ax} \cup D]\ (\textsc{Fact}) \qquad \frac{}{D; \Gamma \vdash \varphi}[\varphi \in \Gamma]\ (\textsc{Assum}) \qquad \frac{D; \Gamma \vdash \varphi}{D; \Gamma \vdash \varphi[\sigma/\alpha]}[\alpha \notin \Gamma]\ (\text{T-}\textsc{Inst})$$

$$\frac{D; \Gamma \vdash \varphi}{D; \Gamma \vdash \varphi[t/x_\sigma]}[x_\sigma \notin \Gamma]\ (\textsc{Inst}) \qquad \frac{}{D; \Gamma \vdash (\lambda x_\sigma.\ t)\ s = t[s/x_\sigma]}\ (\textsc{Beta}) \qquad \frac{D; \Gamma \vdash f\ x_\sigma = g\ x_\sigma}{D; \Gamma \vdash f = g}[x_\sigma \notin \Gamma]\ (\textsc{Ext})$$

$$\frac{D; \Gamma \cup \{\varphi\} \vdash \chi}{D; \Gamma \vdash \varphi \longrightarrow \chi}\ (\textsc{ImpI}) \qquad \frac{D; \Gamma \vdash \varphi \longrightarrow \chi \quad D; \Gamma \vdash \varphi}{D; \Gamma \vdash \chi}\ (\text{MP})$$

The axioms and the deduction rules we gave here are (a variant of) the standard ones for HOL (as in, e.g., [Gordon and Melham 1993; Harrison 2009]). Different provers implementing standard HOL, such as HOL4, HOL Light, HOL-ProofPower and HOL Zero, may use slightly different sets of logical primitives and slightly different rules and axioms; moreover, they of course differ in their implementation details. However, they all implement the same logic, up to logical equivalence.

We write $D \vdash \varphi$ instead of $D; \emptyset \vdash \varphi$ and $\vdash \varphi$ instead of $\emptyset; \emptyset \vdash \varphi$ (that is, we omit empty contexts and theories). Note that the HOL axioms are not part of the parameter theory D, but are wired together with D in the (Fact) axiom. So $\vdash \varphi$ indicates that φ is provable from the HOL axioms only.

3.3 HOL Definitions and Declarations

Besides deduction, another main component of the HOL logic is a mechanism for introducing new constants and types by spelling out their definitions.

The *built-in type constructors* are bool, ind and $\Rightarrow$. The *built-in constants* are $=$, ε, zero and suc. Since the built-in items have an already specified behavior (by the HOL axioms), only non-built-in items can be defined.

Definition 1.

Constant Definitions: Given a non-built-in constant c such that $\mathsf{tpOf}(c) = \sigma$ and a closed term $t : \sigma$, we let $c_\sigma \equiv t$ denote the formula $c_\sigma = t$. We call $c_\sigma \equiv t$ a *constant definition* provided $\mathsf{TV}(t) \subseteq \mathsf{TV}(c_\sigma)$ (i.e., $\mathsf{TV}(t) \subseteq \mathsf{TV}(\sigma)$).

Type Definitions: Given types τ and σ and a closed term $t : \sigma \Rightarrow \mathsf{bool}$, we let $\tau \equiv t$ denote the formula

$$\exists rep_{\tau \Rightarrow \sigma}.\ \mathsf{One\_One}_{rep} \wedge (\forall y_\sigma.\ t\ y \longleftrightarrow (\exists x_\tau.\ y = rep\ x))$$

where $\mathsf{One\_One}_{rep}$ is the formula stating that *rep* is one-to-one (injective), namely, $\forall x_\tau, y_\tau.\ rep\ x = rep\ y \longrightarrow x = y$. We call $\tau \equiv t$ a *type definition*, provided τ has the form $(\alpha_1, \ldots, \alpha_m)\ k$ such that k is a non-built-in type constructor, the α_i's are all distinct type variables and $\mathsf{TV}(t) \subseteq \{\alpha_1, \ldots, \alpha_m\}$. (Hence, we have $\mathsf{TV}(t) \subseteq \mathsf{TV}(\tau)$, which also implies $\mathsf{TV}(\sigma) \subseteq \mathsf{TV}(\tau)$.)

A type definition expresses the following: The new type $(\alpha_1, \ldots, \alpha_m)\ k$ is embedded in its host type σ via some one-to-one function *rep*, and the image of this embedding consists of the elements of σ for which t holds. Since types in HOL are required to be nonempty, the definition is only accepted if the user provides a proof that $\exists x_\sigma.\ t\ x$ holds. Thus, *to perform a type definition, one must give a nonemptiness proof.*

Type and Constant Declarations: Declarations in HOL are a logical extension mechanism which is significantly milder than definitions—they simply add new items to the signature as "uninterpreted," without providing any definition.

3.4 Signature Extensions and the Initial Signature

In the remainder of this paper, when necessary for disambiguation, we will indicate the signature Σ as a subscript when denoting sets and relations associated to it: Type_Σ, Term_Σ, CInst_Σ, $\vdash_\Sigma$, etc.

Given a signature $\Sigma = (\mathsf{K}, \mathsf{arOf}, \mathsf{Const}, \mathsf{tpOf})$ and an item u, we write $u \in \Sigma$ to mean that $u \in \mathsf{K}$ or $u \in \mathsf{Const}$. Given signatures $\Sigma = (\mathsf{K}, \mathsf{arOf}, \mathsf{Const}, \mathsf{tpOf})$ and $\Sigma' = (\mathsf{K}', \mathsf{arOf}', \mathsf{Const}', \mathsf{tpOf}')$, we say Σ *is included in* Σ', or Σ' *extends* Σ, written $\Sigma \subseteq \Sigma'$, if $\mathsf{K} \subseteq \mathsf{K}'$, $\mathsf{Const} \subseteq \mathsf{Const}'$ and the functions arOf' and tpOf' are extensions of arOf and tpOf, respectively. We write $u \in \Sigma' \setminus \Sigma$ to mean $u \in \Sigma'$ and $u \notin \Sigma$. If $c \notin \mathsf{Const}$ and $\sigma \in \mathsf{Type}_\Sigma$, we write $\Sigma \cup \{(c, \sigma)\}$ for the extension of Σ with a new constant c of type σ. Similarly, if $k \notin \mathsf{K}$, we write $\Sigma \cup \{(k, n)\}$ for the extension of Σ with a new type constructor k of arity n.

We write Σ_{init} for the *initial signature*, containing only built-in type constructors and constants. Note that, by definition, any signature extends the initial signature.

4 CONSERVATIVITY OF HOL DEFINITIONS

A HOL development, i.e., a session of interaction with the HOL logic from a user's perspective, consists of intertwining definitions, declarations and (statements and proofs of) theorems. Since theorems are merely consequences of definitions, we will not model them explicitly, but focus on definitions and declarations.

Let $\Sigma = (\mathsf{K}, \mathsf{arOf}, \mathsf{Const}, \mathsf{tpOf})$ be a signature and let D be a finite theory over Σ.

Definition 2. D is said to be a *well-formed definitional theory* if $D = \{def_1, \ldots, def_n\}$, where each def_i is a (type or constant) definition of the form $u_i \equiv t_i$, and there exist the signatures $\Sigma^1, \ldots, \Sigma^n$ and $\Sigma_0, \Sigma_1, \ldots, \Sigma_n$ such that $\Sigma_0 = \Sigma_{\mathsf{init}}$, $\Sigma_n = \Sigma$ and the following hold for all $i \in \{1, \ldots, n\}$:

(1) $t_i \in \mathsf{Term}_{\Sigma^i}$ and Σ_i is the extension of Σ^i with a fresh item defined by def_i, namely:
(1.1) If u_i has the form $(\alpha_1, \ldots, \alpha_m)\ k$, then $k \notin \Sigma^i$ and $\Sigma_i = \Sigma^i \cup \{(k, m)\}$
(1.2) If u_i has the form c_σ, then $c \notin \Sigma^i$ and $\Sigma_i = \Sigma^i \cup \{(c, \sigma)\}$
(2) If def_i is a type definition, meaning u_i is a type and $t_i : \sigma \Rightarrow \mathsf{bool}$, then $\{def_1, \ldots, def_{i-1}\} \vdash_{\Sigma^i} \exists x_\sigma.\ t_i\ x$

(3) $\Sigma_{i-1} \subseteq \Sigma^i$

These conditions express that the theory D consists of intertwined definitions and declarations. The chain of extensions

$$\Sigma_{\mathsf{init}} = \Sigma_0 \subseteq \Sigma^1 \subseteq \Sigma_1 \subseteq \Sigma^2 \subseteq \Sigma_2 \ldots \subseteq \Sigma^n \subseteq \Sigma_n = \Sigma,$$

starting from the initial signature and ending with Σ, alternates sets of declarations (the items in $\Sigma^i \setminus \Sigma_{i-1}$) with definitions (the unique item u_i in $\Sigma_i \setminus \Sigma^i$ being defined by def_i, i.e., as $u_i \equiv t_i$). As shown by condition (2), in the case of type definitions, we also require proofs of non-emptiness of the defining predicate t (from the definitions available so far).

In short, the above conditions state something very basic: Definitions are introduced one at a time and the defined symbols are fresh. This is clearly obeyed by correct implementations of standard HOL, such as HOL4 and HOL Light. (By contrast, the Isabelle/HOL-specific conditions in Section 5 will involve the more complex notions of orthogonality and termination.)

Definition 3. A theory E over Σ is said to be a *(proof-theoretic) conservative extension of initial HOL* if any formula proved from E that belongs to the initial signature Σ_{init} could have been proved without E or the types and constants from outside of Σ. Formally: For all $\varphi \in \mathsf{Fmla}_{\Sigma_{\mathsf{init}}}$, $E \vdash_\Sigma \varphi$ implies $\vdash_{\Sigma_{\mathsf{init}}} \varphi$.

4.1 Roadmap

In what follows, we fix a well-formed definitional theory D and use for it the notations introduced in Def. 2, e.g., Σ, Σ_i. We first sketch the main ideas of our development, motivating the choice of the concepts. The more formal definitions and proofs will be given in the following subsections.

Our two main goals are to *formulate and prove D's meta-safety* and to *prove D's conservativity*. As with any respectable notion of its kind, meta-safety will easily yield conservativity, so we concentrate our efforts on the former.[3]

4.1.1 Unfolding the Definitions. Recall that, for a Σ-formula φ provable from D, meta-safety should allow us to replace all the defined items in φ with items in the initial signature without losing provability, i.e., obtaining a deducible Σ_{init}-formula φ'. For constants, the procedure is clear: Any defined constant c appearing in φ is replaced with its defining term t, then any defined constant d appearing in t is replaced with its defining term, and so on, until (hopefully) the process terminates and we are left with built-in items only.

But how about for types τ occurring in φ? A HOL type definition $\tau \equiv t$ where $t : \sigma \Rightarrow \mathsf{bool}$, is not an equality (there is no type equality in HOL), but a formula asserting the existence of a bijection between τ and the set of elements of σ for which the predicate t holds. So it cannot be "unfolded." First, let us make the simplifying assumption that $\sigma \in \mathsf{Type}_{\Sigma_{\mathsf{init}}}$ and $t \in \mathsf{Term}_{\Sigma_{\mathsf{init}}}$. Then the only reasonable Σ_{init}-substitute for τ is its host type σ; however, after the replacement of τ by σ, the formula needs to be adjusted not to refer to the whole σ, but only to the isomorphic copy of τ—in other words, the formula needs to be relativized to the predicate t. In general, σ or t may themselves contain defined types or constants, which will need to be processed similarly, and so on, recursively. In summary:

- for each type τ, we define its host type $\mathsf{HOST}(\tau) \in \mathsf{Type}_{\Sigma_{\mathsf{init}}}$ and its relativization predicate on that type, $\mathsf{REL}(\tau) : \mathsf{HOST}(\tau) \Rightarrow \mathsf{bool}$ (where $\mathsf{REL}(\tau) \in \mathsf{Term}_{\Sigma_{\mathsf{init}}}$)

[3] A note on terminology: In this paper's title, abstract and introduction, we use the term *safety* to refer to the informal notion of a definition being "safe," i.e., being treatable as a form of abbreviation. On the other hand, *meta-safety* is a technical term introduced by Wenzel for a mathematical formulation of safety for constant definitions. We will introduce our own notion of meta-safety, extending Wenzel's.

- for each term $t : \tau$, we define its unfolding $\mathsf{UNF}(t) : \mathsf{HOST}(\tau)$ (where $\mathsf{UNF}(t) \in \mathsf{Term}_{\Sigma_{\mathsf{init}}}$)

We will illustrate our design choices for the various cases in defining the above translation functions with the help of a running example.

Example 4. Let Σ be the extension of the initial signature with:

- the nullary type constructors nat and zfun
- the constants absnat : ind $\Rightarrow$ nat and z : nat

Let $D = \{def_i \mid i \in \{1, \ldots, 4\}\}$, such that:

- def_1 is nat $\equiv t_1$, where t_1 : ind $\Rightarrow$ bool is a predicate taking the intersection of all predicates that hold for 0 and are closed under Suc. Formally, t_1 is $\lambda i_{\mathsf{ind}}.\ (\forall P_{\mathsf{ind}\Rightarrow\mathsf{bool}}.\ P\ 0 \wedge (\forall j_{\mathsf{ind}}.\ P\ j \longrightarrow P\ (\mathsf{Suc}\ j))) \longrightarrow P\ i$; but the precise form of t_1 will not be important in our discussion, beyond the fact that it is a term in the initial signature.
- def_2 is absnat $\equiv \varepsilon\ t_2$, where t_2 : (ind $\Rightarrow$ nat) $\Rightarrow$ bool is a predicate stating about its argument function that it is a bijection between the elements of ind that satisfy t_1 and nat. Formally, t_2 is $\lambda f_{\mathsf{ind}\Rightarrow\mathsf{nat}}.\ \varphi_1 \wedge \varphi_2$, where:
  - φ_1 states that f is one-to-one on the elements satisfying t_1, namely $\forall i_{\mathsf{ind}}\ j_{\mathsf{ind}}.\ t_1\ i\ \wedge\ t_1\ j\ \wedge\ f\ i = f\ j \longrightarrow i = j$
  - φ_2 states that f maps t_1 onto nat, namely $\forall n_{\mathsf{nat}}.\ \exists i_{\mathsf{ind}}.\ t_1\ i\ \wedge\ f\ i = n$
- def_3 is z $\equiv t_3$, where t_3 is absnat 0.
- def_4 is zfun $\equiv t_4$, where t_4 : (nat $\Rightarrow$ nat) $\Rightarrow$ bool is $\lambda f_{\mathsf{nat}\Rightarrow\mathsf{nat}}.\ f\ \mathsf{z} = \mathsf{z}$.

Thus, there are no (non-defined but) declared items, and the chain $\Sigma_{\mathsf{init}} = \Sigma_0 \subseteq \Sigma^1 \subseteq \Sigma_1 \subseteq \ldots \subseteq \Sigma^4 \subseteq \Sigma_4 = \Sigma$ consists of the following signatures, where we do not repeat the arities and the types:

$$\Sigma^1 = \Sigma_0 = \Sigma_{\mathsf{init}} \qquad \Sigma^3 = \Sigma_2 = \Sigma^2 \cup \{\mathsf{absnat}\} \qquad \Sigma = \Sigma_4 = \Sigma^4 \cup \{\mathsf{zfun}\}$$
$$\Sigma^2 = \Sigma_1 = \Sigma^1 \cup \{\mathsf{nat}\} \qquad \Sigma^4 = \Sigma_3 = \Sigma^3 \cup \{\mathsf{z}\}$$

Incidentally, this example shows the standard procedure of bootstrapping natural numbers in HOL: The type nat is defined by carving out, from HOL's built-in infinite type ind, the smallest set closed under zero and successor. Using the Choice operator, we define the abstraction function absnat as a surjection whose restriction to nat's defining predicate t_1 is a bijection to nat. (The opposite injection can of course also be defined, but is omitted here.) The version of zero for naturals, z : nat, is defined by applying the abstraction to the built-in zero from ind. Subsequently, another type is introduced, zfun, of zero-preserving functions between naturals, defined by carving out from the type nat $\Rightarrow$ nat the set of those functions that map z to z.

The simplest of the three translation functions will be HOST, which will track recursively, for each defined type, the built-in type that represents its defining ancestor. For example, following the type definitions def_1 and def_4, we can compute the host of zfun:

$$\mathsf{HOST}(\mathsf{zfun}) \ = \ \mathsf{HOST}(\mathsf{nat} \Rightarrow \mathsf{nat}) \ = \ \mathsf{HOST}(\mathsf{nat}) \Rightarrow \mathsf{HOST}(\mathsf{nat}) \ = \ \mathsf{ind} \Rightarrow \mathsf{ind}$$

The UNF function will be more challenging to define. A clearly desirable feature is that UNF should leave built-in constants unchanged, e.g., UNF(=) should be =. Moreover, for instances c_σ of constants $c : \tau$ defined by equations $c_\tau \equiv t$, $\mathsf{UNF}(c_\sigma)$ will naturally be recursively defined as $\mathsf{UNF}(t[\rho])$ where ρ is the substitution that makes σ an instance of τ (i.e., $\sigma \leq_\rho \tau$). In other words, we unfold c_σ with the appropriately substituted equation defining c. Since UNF is applied to arbitrary terms, not only to constants, we must indicate its recursive behavior for all term constructs. Abstraction and application are handled as expected, in that UNF distributes over them—with changing the type of the bound variables through the HOST function. For example, starting with t_4 which is $\lambda f_{\mathsf{nat}\Rightarrow\mathsf{nat}}.\ f_{\mathsf{nat}\Rightarrow\mathsf{nat}}\ \mathsf{z} = \mathsf{z}$, we have the following equalities, where UNF delves recursively

into abstractions and applications, unfolds the definitions def_3 of z and def_2 of absnat, and leaves the built-in constants =, 0 and ε unchanged:

$$\begin{array}{rcl}
\mathsf{UNF}(t_4) & = & \lambda f_{\mathsf{HOST(nat\Rightarrow nat)}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}}\ \mathsf{z} = \mathsf{z}) \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ \mathsf{UNF(z)}\ \mathsf{UNF}(=)\ \mathsf{UNF(z)} \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ \mathsf{UNF(absnat\ 0)} = \mathsf{UNF(absnat\ 0)} \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ (\mathsf{UNF(absnat)}\ \mathsf{UNF(0)}) = \mathsf{UNF(absnat)}\ \mathsf{UNF(0)} \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ (\mathsf{UNF}(\varepsilon\ t_2)\ 0) = \mathsf{UNF}(\varepsilon\ t_2)\ 0 \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ (\mathsf{UNF}(\varepsilon)\ \mathsf{UNF}(t_2)\ 0) = \mathsf{UNF}(\varepsilon)\ \mathsf{UNF}(t_2)\ 0 \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})\ (\varepsilon\ \mathsf{UNF}(t_2)\ 0) = \varepsilon\ \mathsf{UNF}(t_2)\ 0
\end{array}$$

Unlike applications and abstractions, variables raise a subtle issue, with global implications on our overall proof strategy. But before discussing them, we must look at how to define the relativization predicates. Clearly, REL should send a defined type such as nat to the unfolding of its defining predicate, here, $\mathsf{UNF}(t_1)$. (Note that in this case t_1 happens to contain only built-in items, meaning $\mathsf{UNF}(t_1) = t_1$.) Moreover, REL should "distribute" over $\Rightarrow$ in that $\mathsf{REL}(\sigma_1 \Rightarrow \sigma_2) = \mathsf{REL}(\sigma_1) \Rightarrow \mathsf{REL}(\sigma_2)$ where, for $p_1 : \tau_1 \Rightarrow \mathsf{bool}$ and $p_2 : \tau_2 \Rightarrow \mathsf{bool}$, $p_1 \Rightarrow p_2$ is the predicate on $\tau_1 \Rightarrow \tau_2$ stating about its argument function that it maps elements satisfying p_1 to elements satisfying p_2—i.e., $p_1 \Rightarrow p_2$ is the lifting of p_1 and p_2 to the function space. For example:

$$\mathsf{REL(nat \Rightarrow nat)} = \mathsf{REL(nat)} \Rightarrow \mathsf{REL(nat)} = t_1 \Rightarrow t_1$$

But what if a type is defined from a type that itself contains other defined types, as is the case of zfun defined from nat $\Rightarrow$ nat (according to def_4)? Then we must accumulate the defining predicates of all intermediate types, each lifted if necessary along the encountered function-space structure:

$$\begin{array}{rcl}
\mathsf{REL(zfun)} & = & \lambda f_{\mathsf{HOST(zfun)}}.\ \mathsf{REL(nat \Rightarrow nat)}\ f_{\mathsf{HOST(zfun)}}\ \wedge\ \mathsf{UNF}(t_4)\ f_{\mathsf{HOST(zfun)}} \\
& = & \lambda f_{\mathsf{ind\Rightarrow ind}}.\ (t_1 \Rightarrow t_1)\ f_{\mathsf{ind\Rightarrow ind}}\ \wedge\ \mathsf{UNF}(t_4)\ f_{\mathsf{ind\Rightarrow ind}}
\end{array}$$

Thus, $\mathsf{REL(zfun)}\ f_{\mathsf{ind\Rightarrow ind}}$ states that f preserves t_1 (the defining predicate of nat from ind) and that $\mathsf{UNF}(t_4)\ f$ holds, where t_4 is the defining predicate of zfun from nat $\Rightarrow$ nat.

Back to the unfolding of variables, we are now ready to ask what should $\mathsf{UNF}(x_\sigma)$ be. An obvious candidate is $x_{\mathsf{HOST}(\sigma)}$. However, this will not work, since a crucial property that we will need about our translation is that it observes membership to types, in that it maps terms of a given type to terms satisfying that type's representing predicate:

> (F1) The relativization predicates hold on translated items, i.e., $\mathsf{REL}(\sigma)\ \mathsf{UNF}(t)$ is deducible (in initial HOL) for each term $t : \sigma$.

In particular, any $\mathsf{REL}(\sigma)\ \mathsf{UNF}(x_\sigma)$, e.g., $\mathsf{REL(nat \Rightarrow nat)}\ \mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})$, should be deducible. To enforce this, we define $\mathsf{UNF}(x_\sigma)$ to be either $x_{\mathsf{HOST}(\sigma)}$ if $\mathsf{REL}(\sigma)\ x_{\mathsf{HOST}(\sigma)}$ holds, or else any item for which $\mathsf{REL}(\sigma)$ holds. This is expressible using the if-then-else and Choice operators: $\mathsf{if\_t\_e}\ (\mathsf{REL}(\sigma)\ x_{\mathsf{HOST}(\sigma)})\ x_{\mathsf{HOST}(\sigma)}\ (\varepsilon\ \mathsf{REL}(\sigma))$. For example:

$$\begin{array}{rcl}
\mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}}) & = & \mathsf{if_t_e}\ (\mathsf{REL(nat \Rightarrow nat)}\ f_{\mathsf{HOST(nat\Rightarrow nat)}})\ f_{\mathsf{HOST(nat\Rightarrow nat)}}\ (\varepsilon\ \mathsf{REL(nat \Rightarrow nat)}) \\
& = & \mathsf{if_t_e}\ ((t_1 \Rightarrow t_1)\ f_{\mathsf{ind\Rightarrow ind}})\ f_{\mathsf{ind\Rightarrow ind}}\ (\varepsilon\ (t_1 \Rightarrow t_1))
\end{array}$$

In other words, $\mathsf{UNF}(f_{\mathsf{nat\Rightarrow nat}})$ is either $f_{\mathsf{ind\Rightarrow ind}}$ if $f_{\mathsf{ind\Rightarrow ind}}$ happens to preserve t_1, or otherwise some element that preserves t_1.

By the Choice axiom, $\mathsf{REL}(\sigma)$ holds for $\varepsilon\ \mathsf{REL}(\sigma)$ just in case $\mathsf{REL}(\sigma)$ is nonempty. So to achieve the goal of ensuring $\mathsf{REL}(\sigma)$ holds for x_σ, we need:

> (F2) The relativization predicates are nonempty, i.e., $\exists x_{\mathsf{HOST}(\sigma)}.\ \mathsf{REL}(\sigma)\ x$ is deducible.

(For our example, this would mean that there exists an element of ind ⇒ ind that preserves t_1.) Another way to regard this property is as a reflection of the HOL types being nonempty—a faithful relativization should of course follow suit.

With our chosen behavior of UNF on variables, the formula connectives and quantifiers will be treated as desired, i.e., yielding (modulo HOL deduction) standard relativization with respect to the REL predicates—for the universal and existential quantifiers, this means bounded quantification. For example, writing $t_1 =_{\mathrm{HOL}} t_2$ for $\vdash_{\Sigma_{\mathsf{init}}} t_1 = t_2$, i.e., for the fact that the equality $t_1 = t_2$ is deducible in initial HOL, we have:

$$\begin{array}{rcl}
\mathsf{UNF}(\forall x_\sigma.\ \varphi\ x_\sigma) & = & \mathsf{UNF}((\lambda x_\sigma.\ \varphi\ x_\sigma) = (\lambda x_\sigma.\ \mathsf{true})) \\
& = & (\lambda x_{\mathsf{HOST}(\sigma)}.\ \mathsf{UNF}(\varphi)\ \mathsf{UNF}(x_\sigma)) = (\lambda x_{\mathsf{HOST}(\sigma)}.\ \mathsf{true}) \\
& = & \forall x_{\mathsf{HOST}(\sigma)}.\ \mathsf{UNF}(\varphi)\ \mathsf{UNF}(x_\sigma) \\
& = & \forall x_{\mathsf{HOST}(\sigma)}.\ \mathsf{UNF}(\varphi)\ (\mathsf{if_t_e}\ (\mathsf{REL}(\sigma)\ x_{\mathsf{HOST}(\sigma)})\ x_{\mathsf{HOST}(\sigma)}\ (\epsilon\ \mathsf{REL}(\sigma))) \\
& =_{\mathrm{HOL}} & \forall x_{\mathsf{HOST}(\sigma)}.\ \mathsf{if_t_e}\ (\mathsf{REL}(\sigma)\ x_{\mathsf{HOST}(\sigma)})\ (\mathsf{UNF}(\varphi)\ x_{\mathsf{HOST}(\sigma)})\ (\mathsf{UNF}(\varphi)\ (\epsilon\ \mathsf{REL}(\sigma))) \\
& =_{\mathrm{HOL}} & \forall x_{\mathsf{HOST}(\sigma)}.\ \mathsf{REL}(\sigma)\ x_{\mathsf{HOST}(\sigma)} \longrightarrow \mathsf{UNF}(\varphi)\ x_{\mathsf{HOST}(\sigma)}
\end{array}$$

The last $=_{\mathrm{HOL}}$ step in the above chain follows from the non-emptiness of $\mathsf{REL}(\sigma)$. Similarly, the unfolding of $\exists x_\sigma.\ \varphi\ x$ will be equal modulo HOL deduction to $\exists x_{\mathsf{HOST}(\sigma)}.\ \mathsf{REL}(\sigma)\ x\ \wedge\ \mathsf{UNF}(\varphi)\ x$.

We can take advantage of the above observation to obtain a palatable form for $\mathsf{UNF}(t_2)$:

$$\begin{array}{rcl}
\mathsf{UNF}(t_2) & =_{\mathrm{HOL}} & \lambda f_{\mathsf{ind}\Rightarrow\mathsf{ind}}.\ \mathsf{UNF}(\varphi_1) \wedge \mathsf{UNF}(\varphi_2) \\
\mathsf{UNF}(\varphi_1) & =_{\mathrm{HOL}} & \forall i_{\mathsf{ind}}\ j_{\mathsf{ind}}.\ t_1\ i\ \wedge\ t_1\ j\ \wedge\ s\ i = s\ j \longrightarrow i = j \\
\mathsf{UNF}(\varphi_2) & =_{\mathrm{HOL}} & \forall n_{\mathsf{ind}}.\ t_1\ n \longrightarrow \exists i_{\mathsf{ind}}.\ t_1\ i \wedge s\ i = n
\end{array}$$

where s is the term $\mathsf{if\_t\_e}\ ((t_1 \Rightarrow t_1)f)\ f\ (\varepsilon\ (t_1 \Rightarrow t_1))$. Thus, $\mathsf{UNF}(t_2)$ states about its argument f that, if it preserves t_1, then it is in fact a bijection on the set of elements of ind that satisfy t_1.

The above discussion suggests that the desired Σ_{init}-formula φ' corresponding to a Σ-formula φ should be $\mathsf{UNF}(\varphi)$. Hence, for us meta-safety over initial HOL will mean:

(MS) For all $\varphi \in \mathsf{Fmla}_\Sigma$, $D \vdash_\Sigma \varphi$ implies $\vdash_{\Sigma_{\mathsf{init}}} \mathsf{UNF}(\varphi)$.

This property is indeed a type-aware version of what Wenzel calls meta-safety: $\mathsf{UNF}(\varphi)$ replaces each defined constant with a term as in Wenzel's concept, and replaces each defined type with a tandem of a host type and a relativization predicate.

For our running example, we can prove $D \vdash_\Sigma \varphi$, where φ is $\forall f_{\mathsf{nat}\Rightarrow\mathsf{nat}}.\ \exists g_{\mathsf{nat}\Rightarrow\mathsf{nat}}.\ \neg\ f = g$, which is a way of saying that nat ⇒ nat, if it is not empty (which is true for all HOL types) then it is not a singleton. By our meta-safety result, we will infer $\vdash_{\Sigma_{\mathsf{init}}} \mathsf{UNF}(\varphi)$, where

$$\begin{array}{rcl}
\mathsf{UNF}(\varphi) & =_{\mathrm{HOL}} & \forall f_{\mathsf{HOST}(\mathsf{nat}\Rightarrow\mathsf{nat})}.\ \mathsf{REL}(\mathsf{nat}\Rightarrow\mathsf{nat})\ f \longrightarrow \exists g_{\mathsf{HOST}(\mathsf{nat}\Rightarrow\mathsf{nat})}.\ \mathsf{REL}(\mathsf{nat}\Rightarrow\mathsf{nat})\ g \wedge \neg\ f = g \\
& = & \forall f_{\mathsf{ind}\Rightarrow\mathsf{ind}}.\ (t_1 \Rightarrow t_1)\ f \longrightarrow \exists g_{\mathsf{ind}\Rightarrow\mathsf{ind}}.\ (t_1 \Rightarrow t_1)\ g \wedge \neg\ f = g
\end{array}$$

This is indeed a tautology (provable in initial HOL): It says that for any function that preserves the natural-number predicate (i.e., t_1) there exists a different function with the same property. This follows from the fact that there are two distinct elements of ind satisfying t_1, e.g., 0 and Suc 0.

To help proving (MS), we will also have lemmas about the good behavior of the translation functions HOST, UNF and REL with respect to the main ingredients of HOL deduction:

(F3) The translation functions preserve variable freshness and commute with substitution.

The order in which we will have to prove these facts has superficially circular dependencies. As discussed, we need (F2) for proving (F1). Moreover, (F1) is needed to prove (F3), more precisely, to make sure that UNF commutes with substitution for the delicate case of variables x_σ. In turn, (F3) is used for (MS). But to prove (F2), the nonemptiness of the relativization predicates, we seem to need (MS). Indeed, for the case of a type τ defined by $\tau \equiv t$ with $t : \sigma \Rightarrow \mathsf{bool}$, $\mathsf{REL}(\tau)$ is the conjunction of $\mathsf{REL}(\sigma)$ and $\mathsf{UNF}(t)$. So, in an inductive proof of (F2), we will need to deduce

$\exists x_{\mathsf{HOST}(\sigma)}.\ \mathsf{REL}(\sigma)\ x \wedge \mathsf{UNF}(t)\ x$. The only fact that can help here is that this formula is (equivalent to) $\mathsf{UNF}(\varphi)$, where φ is $\exists x_\sigma.\ t\ x$. Since φ is the non-emptiness claim for the new type τ, it is deducible (according to Def. 2(2)). So we would like to apply (MS) here for obtaining that $\mathsf{UNF}(\varphi)$ is deducible.

Thus, we have the apparent dependency loop

$$\text{(MS)} \Longrightarrow \text{(F3)} \Longrightarrow \text{(F1)} \Longrightarrow \text{(F2)} \Longrightarrow \text{(MS)}$$

The way out of this loop is a *gradual* approach: We will not define a single version of the translation functions, but one version, HOST_i, UNF_i and REL_i, for each subset $\{def_1, \ldots, def_i\}$ of D with $i \leq n$. This way, we can use (MS) for i to prove (F2) for $i + 1$.

4.1.2 Dealing With Declarations. Lastly, we must take into account a phenomenon we have ignored so far: the presence of declarations in addition to definitions.

Example 5. Consider the following extension of Example 4: After def_4, a declaration of a constant c : zfun is performed. Thus, we have a new signature $\Sigma^5 = \Sigma_4 \cup \{(c, \mathsf{zfun})\}$.

What should the unfolding $\mathsf{UNF}(c_\sigma)$ of the declared constant c : zfun be? A possibility is to acknowledge c as an irreducible entity, and define $\mathsf{UNF}(c_{\mathsf{zfun}}) = c_{\mathsf{HOST(zfun)}}$. However, this way our desirable property (F1), here, $\mathsf{REL(zfun)}\ c_{\mathsf{HOST(zfun)}}$, will not be provable, since nothing prevents the "uninterpreted" items $c_{\mathsf{HOST(zfun)}}$ from being outside of the relativization predicate. Another alternative is to define $\mathsf{UNF}(c_{\mathsf{zfun}})$ as an arbitrary element satisfying REL(zfun), via Choice, i.e., as ϵ REL(zfun). But this would mean that UNF will artificially identify several distinct constants, e.g., $\mathsf{UNF}(c_\sigma) = \mathsf{UNF}(d_\sigma)$ for any two declared constants c_σ and d_σ—besides being unnatural, this situation would become difficult to handle later, for Isabelle/HOL, since it would introduce a breach in monotony: When declaring c_σ and d_σ, their unfoldings would be equal, but at a later stage one of them could get defined, breaking this equality.

In summary, we wish to preserve the identity of the declared constants such as c_{zfun}, while still enforcing $\mathsf{REL(zfun)}\ \mathsf{UNF}(c_{\mathsf{zfun}})$. We achieve this by treating c_{zfun} in a guarded fashion, similarly to the variables—here, taking $\mathsf{UNF}(c_{\mathsf{zfun}})$ to be $\mathsf{if\_t\_e}\ (\mathsf{REL(zfun)}\ c_{\mathsf{HOST(zfun)}})\ c_{\mathsf{HOST(zfun)}}\ (\varepsilon\ \mathsf{REL(zfun)})$, i.e., $\mathsf{if\_t\_e}\ (\mathsf{REL(zfun)}\ c_{\mathsf{ind}\Rightarrow\mathsf{ind}})\ c_{\mathsf{ind}\Rightarrow\mathsf{ind}}\ (\varepsilon\ \mathsf{REL(zfun)})$.

Another subtlety concerning declared constants lies in the question: What should be the *signature* of UNF (c_{zfun})? Since c has no definition, it will not be compiled away by unfolding. However, its type zfun, which is a defined type, must be translated into its host $\mathsf{ind} \Rightarrow \mathsf{ind}$. But none of the existing signatures contains a constant $c : \mathsf{ind} \Rightarrow \mathsf{ind}$. Consequently, we must create a signature Δ that extends Σ_{init} with all the declared constants but having HOST-translated types, and, similarly, with all the declared type constructors. As for the declared (but not defined) types, these can be kept in the signature Δ without causing any problems. Our translations, as well as the statement of (MS), will target this extended signature Δ rather than Σ_{init}.

4.2 Formal Definition of the Translations and Meta-Safety

We will write D_i for the current definitional theory at moment i, $\{def_1, \ldots, def_i\}$. Thus, we have $D = D_n$. As discussed in the previous subsection, we will define deduction-preserving translations of the Σ-types and Σ-terms into Δ-types and Δ-terms, where Δ will be a suitable signature that collects all the declared items—namely, for each Σ-type we define its host Δ-type and its relativization predicate (which is a Δ-term) and for each Σ-term we define its unfolding (which is a Δ-term). We proceed gradually, considering the Σ_i's one i at a time, eventually reaching $\Sigma = \Sigma_n$.

For each $i \in \{1, \ldots, n\}$, we define the signature Δ^i (collecting the declared items from Σ^i with their types translated to their host types), together with the function $\mathsf{HOST}_i : \mathsf{Type}_{\Sigma_i} \Rightarrow \mathsf{Type}_{\Delta^i}$ (producing the host types) as follows:

(H1) $\mathsf{HOST}_i(\alpha) = \alpha$
(H2) $\mathsf{HOST}_i((\sigma_1, \ldots, \sigma_m)\, k) = (\mathsf{HOST}_i(\sigma_1), \ldots, \mathsf{HOST}_i(\sigma_m))\, k$,
if $k \in \Sigma^1 \cup \bigcup_{i'=2}^{i} (\Sigma^{i'} \setminus \Sigma_{i'-1})$
(H3) $\mathsf{HOST}_i((\sigma_1, \ldots, \sigma_m)\, k) = \mathsf{HOST}_i(\sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m])$,
if $(\alpha_1, \ldots, \alpha_m)\, k \equiv t$ is in D_i and $t : \sigma \Rightarrow \mathsf{bool}$

(R1) $\mathsf{REL}_i(\sigma) = \lambda x_\sigma.\ \mathsf{true}$, if $\sigma \in \mathsf{TVar} \cup \{\mathsf{bool}, \mathsf{ind}\}$
(R2) $\mathsf{REL}_i(\sigma_1 \Rightarrow \sigma_2) = \lambda f_{\mathsf{HOST}_i(\sigma_1) \Rightarrow \mathsf{HOST}_i(\sigma_2)}.\ \forall x_{\mathsf{HOST}_i(\sigma_1)}.\ \mathsf{REL}_i(\sigma_1)\, x \longrightarrow \mathsf{REL}_i(\sigma_2)\, (f\ x)$
(R3) $\mathsf{REL}_i((\sigma_1, \ldots, \sigma_m)\, k) = \lambda x_{(\mathsf{HOST}_i(\sigma_1), \ldots, \mathsf{HOST}_i(\sigma_m))\, k}.\mathsf{true}$, if $k \in \bigcup_{i'=1}^{i} (\Sigma^{i'} \setminus \Sigma_{i'-1})$
(R4) $\mathsf{REL}_i((\sigma_1, \ldots, \sigma_m)k) = \lambda x_{\mathsf{HOST}_i(\sigma')}.\ \mathsf{REL}_i(\sigma')\, x\ \wedge\ \mathsf{UNF}_i(t')\, x$,
if $(\alpha_1, \ldots, \alpha_m)\, k \equiv t$ is in D_i and $t : \sigma \Rightarrow \mathsf{bool}$,
where $\sigma' = \sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m]$ and $t' = t[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m]$

(U1) $\mathsf{UNF}_i(x_\sigma) = \mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)\, x_{\mathsf{HOST}_i(\sigma)})\, x\ (\varepsilon\ \mathsf{REL}_i(\sigma))$
(U2) $\mathsf{UNF}_i(c_\sigma) = c_{\mathsf{HOST}_i(\sigma)}$, if $c \in \Sigma_{\mathsf{init}}$
(U3) $\mathsf{UNF}_i(c_\sigma) = \mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)\, c_{\mathsf{HOST}_i(\sigma)})\, c_{\mathsf{HOST}_i(\sigma)}\ (\varepsilon\ \mathsf{REL}_i(\sigma))$, if $c \in \bigcup_{i'=1}^{i} (\Sigma^{i'} \setminus \Sigma_{i'-1})$
(U4) $\mathsf{UNF}_i(c_\sigma) = \mathsf{UNF}_i(t[\rho])$, if $c_\tau \equiv t$ is in D_i and $\sigma \leq_\rho \tau$
(U5) $\mathsf{UNF}_i(t_1\ t_2) = \mathsf{UNF}_i(t_1)\ \mathsf{UNF}_i(t_2)$
(U6) $\mathsf{UNF}_i(\lambda x_\sigma.\ t) = \lambda x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{UNF}_i(t)$

Fig. 2. Definition of the translation functions

- Δ^1 is Σ^1
- Δ^{i+1} is Δ^i extended with:
  - all the type constructors $k \in \Sigma^{i+1} \setminus \Sigma_i$
  - for all constants $c \in \Sigma^{i+1} \setminus \Sigma_i$ of type σ, a constant c of type $\mathsf{HOST}_i(\sigma)$
- HOST_i is defined as in Fig. 2, recursively on types

On defined types (i.e., types having a defined type constructor on top), HOST_i behaves as prescribed in Section 4.1, recursively calling itself for the defining type (clause (H3)). Upon encountering built-in or declared type constructors, i.e., belonging to some $\Sigma^{i'}$ for $i' \leq i$, but not to the corresponding $\Sigma_{i'-1}$, HOST_i delves into the subexpressions (clause (H2)).

Next, *mutually* recursively on Σ_i-types and Σ_i-terms, we define a function returning the relativization predicate of a type, $\mathsf{REL}_i : \mathsf{Type}_{\Sigma_i} \to \mathsf{Term}_{\Delta^i}$, and one returning the unfolded term, $\mathsf{UNF}_i : \mathsf{Term}_{\Sigma_i} \to \mathsf{Term}_{\Delta^i}$. Their definition is shown in Fig. 2 (where we again make use of the convention that we don't show the type labels of variables when they can be inferred). Again, they behave as prescribed in Section 4.1. In particular, REL_i is naturally lifted to function spaces (clause (R2)) and accumulates defining predicates, as shown in clause (R4)—here, the substitution $\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m$ stems from an instance of the defined type, $(\alpha_1, \ldots, \alpha_m)\, k$. Type variables and declared types are treated as black boxes, so REL_i is vacuously true for them, just like for the built-in types bool and ind (clauses (R1) and (R3)). Note that, while (H2) refers to declared or built-in type constructors, (R3) only refers to declared ones—it explicitly excludes Σ_{init}.

As discussed in Section 4.1, UNF_i treats type variables and declared constants in a guarded fashion (clauses (U1) and (U3)), and distributes over application and abstraction (clauses (U5) and (U6)). Moreover, UNF_i merely calls HOST_i for built-in constants (clause (U2)). Finally, UNF_i unfolds the definitions of defined constants, as shown in clause (U4). In that clause, c_τ and $\rho\restriction_{\mathsf{TV}(c_\tau)}$ (the restriction of ρ to $\mathsf{TV}(c_\tau)$) are uniquely determined by c_σ; and since $\mathsf{TV}(t) \subseteq \mathsf{TV}(c_\sigma)$ (by Def. 1), it follows that $t[\rho]$ is also uniquely determined by c_σ.

Obviously, these functions can reach their purpose only if they are total functions. i.e., their recursive evaluation process terminates for all inputs. This is what we prove in the next subsection.

Assuming totality, we have all the prerequisites to formulate meta-safety. We let UNF be UNF_n, the function that unfolds all definitions in $D = D_n$, and Δ be Δ^n, the signature collecting all the declared items in Σ.

Definition 6. D is said to be a *meta-safe extension of HOL-with-declarations* if, for all $\varphi \in \mathsf{Fmla}_\Delta$, it holds that $D \vdash_\Sigma \varphi$ implies $\vdash_\Delta \mathsf{UNF}(\varphi)$.

4.3 Totality of the Translations

The goal of this subsection is to prove:

Prop 7. The following hold:

(1) The function HOST_i is total, i.e., its recursive calls terminate.
(2) The functions REL_i and UNF_i are total, i.e., their mutually recursive calls terminate.

As discussed, these functions combine structural recursion with the unfolding of constant and type definitions. Roughly speaking, the reason why this recursion terminates is the following: The structural calls are clearly terminating, and the unfoldings are terminating thanks to the freshness condition imposed on the HOL definitional theories (Def. 2(1)), which means the new item on the left of the definition is reduced to existing items. But in order to make this rough intuition precise, we will also need to show that the structural calls and the unfoldings do not somehow interfere in a non-terminating manner.

Note that, if freshness is violated, the functions can become non-terminating. For example, a definition $c_\sigma \equiv c_\sigma$ immediately makes UNF_i non-terminating (due to clause (U4)), also leading to the non-termination of REL_i (which depends on UNF_i via clause (R4)); and similarly for type definitions. Of course, freshness is only a sufficient condition for termination. For example, defining $c_{\alpha\ \mathsf{list}}$ in terms of c_α violates freshness, but locally exhibits a form of terminating recursion, since it descends on the constant's type. As we discuss in Section 5, Isabelle/HOL takes advantage of this observation to replace freshness by a weaker condition. We have designed the concepts we use in the following proof of termination, in particular, the definitional dependency relation, to also be relevant later, when we attend to Isabelle/HOL.

To prove (1), we must show that the call graph of HOST_i, namely, the relation $\blacktriangleright\!\!\blacktriangleright_i$ defined by:

$$(\sigma_1, \ldots, \sigma_m)\, k \blacktriangleright\!\!\blacktriangleright_i \sigma_j \quad \text{if } k \in \Sigma^i$$
$$(\sigma_1, \ldots, \sigma_m)\, k \blacktriangleright\!\!\blacktriangleright_i \sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m] \quad \text{if } (\alpha_1, \ldots, \alpha_m)\, k \equiv t \text{ is in } D_i \text{ and } t : \sigma \Rightarrow \mathsf{bool}$$

is terminating. This is easily done by defining a lexicographic order based on the order in which the items were defined, i.e., the indexes of the definitions def_i in which they appear. (More details are given in the appendix.)

To prove (2), we will exhibit a terminating relation $\blacktriangleright_i$ that captures the mutual call graph of REL_i and UNF_i. We take $\blacktriangleright_i$ to be the union $\equiv_i^{\downarrow} \cup \rhd$, where $\equiv_i^{\downarrow}$ and $\rhd$ are defined below. The relation $\rhd$ consists of the structurally recursive calls of REL_i and UNF_i, from clauses (R2), (U1), (U5) and (U6):

$$\sigma_1 \Rightarrow \sigma_2 \rhd \sigma_1 \qquad \sigma_1 \Rightarrow \sigma_2 \rhd \sigma_2 \qquad x_\sigma \rhd \sigma \qquad t_1 t_2 \rhd t_1 \qquad t_1 t_2 \rhd t_2 \qquad \lambda x_\sigma.\, t \rhd t$$

Moreover, $\equiv_i^{\downarrow}$ captures the recursive calls corresponding to defined items, from (R4) and (U4). Given $u, v \in \mathsf{Type}_{\Sigma_i} \cup \mathsf{Term}_{\Sigma_i}$, $u \equiv_i^{\downarrow} v$ states that there exists a definition $u' \equiv v'$ in D_i and a type substitution ρ such that $u = \rho(u')$ and $v = \rho(v')$.

Thus, the totality of REL_i and UNF_i is reduced to the termination of $\blacktriangleright_i$. In order to prove the latter, we will introduce a more basic relation: the dependency relation between non-built-in items

induced by definitions in D_i. We let $\mathsf{Type}^\bullet_{\Sigma_i}$ be the set of Σ_i-types that have a non-built-in type constructor at the top, and $\mathsf{CInst}^\bullet_{\Sigma_i}$ be the set of instances of non-built-in constants. Given any term t, we let $\mathsf{types}^\bullet(t)$ be the set of all types from $\mathsf{Type}^\bullet_{\Sigma_i}$ appearing in t and $\mathsf{cinsts}^\bullet(t)$ be the set of all constant instances from $\mathsf{CInst}^\bullet_{\Sigma_i}$ appearing in t. (The appendix gives the formal definition of these operators.)

Definition 8. The *dependency relation* $\rightsquigarrow_i$ on $\mathsf{Type}^\bullet_{\Sigma_i} \cup \mathsf{CInst}^\bullet_{\Sigma_i}$ is defined as follows: $u \rightsquigarrow_i v$ iff there exists in D_i a definition of the form $u \equiv t$ such that $v \in \mathsf{cinsts}^\bullet(t) \cup \mathsf{types}^\bullet(t)$.

We write $\rightsquigarrow_i^\downarrow$ for the (type-)substitutive closure of $\rightsquigarrow_i$, defined as follows: $u \rightsquigarrow_i^\downarrow v$ iff there exist u', v' and a type substitution ρ such that $u = u'[\rho]$, $v = v'[\rho]$ and $u' \rightsquigarrow_i v'$. Since HOL with definitions is well-known to be consistent, one would expect that definitions cannot introduce infinite (including cyclic) chains of dependencies. This can indeed be proved by a lexicographic argument, again taking advantage of the definitional order:

Lemma 9. The relation $\rightsquigarrow_i^\downarrow$ is terminating.

The next observation connects $\blacktriangleright_i$ and $\rightsquigarrow_i^\downarrow$, via $\rhd^*$ (the transitive closure of $\rhd$):

Lemma 10. If $u, v \in \mathsf{Type}^\bullet_{\Sigma_i} \cup \mathsf{CInst}^\bullet_{\Sigma_i}$ and $u \equiv_i^\downarrow t \rhd^* v$, then $u \rightsquigarrow_i^\downarrow v$

Now we can reduce the termination of $\blacktriangleright_i$ to that of $\rightsquigarrow_i^\downarrow$, hence prove the former:

Lemma 11. The relation $\blacktriangleright_i$ is terminating.

This concludes the proof of Prop. 7.

4.4 Basic Properties of the Translations

As envisioned in Section 4.1, the translations are extensions of each other and preserve type membership:

Lemma 12. Assume $i \leq n - 1$. The following hold:

(1) If $\sigma \in \mathsf{Type}_{\Sigma_i}$, then $\mathsf{HOST}_{i+1}(\sigma) = \mathsf{HOST}_i(\sigma)$
(2) If $\sigma \in \mathsf{Type}_{\Sigma_i}$, then $\mathsf{REL}_{i+1}(\sigma) = \mathsf{REL}_i(\sigma)$.
(3) If $t \in \mathsf{Term}_{\Sigma_i}$, then $\mathsf{UNF}_{i+1}(t) = \mathsf{UNF}_i(t)$.

Lemma 13. If $\sigma \in \mathsf{Type}_{\Sigma_i}$, $t \in \mathsf{Type}_{\Sigma_i}$ and $t : \sigma$, then $\mathsf{REL}_i(\sigma) : \mathsf{HOST}_i(\sigma) \Rightarrow \mathsf{bool}$ and $\mathsf{UNF}_i(t) : \mathsf{HOST}_i(\sigma)$.

For items in the initial signature, the behavior of the translations is either idle (for HOST_i and UNF_i) or trivial (for REL_i):

Lemma 14. The following hold:

(1) If $\sigma \in \mathsf{Type}_{\Sigma_{\mathsf{init}}}$, then $\mathsf{HOST}_i(\sigma) = \sigma$
(2) If $\sigma \in \mathsf{Type}_{\Sigma_{\mathsf{init}}}$, then $\vdash_{\Sigma_{\mathsf{init}}} \mathsf{REL}_i(\sigma) = \lambda x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{true}$
(3) If $t \in \mathsf{Term}_{\Sigma_{\mathsf{init}}}$ and t is well-typed, then $\vdash_{\Sigma_{\mathsf{init}}} \mathsf{UNF}_i(t) = t$

Other easy, but important properties state that the translations do not introduce new variables or type variables and commute with *type* substitution:

Lemma 15. The following hold for all $\sigma \in \mathsf{Type}_{\Sigma_i}$ and $t \in \mathsf{Term}_{\Sigma_i}$:

(1) $\mathsf{TV}(\mathsf{HOST}_i(\sigma)) \subseteq \mathsf{TV}(\sigma)$
(2) $\mathsf{TV}(\mathsf{REL}_i(\sigma)) \subseteq \mathsf{TV}(\sigma)$ and $\mathsf{FV}(\mathsf{REL}_i(\sigma)) = \emptyset$

(3) $\mathsf{TV}(\mathsf{UNF}_i(t)) \subseteq \mathsf{TV}(t)$ and $\mathsf{FV}(\mathsf{UNF}_i(t)) = \{x_{\mathsf{HOST}_i(\sigma)} \mid x_\sigma \in \mathsf{FV}(t)\}$

Lemma 16. The following hold for all $\sigma, \tau \in \mathsf{Type}_{\Sigma_i}$ and $t \in \mathsf{Term}_{\Sigma_i}$:

(1) $\mathsf{HOST}_i(\sigma[\tau/\alpha]) = \mathsf{HOST}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]$
(2) $\mathsf{REL}_i(\sigma[\tau/\alpha]) = \mathsf{REL}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]$
(3) $\mathsf{UNF}_i(t[\tau/\alpha]) = \mathsf{UNF}_i(t)[\mathsf{HOST}_i(\tau)/\alpha]$

4.5 Main Results

We are now ready to finalize the plan set out in Section 4.1. The following facts in Lemma 17 are stated and proved in the delicate order prescribed there. Fact (4) corresponds to part of (F3) (the remaining parts being covered by Lemmas 15 and 16). Moreover, (2) corresponds to (F2), (3) to (F1), and (5) to (MS). Finally, (1) states deducibility of the translated nonemptiness statement, identified in Section 4.1 as an intermediate fact leading from (MS) to (F2).

Lemma 17. Let $i \in \{1, \ldots, n\}$. The following hold for all $\sigma, \tau \in \mathsf{Type}_{\Sigma_i}$, $t, t' \in \mathsf{Term}_{\Sigma_i}$ and $\varphi \in \mathsf{Fmla}_{\Sigma_i}$:

(1) If $\tau \equiv t$ is a type definition in D_i with $t : \sigma \Rightarrow \mathsf{bool}$, then $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\, x \wedge \mathsf{UNF}_i(t)\, x$
(2) $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\, x$
(3) If $t : \sigma$, then $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma)\, \mathsf{UNF}_i(t)$
(4) If $t' : \sigma$, then $\vdash_{\Delta^i} \mathsf{UNF}_i(t[t'/x_\sigma]) = \mathsf{UNF}_i(t)[\mathsf{UNF}_i(t')/x_{\mathsf{HOST}_i(\sigma)}]$
(5) If $D_i \vdash_{\Sigma_i} \varphi$, then $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$

Proof. The facts follow by induction on i. More precisely, let $(\mathrm{j})_i$ denote fact (j) for a given layer i. We prove:

- that $(1)_1$ holds;
- that, for any $i \in \{1, \ldots, n\}$:
  - $(1)_i$ implies $(2)_i$ implies $(3)_i$ implies $(4)_i$;
  - $(2)_i$ and $(4)_i$ imply $(5)_i$;
- that, for any $i \in \{1, \ldots, n-1\}$, $(5)_i$ implies $(1)_{i+1}$.

$(1)_1$: By the well-formedness of D (Def. 2), we have that $t \in \mathsf{Term}_{\Delta^i}$ and $\sigma \in \mathsf{Term}_{\Delta^i}$, hence $\mathsf{HOST}_0(\sigma) = \sigma$, $\vdash_{\Delta^i} \mathsf{REL}_0(\sigma) = \lambda x_\sigma.\ \mathsf{true}$ and $\vdash_{\Delta^i} \mathsf{UNF}_0(t) = t$. From this, we obtain that the fact to be proved is equivalent to $\vdash_{\Delta^i} \exists x_\sigma.\ t\ x$, which is again true by the well-formedness of D.

Next, we fix $i \in \{1, \ldots, n\}$.

$(1)_i$ implies $(2)_i$: Assuming $(1)_i$, we prove $(2)_i$ by structural induction on σ. The only interesting case is when the type is defined, i.e., has a defined type constructor on top (dealt with in clause (R4)). We need to show $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma')}.\ \mathsf{REL}_i(\sigma')\, x \wedge \mathsf{UNF}_i(t')\, x$, where $(\alpha_1, \ldots, \alpha_m)\, k \equiv t$ is in D_i and $t : \sigma \Rightarrow \mathsf{bool}$, $\sigma' = \sigma[(\sigma_j/\alpha_j)_j]$, and $t' = t[(\sigma_j/\alpha_j)_j]$.

By $(1)_i$, we have $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\, x \wedge \mathsf{UNF}_i(t)\, x$. By the type substitution rule (T-Inst) applied m times (once for each $\mathsf{HOST}_i(\sigma_j)/\alpha_j$), we have $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)[(\mathsf{HOST}_i(\sigma_j)/\alpha_j)_j]}.\ \mathsf{REL}_i(\sigma)\,[(\mathsf{HOST}_i(\sigma_j)/\alpha_j)_j]\, x \wedge \mathsf{UNF}_i(t)[(\mathsf{HOST}_i(\sigma_j)/\alpha_j)_j]\, x$. Using Lemma 16 m times (once for each σ_j/α_j),

we obtain $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma[(\sigma_j/\alpha_j)_j])}.\ \mathsf{REL}_i(\sigma[(\sigma_j/\alpha_j)_j])\ x\ \wedge\ \mathsf{UNF}_i(t[(\sigma_j/\alpha_j)_j])\ x$, which implies $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma')}.\ \mathsf{REL}_i(\sigma')\ x\ \wedge\ \mathsf{UNF}_i(t')\ x$, as desired.

$(2)_i$ implies $(3)_i$: Assume $(2)_i$. Then $(3)_i$ follows by rule induction on the definition of typing. For the variable case, we use $(2)_i$ and the Choice axiom, which ensure us that $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma)(\epsilon\ \mathsf{REL}_i(\sigma))$ holds, hence $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma)(\mathsf{UNF}_i(x_\sigma))$ holds.

$(3)_i$ implies $(4)_i$: Assume $(3)_i$. Then $(4)_i$ follows by well-founded induction on t w.r.t. $\blacktriangleright_i$. The only interesting case is in the variable case (clause (U1)), when the variable coincides with the to-be substituted variable x_σ. Thus, $t = x_\tau$. Here, we need to show $\vdash_{\Delta^i} \mathsf{UNF}_i(t') = \mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)\ \mathsf{UNF}_i(t'))\ (\mathsf{UNF}_i(t'))\ (\varepsilon\ \mathsf{REL}_i(\sigma))$. This follows from the fact that, thanks to $(3)_i$ and $t' : \sigma$, we have $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma)\ \mathsf{UNF}_i(t')$.

$(2)_i$ and $(4)_i$ imply $(5)_i$: Assume $(2)_i$ and $(4)_i$. By induction on the definition of HOL deduction ($\vdash$), we prove a slight generalization of $(5)_i$, namely: We assume $\Gamma \cup \{\varphi\} \subseteq \mathsf{Fmla}_{\Sigma_i}$ and $D_i; \Gamma \vdash_{\Sigma_i} \varphi$, and prove $\emptyset; \mathsf{UNF}_i(\Gamma) \vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$. We distinguish different cases, according to the last applied rule in inferring $\Gamma \cup \{\varphi\} \subseteq \mathsf{Fmla}_{\Sigma_i}$:

(Fact): We need to prove $\emptyset; \mathsf{UNF}_i(\Gamma) \vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$, assuming $\varphi \in \mathsf{Ax} \cup D_i$. First, assume $\varphi \in D$. Then $\varphi = u \equiv t \in D_i$. We have two subcases:

(A) u is a constant c_σ. Then $\mathsf{UNF}_i(\varphi)$ is the formula $\mathsf{UNF}_i(c_\sigma) = \mathsf{UNF}_i(t)$. And since $\mathsf{UNF}_i(c_\sigma)$ and $\mathsf{UNF}_i(t)$ are (syntactically) equal, the desired fact follows by the HOL reflexivity rule.

(B) u is a type τ of the form $(\alpha_1, \ldots, \alpha_m)\ k$ and $t : \sigma \Rightarrow \mathsf{bool}$. Then, by the definition of UNF_i and of the $\forall$ and $\exists$ constructs, $\mathsf{UNF}_i(\varphi)$ is equivalent (modulo HOL deduction) to the formula

$$\exists rep_{\mathsf{HOST}_i(\sigma)\Rightarrow\mathsf{HOST}_i(\sigma)}.\ (\forall x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\ x\ \wedge\ \mathsf{UNF}_i(t)\ x \longrightarrow \mathsf{REL}_i(\sigma)\ (rep\ x))$$
$$\wedge$$
$$\forall x_{\mathsf{HOST}(\sigma)}, y_{\mathsf{HOST}(\sigma)}.\ \mathsf{REL}_i(\sigma)\ x \wedge \mathsf{UNF}_i(t)\ x \wedge \mathsf{REL}_i(\sigma)\ y \wedge \mathsf{UNF}_i(t)\ y \wedge rep\ x = rep\ y \longrightarrow x = y$$
$$\wedge$$
$$\forall y_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\ y \longrightarrow (\mathsf{UNF}_i(t)\ y\ \longleftrightarrow\ (\exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\ x\ \wedge\ \mathsf{UNF}_i(t)\ x\ \wedge\ y = rep\ x))$$

where the first conjunct comes from the relativization of $\tau \Rightarrow \sigma$, the second from unfolding $\mathsf{One\_One}_{rep}$, and the third from unfolding $\forall y_\sigma.\ t\ y\ \longleftrightarrow\ (\exists x_\tau.\ y = rep\ x)$ (in Def. 1). This states the following (in a verbose fashion): There exists $rep : \mathsf{HOST}_i(\sigma) \Rightarrow \mathsf{HOST}_i(\sigma)$ which is one-to-one on the intersection of $\mathsf{REL}_i(\sigma)$ and $\mathsf{UNF}_i(t)$ and the image of this intersection through rep is the intersection itself. This is of course deducible in HOL, taking rep as the identity function.

Now, assume $\varphi \in \mathsf{Ax}$. Then $\varphi \in \mathsf{Fmla}_{\Sigma_{\mathsf{init}}}$, hence, by Lemma 14(3), $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi) = \varphi$. And since also $\emptyset; \mathsf{UNF}_i(\Gamma) \vdash_{\Delta^i} \varphi$ is true by (Fact), the desired fact follows using the HOL equality rules.

(Assum): Follows by applying (Assum).

(T-Inst): Courtesy of UNF_i commuting with type substitution (Lemma 16(3)) and preserving freshness (Lemma 15(3)).

(Inst): Courtesy of UNF_i commuting with substitution (point $(4)_i$) and preserving freshness (Lemma 15(3)).

(Beta), (Ext), (ImpI) and (MP): Courtesy of UNF_i commuting with substitution, preserving freshness, and distributing (by definition) over abstractions, applications and implications.

Next, we fix $i \in \{1, \ldots, n-1\}$.

$(5)_i$ implies $(1)_{i+1}$: Assume $(5)_i$ and let σ, t be as in the formulation of $(1)_{i+1}$, namely, $def_{i+1} = \sigma \equiv t$. By the well-formedness of D (Def. 2), we have $D_i \vdash_{\Sigma_i} \exists x_\sigma.\ t\ x$. Applying $(5)_i$, we obtain $\vdash_{\Delta^i} \mathsf{UNF}_i(\exists x_\sigma.\ t\ x)$. By the definition of the $\exists$ quantifier and the definition of UNF_i, the above is equivalent to $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\ x_{\mathsf{HOST}_i(\sigma)}\ \wedge\ \mathsf{UNF}_i(t)\ t'$, where t' is $\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)\ x_{\mathsf{HOST}_i(\sigma)})\ x\ (\varepsilon\ \mathsf{REL}_i(\sigma))$. By the definition of the if-then-else operator, we can replace t' by x. So the above

is further equivalent to $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\ x\ \wedge\ \mathsf{UNF}_i(t)\ x$. By Lemma 12 and the fact that $\Delta^i \subseteq \Delta^{i+1}$, the above implies $\vdash_{\Delta^{i+1}} \exists x_{\mathsf{HOST}_{i+1}(\sigma)}.\ \mathsf{REL}_{i+1}(\sigma)\ x\ \wedge\ \mathsf{UNF}_{i+1}(t)\ x$, as desired. □

Note that, in our sequence of translations, each translation unfolds not only the i'th definition, but all definitions up to the i'th. One might wonder if the proof could not go more smoothly if we unfolded only the i'th definition and worked with a base theory consisting of all previous definitions, on which we maintained an invariant. That would make a more elegant alternative, but we cannot find an invariant apart from "base theory is definitional," which does not seem to help.

As a particular case of Lemma 17(5), we have:

Theorem 18. *D is a meta-safe extension of HOL-with-declarations.*

Thus, we can compile away all the definitions of D, which leaves us with types and terms over the signature Δ containing declarations only. With the definitions out of our way, it remains to show that *declarations* are conservative, which is much easier:

Lemma 19. *If $\varphi \in \mathsf{Fmla}_{\Sigma_{\mathsf{init}}}$ and $\vdash_\Delta \varphi$, then $\vdash_{\Sigma_{\mathsf{init}}} \varphi$.*

Proof. Assume $\vdash_\Delta \varphi$. In the proof tree for this fact, we replace:

(1) all occurrences of any declared constant instance c_σ by a fresh variable x_σ
(2) all occurrences of any declared type constructor k of arity m by a built-in type expression of arity m, e.g., $(\sigma_1, \ldots, \sigma_m)k$ is replaced by $\sigma_1 \Rightarrow \ldots \Rightarrow \sigma_m$

When performing the indicated replacements, all applications of the HOL rules remain valid; in particular, the application of (Fact) remains unchanged, since the underlying theory D is empty and no HOL axiom (in Ax) refers to declared-only constants or types. Hence these replacements yield a valid proof tree. Since φ, being in $\mathsf{Fmla}_{\Sigma_{\mathsf{init}}}$, is not affected by the replacements, this proof tree constitutes a proof of $\vdash_{\Sigma_{\mathsf{init}}} \varphi$. □

Finally, we can prove overall conservativity:

Theorem 20. *D is a conservative extension of initial HOL.*

Proof. Assume $D \vdash_\Sigma \varphi$, where $\varphi \in \mathsf{Fmla}_{\Sigma_{\mathsf{init}}}$. By Theorem 18, we have $\vdash_\Delta \mathsf{UNF}(\varphi)$. Moreover, by Lemma 14(3), we have $\vdash_{\Sigma_{\mathsf{init}}} \mathsf{UNF}(\varphi) = \varphi$, hence, a fortiori, $\vdash_\Delta \mathsf{UNF}(\varphi) = \varphi$. From these two, we obtain $\vdash_\Delta \varphi$. With Lemma 19, we obtain $\vdash_{\Sigma_{\mathsf{init}}} \varphi$, as desired. □

4.6 Abstract Constant Definition Mechanisms

As definitional schemes for constants, we have only looked into the traditional *equational* ones, implemented in most HOL provers. Two non-equational schemes have also been designed [Arthan 2014], and are available in HOL4, HOL Light and ProofPower-HOL: "new specification" and "gen new specification." They allow for more abstract (under)specification of constants.

However, these schemes have been shown not to increase expressiveness: "new specification" can be over-approximated by traditional definitions and the use of the Choice operator, and "gen new specification" is an admissible rule in HOL with "new specification" [Arthan 2014; Kumar et al. 2014]. Hence our results cater for them.

5 CONSERVATIVITY OF ISABELLE/HOL DEFINITIONS

As mentioned in the introduction, Isabelle/HOL allows more flexible constant definitions than HOL, in that it enables ad hoc overloaded definitions. For example, one can declare a polymorphic constant, such as $\leq : \alpha \Rightarrow \alpha$ bool, and at later times (perhaps after some other type and constant definitions and declarations have been performed) define different, non-overlapping instances of it:

$\leq_{\mathsf{nat}}$ as the standard order on natural numbers, $\leq_{\mathsf{bool}}$ as implication, etc. Even recursive overloading is allowed, e.g., one can define $\leq_{\alpha\ \mathsf{list}}$ as the component-wise extension of $\leq_\alpha$ to α list:

$$xs \leq_{\alpha\ \mathsf{list}} ys \quad \equiv \quad \mathsf{length}\ xs = \mathsf{length}\ ys \ \wedge\ (\forall i < \mathsf{length}\ xs.\ xs_i \leq_\alpha ys_i)$$

This means that now constant definitions no longer require the constant to be *fresh*. In fact, we are no longer speaking of constant definitions, but of constant *instance* definitions: The above examples do not define the overall constant $\leq$, but various instances of it, $\leq_{\mathsf{nat}}$, $\leq_{\mathsf{bool}}$ and $\leq_{\mathsf{list}}$.

Definition 21. Given a non-built-in constant c, a type $\sigma \leq \mathsf{tpOf}(c)$ and a closed term $t : \sigma$, we let $c_\sigma \equiv t$ denote the formula $c_\sigma = t$. We call $c_\sigma \equiv t$ a *constant-instance definition* provided $\mathsf{TV}(t) \subseteq \mathsf{TV}(c_\sigma)$.

To compensate for the lack of freshness from constant-instance definitions, the Isabelle/HOL system performs some global syntactic checks, making sure that defined instances do not overlap (i.e., definitions are *orthogonal*) and that the dependency relation $\rightsquigarrow_n$ from Def. 8, terminates [Kunčar 2015; Kunčar and Popescu 2015, 2017a].[4] (Recall that $D = D_n$, hence $\rightsquigarrow_n$ is the dependency induced by D, i.e., by all the considered definitions.) Formally:

Definition 22. An *Isabelle/HOL-well-formed definitional theory* is a set D of type and constant-instance definitions over Σ such that:

- It satisfies all the conditions of Def. 2, except that it is *not* required that, in condition (1.2), c be fresh, i.e., it is *not* required that $c \notin \Sigma^i$
- It is orthogonal: For all constants c, if c_σ and c_τ appear in two definitions in D, then $\sigma \# \tau$
- Its induced dependency relation $\rightsquigarrow_n$ is terminating

We wish to prove meta-safety and conservativity results similar to the ones for traditional HOL. To this end, we fix an Isabelle/HOL-well-formed definitional theory D and look into the results of Section 4 to see what can be reused—as it turns out, quite a lot.

First, the (type-translated) declaration signatures Δ^i and the translation functions HOST_i, REL_i and UNF_i are defined in the same way. The orthogonality assumption in Def. 22 ensures that, in clause (U4) from the definition of UNF_i, the choice of t is unique (whereas before, this was simply ensured by c appearing on the left in at most one definition). The notion of meta-safety is then defined in the same way. Thanks to $\rightsquigarrow_n$ being terminating, all the dependency relations $\rightsquigarrow_i$, which are included in $\rightsquigarrow_n$, are also terminating. Then all the results in Section 4.3 hold, leading to the totality of the translation functions. Furthermore, almost all the lemmas in Section 4.4 go through undisturbed, because they do not need the freshness assumption $c \notin \Sigma^i$.

The only losses are parts of Lemmas 12 (extension of the translations from i to $i + 1$) and 16 (commutation with type substitution), namely, points (2) and (3) of these lemmas—which deal with REL_i and UNF_i. We first look at Lemma 16.

While HOST_i still commutes with substitution, this is no longer the case for REL_i and UNF_i. Essentially, $\mathsf{UNF}_i(\sigma[\tau/\alpha]) = \mathsf{UNF}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]$ now fails because $\mathsf{UNF}_i(\sigma[\tau/\alpha])$ gets to unfold more constant-instance definitions than $\mathsf{UNF}_i(\sigma)$. So the difference is that, for the constant instances $c_{\sigma'}$ occurring in σ that happen to have a definition of one of their instances, say, $c_{\sigma''} \equiv t$ with $\sigma'' \leq \sigma'$, activated by the substitution τ/α (meaning we have $\sigma'[\tau/\alpha] \leq \sigma''$, but $\sigma' \not\leq \sigma''$), $\mathsf{UNF}_i(\sigma[\tau/\alpha])$ will unfold $c_{\sigma'}$ into the corresponding instance of $\mathsf{UNF}(t)$, whereas $\mathsf{UNF}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]$ will replace $c_{\sigma'}$ with $\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma')\ c_{\mathsf{HOST}_i(\sigma')})\ c_{\mathsf{HOST}_i(\sigma')}\ (\varepsilon\ \mathsf{REL}_i(\sigma'))$. (And since REL_i depends recursively on UNF_i, the former will also fail to commute with type substitution.)

[4] These syntactic checks are part of Isabelle/HOL's logical kernel, just like the local checks for standard HOL definitions are part of HOL's kernel.

Example 23. To Example 4's signature, we add a declared constant c of polymorphic type α and a definition of its nat-instance, $c_{\mathsf{nat}} \equiv \mathsf{z}$. We have $\mathsf{UNF}(c_\alpha[\mathsf{nat}/\alpha]) = \mathsf{UNF}(c_{\mathsf{nat}}) = \mathsf{UNF}(\mathsf{z})$, whereas $\mathsf{UNF}(c_\alpha)\,[\mathsf{HOST}(\mathsf{nat})/\alpha] = (\mathsf{if\_t\_e}\ (\mathsf{REL}(\alpha)\ c_{\mathsf{HOST}(\alpha)})\ c_{\mathsf{HOST}(\alpha)}\ (\varepsilon\ \mathsf{REL}(\alpha)))\,[\mathsf{ind}/\alpha] = (\mathsf{if\_t\_e}\ \mathsf{true}\ c_\alpha\ (\varepsilon\ (\lambda x.\ \mathsf{true})))\,[\mathsf{ind}/\alpha] =_{\mathrm{HOL}} c_\alpha[\mathsf{ind}/\alpha] = c_{\mathsf{ind}}$, where we wrote $=_{\mathrm{HOL}}$ for HOL-provable equality (in the current signature). We do not need to evaluate $\mathsf{UNF}(\mathsf{z})$ in order to see that it cannot be equal, not even HOL-provably equal, to c_{ind}. Indeed, the constant c was not even present in the signature when z was defined, so $\mathsf{UNF}(\mathsf{z})$ cannot be connected to c_{ind}.

Fortunately, we can amend this mismatch "after the fact" by replacing $c_{\mathsf{HOST}_i(\sigma'')}$ with $\mathsf{UNF}_i(c_{\sigma''})$ in $\mathsf{UNF}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]$ for all instances $c_{\sigma''}$ (with $\sigma'' \leq \sigma'$) of all defined constant instances $c_{\sigma'}$. In the above example, this means replacing c_{ind} with $\mathsf{UNF}(c_{\mathsf{nat}})$, i.e., with $\mathsf{UNF}(\mathsf{z})$. To express this formally, we define a *constant-instance substitution* to be a function $\gamma : \mathsf{CInst}^\bullet_{\Delta^i} \Rightarrow \mathsf{Term}_{\Delta^i}$ such that, for all $c_\sigma \in \mathsf{CInst}^\bullet_{\Delta^i}$, $\gamma(c_\sigma)$ is a closed term and $\mathsf{TV}(\gamma(c)) \subseteq \mathsf{TV}(c)$—thus assigning a term to any instance of a non-built-in, i.e., declared constant in Δ^i. Using a notation similar to variable substitution, we write $\sigma[[\gamma]]$ and $t[[\gamma]]$ for the effect of performing γ everywhere inside the type σ or the term t.

Lemma 24. There exists a constant-instance substitution γ such that:

(1) $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma[\tau/\alpha]) = \mathsf{REL}_i(\sigma)[\mathsf{HOST}_i(\tau)/\alpha]\,[[\gamma]]$

(2) $\vdash_{\Delta^i} \mathsf{UNF}_i(t[\tau/\alpha]) = \mathsf{UNF}_i(t)[\mathsf{HOST}_i(\tau)/\alpha]\,[[\gamma]]$

Now, the question is whether the partial consolation offered by Lemma 24, a quasi-commutativity property for REL_i and UNF_i, can replace full commutativity towards the central goal in Lemma 17, namely, point (5) (which ensures meta-safety). Answering this will require some proof mining.

The only usage of Lemma 16 was for $(1)_i$ implies $(2)_i$ (which is part of an implication chain leading to $(4)_i$; and both $(2)_i$ and $(4)_i$ are used for $(5)_i$). There, we used Lemma 16 m times to infer $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma')}.\ \mathsf{REL}_i(\sigma')\,x \wedge \mathsf{UNF}_i(t')\,x$ from $\vdash_{\Delta^i} \exists x_{\mathsf{HOST}_i(\sigma)}.\ \mathsf{REL}_i(\sigma)\,x \wedge \mathsf{UNF}_i(t)\,x$. So we actually need a weaker statement:

Lemma 25. If $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$, then $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi[\sigma/\alpha])$.

For Lemma 12, the situation is quite similar to that of Lemma 16. This time, it is not substitution that can enable additional unfoldings, but a newly added instance definition $c_\sigma \equiv t$ at layer $i+1$ for a constant c that already existed at layer i. Moreover, when we look at how we employed Lemma 12 in the proof of our main chain of results in Lemma 17, we discover a similar pattern: We only use that UNF_{i+1} and REL_{i+1} extend UNF_i and REL_i in the proof of $(5)_i$ implies $(1)_{i+1}$, where we needed that deduction at layer $i+1$ is implied by deduction at layer i. By a similar trick as before, this can be proved using a weaker quasi-commutativity property.

Lemma 26. If $\varphi \in \mathsf{Fmla}_{\Sigma_i}$, and $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$, then $\vdash_{\Delta^{i+1}} \mathsf{UNF}_{i+1}(\varphi)$.

Lemma 25 and 26 reflect a concession made to Isabelle/HOL's ad hoc overloading: We can no longer exhibit a precise structural relationship between $\mathsf{UNF}_i(\varphi)$ on the one hand and and $\mathsf{UNF}_i(\varphi[\sigma/\alpha])$ or $\mathsf{UNF}_{i+1}(\varphi)$ on the other, but we can prove that the latter are "at least as deducible as the former." This would not have been possible had we not treated declared constants in a guarded fashion in the UNF_i clause (U3) (see the discussion on page 14).

Thus, we were able to recover Lemma 17's point (5), leading to meta-safety. And since the other ingredients in the proof of Theorem 20 are also available (including Lemma 19, which is independent of the definitional mechanisms), we infer conservativity. We obtained:

Theorem 27. Theorems 18 and 20 still hold if we assume that D is an Isabelle/HOL-well-formed definitional theory.

A Note on Model-Theoretic Conservativity

Let us revisit some of the aspects of model-theoretic conservativity listed in the introduction's Figure 1. The reason why, for Isabelle/HOL constant instance definitions, model-theoretic conservativity over arbitrary base theories fails is the following: Say we add a constant instance definition $c_\sigma \equiv t$ over a base theory Θ_1 with signature Σ_1, such that Σ_1 contains the constant $c : \tau$ (with $\sigma \leq \tau$) but the formulas in Θ_1 do not refer to c_σ or to any instance of c that is non-orthogonal to c_σ. Thus, $\Sigma_2 = \Sigma_1$ and $\Theta_2 = \Theta_1 \cup \{c_\sigma \equiv t\}$. Then we can easily build a standard model of Θ_1 where $c_\sigma \equiv t$ does not hold, implying that it cannot be extended to a model of Θ_2—which contradicts model-theoretic conservativity. (And (proof-theoretic) conservativity fails for a similar reason: $c_\sigma \equiv t$ is provable from Θ_2 but not from Θ_1.)

We also mentioned in the introduction that, for constant definitions over initial HOL, model-theoretic conservativity follows from conservativity. Here is how the argument goes: Say Θ is a conservative extension of initial HOL with a finite collection of constant (instance) definitions. Then Θ proves a formula φ that encodes these definitions as a conjunction of existentially quantified formulas, where the defined constants (or constant instances) become existentially quantified variables of corresponding types. By conservativity, initial HOL also proves φ. Then any standard model M (of initial HOL) satisfies φ, which implies that the desired constants and types can be defined in M, leading to an extension of M to a standard model of Θ—which proves model-theoretic conservativity.

When trying to apply a similar trick for the case of Θ extending initial HOL with constant *and type* definitions, we face the problem that in HOL we are not allowed to quantify existentially over type variables, to account for the defined types in Θ. Instead, we could appeal to the machinery developed in this paper to perform a more direct proof of model-theoretic conservativity. Namely, starting with a standard model (of initial HOL) M, we could build an extension to a standard model of Θ by well-founded recursion on the terminating relations underlying the definitions of HOST, REL and UNF. The necessary types in M would be introduced taking advantage of the fact that $\mathsf{REL}(\sigma) : \mathsf{HOST}(\sigma) \Rightarrow \mathsf{bool}$ and $\mathsf{REL}(\sigma)$ is provably nonempty. Thus, for mixed constant-type definitions over initial HOL, model-theoretic conservativity would follow not from conservativity, but from the machinery we developed to prove conservativity. We leave a rigorous proof of this as future work—until then, we will not haste to declare the problem closed.

In very recent work, Gengelbach and Weber [2017] prove a form of model-theoretic conservativity for Isabelle/HOL over definitional base theories. However, they do not work with standard models, but employ the ground semantics we had developed for proving Isabelle/HOL's consistency [Kunčar and Popescu 2015]. The connection between ground-model conservativity and standard-model conservativity is yet to be understood.

6 CONCLUDING REMARKS

We have resolved an open problem, relevant for the foundation of HOL-based theorem provers, including our favorite one, Isabelle/HOL: We showed that the definitional mechanisms in such provers are meta-safe and conservative over pure HOL, i.e., are truly "definitional." Our result has for HOL a foundational status analogous to strong normalization results for type theory.

Our translations compile away the constant and type definitions, the latter being significantly more problematic due to the lack of HOL infrastructure for unfolding them. In previous work [Kunčar and Popescu 2017a] we address this infrastructure problem by introducing HOLC, an extension of HOL with comprehension/refinement types. HOL type definitions can be naturally unfolded into HOLC types, yielding a HOL to HOLC translation that was sufficient for showing the *consistency* of Isabelle/HOL definitions. However, that translation would be too coarse for the

stronger results we proved here. Indeed, it is not even conservative, due to HOLC being able to perform type definitions inside proof contexts, unlike HOL. We conjecture that the translation becomes conservative if we enrich HOL with the "local typedef" rule we proposed recently as a non-invasive enhancement of HOL [Kunčar and Popescu 2016].

Our relativization predicates perform an encoding of types as terms, which bears a technical resemblance to the intensional type analysis translations for programming languages introduced in [Crary and Weirich 1999; Crary et al. 1998]. However, they map all (built-in) types to terms, essentially by a structurally recursive traversal. On the other hand, we focus on representing HOL's defined refinement-like types only. Our recursion has a "vertical," structural component (reflecting the structure of the host, built-in types), but also a "horizontal" component, given by unfolding the type definitions.

Our statement of meta-safety is calibrated to what we believe is the key desirable property: that definitions can *all* be compiled away, without loss of provability. An even more general statement would involve compiling away *some* definitions $E \subseteq D$ only, and translating any statement involving all definitions into one involving all definitions but those in E.

However, even the formulation of meta-safety seems problematic here: Say we define the polymorphic type $\alpha\ k$ as the subset $\mathsf{if\_t\_e}\ (\mathsf{cardinal}\ \alpha = 3)\ \{\mathsf{true}, \mathsf{false}\}\ \{\mathsf{true}\}$ of bool. Then we define the type l as the subset $\{1, 2, 3\}$ of ind. Stating that l's definition is meta-safe over k's definition would require us, e.g., to find a host type for $l\ k$ without being allowed to unfold k. The only sensible choice for the host would be $\mathsf{ind}\ k$, which is not suitable since $l\ k$ is larger than $\mathsf{ind}\ k$: The former has two elements, whereas the latter has one. This means that we cannot relativize $l\ k$ as a predicate on $\mathsf{ind}\ k$. Abstractly, the problem is that we cannot lift relativization predicates from the types with which $\alpha\ k$ may be instantiated (such as l). If each HOL type constructor had the structure of a relator (endofunctor on the category of sets having relations as morphisms), the lifting would be possible in a canonical way. And most useful types in HOL, e.g., all combinations of inductive and coinductive datatypes and function spaces, are in fact relators [Traytel et al. 2012]. However, typedef *can* introduce (rather strange looking) non-relators: k is an example of a type constructor that cannot be organized as a relator.

Also, if k were merely *declared*, we would not have a problem, since then we could treat it as a black box that renders $\mathsf{ind}\ k$ and $l\ k$ indistinguishable; so we could take the latter's relativization predicate to be vacuously true. In our meta-safety theorems, we employed this trick to cover declarations intermixed with definitions.

Notwithstanding the difficulty with formulating a more general meta-safety, we believe *conservativity* holds more generally, but requires a different proof technique.

A worthwhile future endeavor will be to certify our results on the foundations of HOL-based proof assistants (in this and our previous papers) by formalizing them in a proof assistant. The main difficulty will involve the notion of recursion for syntax with bindings. The state of the art in recursion principles modulo alpha (as in Nominal Logic) only offers *structural* recursion, hence is not applicable to our functions REL and UNF, which need a more general, *well-founded* recursion. So we could either take a low-level approach (such as working with raw, non-quotiented terms and then prove compatibility with alpha), or use these functions as an inspiration to first design and formalize more powerful principles ourselves, e.g., extending the Horn-based approach to recursion for binders and swapping/substitution [Gheri and Popescu 2017; Norrish 2004; Popescu and Gunter 2011]. Another alternative would be to use higher-order abstract syntax (HOAS), as implemented in Twelf [Pfenning and Schürmann 1999] or Beluga [Pientka and Dunfield 2010]—but this would still leave behind an informal residuum: a pen-and-paper proof of adequacy. We will also explore the possibility to deploy "HOAS on top of FOAS" [Popescu et al. 2010], a framework that enables HOAS while also formalizing adequacy (in the Isabelle/HOL prover).

ACKNOWLEDGMENTS

We thank Rob Arthan, Jasmin Blanchette, Roger Bishop Jones, Ramana Kumar, Tobias Nipkow, Larry Paulson, Dmitriy Traytel, Makarius Wenzel and the members of the Isabelle and HOL mailing lists for inspiring discussions about the logical foundations of theorem proving. We thank Conor McBride and the anonymous reviewers for very useful comments and suggestions which led to the improvement of the paper's presentation. We gratefully acknowledge support from EPSRC through the grant "Verification of Web-based Systems (VOWS)" (EP/N019547/1) and from DFG through the grant "Security Type Systems and Deduction" (Ni 491/13-3) in the priority program "RS^3 – Reliably Secure Software Systems" (SPP 1496).

REFERENCES

Andreas Abel, Thierry Coquand, and Peter Dybjer. 2007. Normalization by Evaluation for Martin-Lof Type Theory with Typed Equality Judgements. In *LICS*. 3–12.

Mark Adams. 2010. Introducing HOL Zero (Extended Abstract). In *ICMS '10*. Springer.

Wolfgang Ahrendt, Bernhard Beckert, Richard Bubel, Reiner Hähnle, Peter H. Schmitt, and Mattias Ulbrich (Eds.). 2016. *Deductive Software Verification - The KeY Book - From Theory to Practice*. Springer.

Thorsten Altenkirch. 1993. Proving Strong Normalization of CC by Modifying Realizability Semantics. In *TYPES*. 3–18.

Rob Arthan. 2014. "HOL Constant Definition Done Right". In *ITP*. 531–536.

R. D. Arthan. 2004. Some Mathematical Case Studies in ProofPower–HOL. In *TPHOLs*.

Andrea Asperti, Wilmer Ricciotti, Claudio Sacerdoti Coen, and Enrico Tassi. 2011. The Matita Interactive Theorem Prover. In *CADE*. 64–69.

Bruno Barras. 2010. Sets in Coq, Coq in Sets. *Journal of Formalized Reasoning* 3, 1 (2010).

Yves Bertot and Pierre Casteran. 2004. *Interactive Theorem Proving and Program Development. Coq'Art: The Calculus of Inductive Constructions*. Springer.

Jasmin Christian Blanchette, Johannes Hölzl, Andreas Lochbihler, Lorenz Panny, Andrei Popescu, and Dmitriy Traytel. 2014. Truly Modular (Co)datatypes for Isabelle/HOL. In *ITP*, Vol. 8558. 93–110.

Ana Bove, Peter Dybjer, and Ulf Norell. 2009. A Brief Overview of Agda—A Functional Language with Dependent Types. In *TPHOLs*.

Alonzo Church. 1940. A Formulation of the Simple Theory of Types. *The Journal of Symbolic Logic* 5, 2 (1940), 56–68.

Thierry Coquand, Jean Gallier, and Le Chesnay Cedex. 1990. A Proof of Strong Normalization For the Theory of Constructions Using a Kripke-Like Interpretation. In *Workshop on Logical Frameworks*.

Thierry Coquand and Arnaud Spiwack. 2006. A Proof of Strong Normalisation using Domain Theory. In *LICS*. 307–316.

Karl Crary and Stephanie Weirich. 1999. Flexible Type Analysis. In *ICFP*. 233–248.

Karl Crary, Stephanie Weirich, and J. Gregory Morrisett. 1998. Intensional Polymorphism in Type-Erasure Semantics. In *ICFP*. 301–312.

Javier Esparza, Peter Lammich, René Neumann, Tobias Nipkow, Alexander Schimpf, and Jan-Georg Smaus. 2013. A Fully Verified Executable LTL Model Checker. In *CAV*. 463–478.

Arve Gengelbach and Tjark Weber. 2017. Model-Theoretic Conservative Extension for Definitional Theories. (2017). Pre-proceedings of LSFA 2017. Available at http://lsfa2017.cic.unb.br/LSFA2017.pdf.

J.H. Geuvers. 1993. *Logics and Type systems*. Ph.D. Dissertation. University of Nijmegen.

Lorenzo Gheri and Andrei Popescu. 2017. A Formalized General Theory of Syntax with Bindings. In *ITP*. 241–261.

M. J. C. Gordon and T. F. Melham (Eds.). 1993. *Introduction to HOL: A Theorem Proving Environment for Higher Order Logic*. Cambridge University Press.

Florian Haftmann and Makarius Wenzel. 2006. Constructive Type Classes in Isabelle.. In *TYPES*. 160–174.

John Harrison. 1996. HOL Light: A Tutorial Introduction. In *FMCAD*. Springer.

John Harrison. 2006. Towards self-verification of HOL Light. In *IJCAR*. Springer.

John Harrison. 2009. HOL Light: An Overview. In *TPHOLs*. 60–66.

Leon Henkin. 1949. The Completeness of the First-Order Functional Calculus. *J. Symbolic Logic* 14, 3 (09 1949), 159–166.

Isabelle. 2016. The Isabelle Library. (2016). https://isabelle.in.tum.de/dist/library/HOL/index.html.

Matt Kaufmann, Panagiotis Manolios, and J Strother Moore. 2000. *Computer-Aided Reasoning: An Approach*. Kluwer Academic Publishers.

Gerwin Klein, June Andronick, Kevin Elphinstone, Gernot Heiser, David Cock, Philip Derrin, Dhammika Elkaduwe, Kai Engelhardt, Rafal Kolanski, Michael Norrish, Thomas Sewell, Harvey Tuch, and Simon Winwood. 2010. seL4: formal verification of an operating-system kernel. *Commun. ACM* 53, 6 (2010), 107–115.

Gerwin Klein, Tobias Nipkow, Larry Paulson, and René Thiemann (eds.). 2016. Isabelle's Archive of Formal Proofs. (2016).

Alexander Krauss. 2009. *Automating recursive definitions and termination proofs in higher-order logic*. Ph.D. Dissertation. Technical University Munich.

Ramana Kumar, Rob Arthan, Magnus O. Myreen, and Scott Owens. 2014. HOL with Definitions: Semantics, Soundness, and a Verified Implementation. In *ITP*. 308–324.

Ramana Kumar, Rob Arthan, Magnus O. Myreen, and Scott Owens. 2016. Self-Formalisation of Higher-Order Logic - Semantics, Soundness, and a Verified Implementation. *J. Autom. Reasoning* 56, 3 (2016), 221–259.

Ondřej Kunčar. 2015. Correctness of Isabelle's Cyclicity Checker: Implementability of Overloading in Proof Assistants. In *CPP*. 85–94.

Ondřej Kunčar and Andrei Popescu. 2015. A Consistent Foundation for Isabelle/HOL. In *ITP*. 234–252.

Ondřej Kunčar and Andrei Popescu. 2016. From Types To Sets By Local Type Definitions in Higher-Order Logic. In *ITP*. 200–218.

Ondřej Kunčar and Andrei Popescu. 2017a. Comprehending Isabelle/HOL's Consistency. In *ESOP*. 724–749.

Ondřej Kunčar and Andrei Popescu. 2017b. Safety and Conservativity of Definitions in HOL and Isabelle/HOL: Isabelle/HOL Implementation. (2017). http://www21.in.tum.de/~kuncar/documents/unf.html.

K. Rustan M. Leino. 2010. Dafny: An Automatic Program Verifier for Functional Correctness. In *LPAR (Dakar)*. 348–370.

Andreas Lochbihler. 2010. Verifying a Compiler for Java Threads. In *ESOP*. 427–447.

Thomas F. Melham. 1989. Automating Recursive Type Definitions in Higher Order Logic. In *Current Trends in Hardware Verification and Automated Theorem Proving*. 341–386.

Magnus O. Myreen and Jared Davis. 2014. The Reflective Milawa Theorem Prover Is Sound - (Down to the Machine Code That Runs It). In *ITP*. 421–436.

Tobias Nipkow and Gerwin Klein. 2014. *Concrete Semantics - With Isabelle/HOL*. Springer.

Tobias Nipkow, Lawrence Paulson, and Markus Wenzel. 2002. *Isabelle/HOL — A Proof Assistant for Higher-Order Logic*. LNCS, Vol. 2283. Springer.

Tobias Nipkow and Gregor Snelting. 1991. Type Classes and Overloading Resolution via Order-Sorted Unification. In *Functional Programming Languages and Computer Architecture*.

Michael Norrish. 2004. Recursive Function Definition for Types with Binders. In *TPHOLs*. 241–256.

Steven Obua. 2006. Checking Conservativity of Overloaded Definitions in Higher-Order Logic.. In *RTA*. 212–226.

Sam Owre and Natarajan Shankar. 1999. The Formal Semantics of PVS. (1999). SRI technical report. http://www.csl.sri.com/papers/csl-97-2/.

Lawrence C. Paulson. 1990. A formulation of the simple theory of types (for Isabelle). In *COLOG-88*. 246–274.

Lawrence C. Paulson. 2010. Three Years of Experience with Sledgehammer, a Practical Link between Automatic and Interactive Theorem Provers. In *PAAR*. 1–10.

Frank Pfenning and Carsten Schürmann. 1999. System Description: Twelf - A Meta-Logical Framework for Deductive Systems. In *CADE*. 202–206.

Brigitte Pientka and Joshua Dunfield. 2010. Beluga: A Framework for Programming and Reasoning with Deductive Systems (System Description). In *IJCAR*. 15–21.

A. Pitts. 1993. *Introduction to HOL: A Theorem Proving Environment for Higher Order Logic*, Chapter The HOL Logic, 191–232. In Gordon and Melham [Gordon and Melham 1993].

Andrei Popescu and Elsa L. Gunter. 2011. Recursion principles for syntax with bindings and substitution. In *ICFP*. 346–358.

Andrei Popescu, Elsa L. Gunter, and Christopher J. Osborn. 2010. Strong Normalization for System F by HOAS on Top of FOAS. In *LICS*. 31–40.

John C. Reynolds. 1983. Types, Abstraction and Parametric Polymorphism. In *IFIP Congress*. 513–523.

Donald Sannella and Andrzej Tarlecki. 2012. *Foundations of Algebraic Specification and Formal Software Development*. Springer. I–XVI, 1–581 pages.

Konrad Slind and Michael Norrish. 2008. "A Brief Overview of HOL4". In *TPHOLs*. 28–32.

Dmitriy Traytel, Andrei Popescu, and Jasmin Christian Blanchette. 2012. Foundational, Compositional (Co)datatypes for Higher-Order Logic: Category Theory Applied to Theorem Proving. In *LICS*. 596–605.

D. A. Turner. 2004. Total Functional Programming. *J. UCS* 10, 7 (2004), 751–768.

Markus Wenzel. 1997. Type Classes and Overloading in Higher-Order Logic.. In *TPHOLs*. 307–322.

Markus Wenzel. 1999. Isar - A Generic Interpretative Approach to Readable Formal Proof Documents. In *TPHOLs*. 167–184.

Makarius Wenzel. 2014. System description: Isabelle/jEdit in 2014. In *UITP*. 84–94.

Freek Wiedijk. 2009. Stateless HOL. In *TYPES*. 47–61.

Burkhart Wolff. 2015. Isabelle Foundation & Certification. (2015). Archived at https://lists.cam.ac.uk/pipermail/cl-isabelle-users/2015-September/thread.html.

A MORE DETAILS ON HOL

It is well-known (and easy to prove) that substitution respects typing:

Lemma 28. If $t : \sigma$, then $t[\rho] : \sigma[\rho]$.

When writing concrete terms or formulas, we take the following conventions:

- We omit redundantly indicating the types of the variables, e.g., we shall write $\lambda x_\sigma .\ x$ instead of $\lambda x_\sigma .\ x_\sigma$.
- We omit redundantly indicating the types of the variables and constants in terms if they can be inferred by typing rules, e.g., we shall write $\lambda x.\ (y_{\sigma\Rightarrow\tau}\ x)$ instead of $\lambda x_\sigma .\ (y_{\sigma\Rightarrow\tau}\ x)$ or $\varepsilon(\lambda x_\sigma .\ P\ x)$ instead of $\varepsilon_{(\sigma\Rightarrow\mathsf{bool})\Rightarrow\sigma}(\lambda x_\sigma .\ P_{\sigma\Rightarrow\mathsf{bool}}\ x)$.
- We write $\lambda x_\sigma\ y_\tau .\ t$ instead of $\lambda x_\sigma .\ \lambda y_\tau .\ t$
- We apply the constant = in an infix manner, i.e., we shall write $t_\sigma = s$ instead of $= t_\sigma\ s$.

The formula connectives and quantifiers are defined as abbreviations in the usual way, starting from the implication and equality primitives:

$$
\begin{array}{lcl}
\mathsf{true} & = & (\lambda x_{\mathsf{bool}} .\ x) = (\lambda x_{\mathsf{bool}} .\ x) \\
\mathsf{all}_\sigma & = & \lambda p_{\sigma\Rightarrow\mathsf{bool}} .\ p = (\lambda x_\sigma .\ \mathsf{true}) \\
\multicolumn{3}{l}{\text{(in what follows, we write } \forall x_\sigma .\ t \text{ instead of } \mathsf{all}_\sigma\ (\lambda x_\sigma .\ t))} \\
\mathsf{and} & = & \lambda p_{\mathsf{bool}}\ q_{\mathsf{bool}} .\ \forall f_{\mathsf{bool}\Rightarrow\mathsf{bool}\Rightarrow\mathsf{bool}} .\ f\ p\ q = f\ \mathsf{true}\ \mathsf{true} \\
\mathsf{implies} & = & \lambda p_{\mathsf{bool}}\ q_{\mathsf{bool}} .\ \mathsf{and}\ p\ q = p \\
\multicolumn{3}{l}{\text{(in what follows, we write } p \longrightarrow q \text{ instead of implies } p\ q)} \\
\mathsf{ex}_\sigma & = & \lambda p_{\sigma\Rightarrow\mathsf{bool}} .\ \forall q_{\mathsf{bool}} .\ (\forall x_\sigma .\ p\ x \longrightarrow q) \longrightarrow q \\
\mathsf{false} & = & \forall p_{\mathsf{bool}} .\ p \\
\mathsf{not} & = & \lambda p_{\mathsf{bool}} .\ p \longrightarrow \mathsf{false} \\
\mathsf{or} & = & \lambda p_{\mathsf{bool}}\ q_{\mathsf{bool}} .\ \forall r_{\mathsf{bool}} .\ (p \longrightarrow r) \longrightarrow ((q \longrightarrow r) \longrightarrow r)
\end{array}
$$

As customary, we write:

- $\exists x_\sigma .\ t$ instead of $\mathsf{ex}_\sigma\ (\lambda x_\sigma .\ t)$
- $\neg\ \varphi$ instead of $\mathsf{not}\ \varphi$
- $\varphi \wedge \chi$ instead of $\mathsf{and}\ \varphi\ \chi$
- $\varphi \vee \chi$ instead of $\mathsf{or}\ \varphi\ \chi$

The HOL axioms, forming the set Ax, are the following:

- Equality Axioms:
  refl: $x_\alpha = x$
  subst: $x_\alpha = y \longrightarrow p\ x \longrightarrow p\ y$
- Infinity Axioms:
  suc_inj: $\mathsf{suc}\ x = \mathsf{suc}\ y \longrightarrow x = y$
  suc_not_zero: $\neg\ \mathsf{suc}\ x = \mathsf{zero}$
- Choice:
  some_intro: $p_{\alpha\Rightarrow\mathsf{bool}}\ x \longrightarrow p\ (\varepsilon\ p)$

Above, refl and subst axiomatize equality. suc_inj and suc_not_zero ensure that ind is an infinite type. some_intro regulates the behavior of the Hilbert Choice operator. The primciple of excluded middle, $(b = \mathsf{true}) \vee (b = \mathsf{false})$, follows from the axiom of choice—this makes HOL a classical logic.

B DETAILED DEFINITION OF THE OPERATORS USED IN THE DEPENDENCY RELATION

Note that the types$^\bullet$ operator is overloaded for types and terms.

$$\begin{aligned}
\mathsf{types}^\bullet(\alpha) &= \{\alpha\} \\
\mathsf{types}^\bullet(\mathsf{bool}) &= \emptyset \\
\mathsf{types}^\bullet(\mathsf{ind}) &= \emptyset \\
\mathsf{types}^\bullet(\sigma_1 \Rightarrow \sigma_2) &= \mathsf{types}^\bullet(\sigma_1) \cup \mathsf{types}^\bullet(\sigma_2) \\
\mathsf{types}^\bullet(\overline{\sigma}\ k) &= \{\overline{\sigma}\ k\},\ \text{if } k \neq \Rightarrow, \mathsf{bool}, \mathsf{ind}
\end{aligned}$$

$$\begin{aligned}
\mathsf{types}^\bullet(x_\sigma) &= \mathsf{types}^\bullet(\sigma) \\
\mathsf{types}^\bullet(c_\sigma) &= \mathsf{types}^\bullet(\sigma) \\
\mathsf{types}^\bullet(t_1\ t_2) &= \mathsf{types}^\bullet(t_1) \cup \mathsf{types}^\bullet(t_2) \\
\mathsf{types}^\bullet(\lambda x_\sigma.\ t) &= \mathsf{types}^\bullet(\sigma) \cup \mathsf{types}^\bullet(t)
\end{aligned}$$

$$\begin{aligned}
\mathsf{cinsts}^\bullet(x_\sigma) &= \emptyset \\
\mathsf{cinsts}^\bullet(c_\sigma) &= \begin{cases} \{c_\sigma\} & \text{if } c_\sigma \in \mathsf{CInst}^\bullet \\ \emptyset & \text{otherwise} \end{cases} \\
\mathsf{cinsts}^\bullet(t_1\ t_2) &= \mathsf{cinsts}^\bullet(t_1) \cup \mathsf{cinsts}^\bullet(t_2) \\
\mathsf{cinsts}^\bullet(\lambda x_\sigma.\ t) &= \mathsf{cinsts}^\bullet(t)
\end{aligned}$$

C PROOF SKETCHES

In the proofs, we use several induction schemas, fit for the purpose:

- *Well-founded induction* on types and/or terms with respect to one of the (known to be terminating) relations $\blacktriangleright\!\!\blacktriangleright_i$ and $\blacktriangleright_i$: Given u, we can assume the property holds for all items u' such that $u \blacktriangleright\!\!\blacktriangleright_i u'$ (or $u \blacktriangleright_i u'$) and need to prove it for u. So whenever we indicate a proof by well-founded induction, we will implicitly refer to one of these two, namely, to the first when proving something about HOST_i and to the second when proving something about REL_i and/or UNF_i.
- *Structural induction* on types and/or terms: Given u, we can assume the property holds for all immediate subtypes/subterms of u and need to prove it for u.
- *Rule induction* with respect to the definition of typing or the definition of HOL deduction: To conclude that typing or deduction implies a property, we prove that the property is closed under the rules defining typing or deduction.

In all these schemas, (IH) denotes the induction hypothesis.

Proof of point (1) of Prop. 7. We first define, for any type constructor $k \in \Sigma_i$, the operator $\mathsf{depth}_k : \mathsf{Type}_{\Sigma_i} \Rightarrow \mathbb{N}$ to return, for any type, the length of the longest nesting of k's appearing in it, namely:

$$\begin{aligned}
&\mathsf{depth}_k(\alpha) = 0 \\
&\mathsf{depth}_k((\sigma_1, \ldots, \sigma_m)\ l) = \\
&\quad \begin{cases} 1 + \max\{\mathsf{depth}_k(\sigma_i) \mid i \in \{1, \ldots, m\}\} & \text{if } l = k \\ \max\{\mathsf{depth}_k(\sigma_i) \mid i \in \{1, \ldots, m\}\} & \text{if } l \neq k \end{cases}
\end{aligned}$$

Let K^i and K_i be the sets of type constructors of $\Sigma_1 \cup \bigcup_{i'=1}^{i} \Sigma^{i'} \setminus \Sigma_{i'-1}$ and Σ_i, respectively. Note that $\mathsf{K}^i \subseteq \mathsf{K}_i$ and that $\mathsf{K}_i \setminus \mathsf{K}^i$ contains the defined type constructors, whereas K^i contains the declared and built-in ones (up to moment i). We chose an arbitrary total order $\succ$ on K^i, and then extend it to a homonymous total order on K_i, as follows:

- If $k \in \mathsf{K}^i$ and $l \in \mathsf{K}_i$, then $l \succ k$
- If $k_1, k_2 \in \mathsf{K}_i$, then $k_1 \succ k_i$ if k_1 was introduced later than k_2, i.e., if the unique $j_1 \leq i$ such that k_1 appears in the lefthand side of def_{j_1} is greater than the unique $j_2 \leq i$ such that k_2 appears on the lefthand side of def_{j_2}

Since K_i is finite, it has the form $\{k_1, \ldots, k_p\}$ with $k_1 \succ \ldots \succ k_p$. We define the measure meas : $\mathsf{Type}_{\Sigma_i} \to \mathbb{N}^p$ by $\mathsf{meas}(\sigma) = (\mathsf{depth}_{k_1}, \ldots, \mathsf{depth}_{k_p})$. Finally, we note that $\blacktriangleright_i$ decreases this measure w.r.t. the lexicographic order on $\mathbb{N}^p$ (which ensures its termination). Indeed, we consider the two cases in the definition of $\blacktriangleright_i$:

- In the first case (given by recursive clause (U3)), all depth_{k_j} remain the same or decrease, and depth_k decreases by 1
- In the second case (given by recursive clause (U4)), we know from the well-foundedness of D that σ only contains type constructors l with $k \succ l$. Therefore, we have:

  $$\mathsf{depth}_k((\sigma_1, \ldots, \sigma_m)\, k) = 1 + \max\{\mathsf{depth}_k(\sigma_j) \mid j \in \{1, \ldots, m\}\} >$$
  $$\max\{\mathsf{depth}_l(\sigma_j) \mid j \in \{1, \ldots, m\} \ \wedge\ \alpha_i \in \mathsf{TV}(\sigma)\} = \mathsf{depth}_k(\sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m])$$

  Moreover, for any l such that $l \succ k$, we have:

  $$\mathsf{depth}_l((\sigma_1, \ldots, \sigma_m)\, k) = \max\{\mathsf{depth}_l(\sigma_j) \mid j \in \{1, \ldots, m\}\} \geq$$
  $$\max\{\mathsf{depth}_l(\sigma_j) \mid j \in \{1, \ldots, m\} \ \wedge\ \alpha_i \in \mathsf{TV}(\sigma)\} = \mathsf{depth}_l(\sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m])$$

  Thus, depth_k decreases strictly and, for $l \succ k$, depth_l remains the same or decreases; this ensures that meas decreases. □

Proof of Lemma 9. We proceed similarly to the proof of termination for the call graph of HOST_i, but considering Σ_i-constants in addition to Σ_i-type constructors. Similarly to there, for each $e \in \mathsf{K}_i \cup \mathsf{Const}_i$, we define $\mathsf{depth}_e : \mathsf{Type}_{\Sigma_i} \cup \mathsf{Term}_{\Sigma_i} \Rightarrow \mathbb{N}$, the u-depth of a type or term, to the length of the longest nesting of u's appearing in it. We similarly order the items in $\mathsf{K}_i \cup \mathsf{Const}_i$ by a relation $\succ$ asking that all defined items are greater than all non-defined ones and a later defined item is greater than an earlier defined one. Assuming $\mathsf{K}_i \cup \mathsf{Const}_i$ has the form $\{e_1, \ldots, e_p\}$ with $e_1 \succ \ldots \succ e_p$, we define the measure meas : $\mathsf{Type}_{\Sigma_i} \cup \mathsf{Term}_{\Sigma_i} \to \mathbb{N}^p$ by $\mathsf{meas}(v) = (\mathsf{depth}_{e_1}(v), \ldots, \mathsf{depth}_{e_p}(v))$.

We show that meas decreases with $\rightsquigarrow_i^{\downarrow}$ w.r.t. the lexicographic order on $\mathbb{N}^p$ (which makes $\rightsquigarrow_i^{\downarrow}$ terminating). Assume $u \rightsquigarrow_i^{\downarrow} v$. Then there exists u', v', ρ such that $u = u'[\rho]$, $v = v'[\rho]$ and $u' \rightsquigarrow_i v'$. Then u' is either a type of the form $(\alpha_1, \ldots, \alpha_m)k$ with $k \in \mathsf{K}_i$, or a constant instance c_σ; meaning u is either $(\rho(\alpha_1), \ldots, \rho(\alpha_m))k$ or $c_{\sigma[\rho]}$. We let e denote either k or c. In both cases, we have $v' \in \mathsf{types}^\bullet(t) \cup \mathsf{cinsts}^\bullet(t)$ for some $t \in \mathsf{Term}_{\Sigma_i}$. Hence $v \in \mathsf{types}^\bullet(t[\rho]) \cup \mathsf{cinsts}^\bullet(t[\rho])$. By the well-formedness of D, e is greater than all the type constructors and constants in t (w.r.t. $\succ$). Then $\mathsf{depth}_e(u) > \mathsf{depth}_e(v)$ and, for all e' such that $e' \succ e$, $\mathsf{depth}_{e'}(u) \geq \mathsf{depth}_{e'}(v)$. This ensures $\mathsf{meas}(u) > \mathsf{meas}(v)$. □

Proof of Lemma 10. By routine structural induction on t. □

Proof of Lemma 11. Let us assume by absurd that $\blacktriangleright_i$ does not terminate. Then there exists an infinite sequence $(w_p)_{p\in\mathbb{N}}$ such that $w_p \blacktriangleright_i w_{p+1}$ for all p. Since $\blacktriangleright_i$ is defined as $\equiv_i^{\downarrow} \cup \rhd$ and $\rhd$ clearly terminates, there must exist an infinite subsequence $(w_{p_j})_{j\in\mathbb{N}}$ such that $w_{p_j} \equiv_i^{\downarrow} w_{p_j+1} \rhd^* w_{p_{j+1}}$ for all j. Since from the definition of $\equiv_i^{\downarrow}$ we have $w_{p_j} \in \mathsf{Type}^\bullet_{\Sigma_i} \cup \mathsf{CInst}^\bullet_{\Sigma_i}$, we obtain from Lemma 10 that $w_{p_j} \rightsquigarrow_i^{\downarrow} w_{p_{j+1}}$ for all p. This contradicts the termination of $\rightsquigarrow_i^{\downarrow}$. □

Proof of Lemma 12. (1): By an easy well-founded induction on σ w.r.t. $\blacktriangleright_i$, distinguishing between the different cases in the definition of HOST_i and HOST_{i+1}. The definitions are identical for the two functions, and for the defined type case (clause (H3)), we know that k is in Σ_i, ensuring that

$(\sigma_1, \ldots, \sigma_m)k \equiv t$ is in D_i.
(2) and (3): Similar to (1), by an easy well-founded induction on σ and t w.r.t. $\blacktriangleright_i$. □

Proof of Lemma 13. By well-founded induction on σ and t, distinguishing between the different cases in the definitions of REL_i and UNF_i. The proof is routine. We only show the two slightly less obvious cases, where we employ the local notations used in the definitions (e.g., σ', t'):
The defined type case for REL_i (clause (R4)): We know that $(\sigma_1, \ldots, \sigma_m)k \blacktriangleright_i \sigma'$ and also that $(\sigma_1, \ldots, \sigma_m)k \blacktriangleright_i t'$. Moreover, from $t : \sigma$ we obtain $t' : \sigma'$. Hence, by (IH), we have $\mathsf{REL}_i(\sigma') : \mathsf{HOST}(\sigma') \Rightarrow \mathsf{bool}$ and $\mathsf{UNF}_i(t') : \mathsf{HOST}(\sigma')$. From this, the definition of REL_i and the HOL typing rules, we obtain $\mathsf{REL}_i((\sigma_1, \ldots, \sigma_m)k) : \mathsf{HOST}_i(\sigma') \Rightarrow \mathsf{bool}$. Finally, from the definition of HOST_i we have $\mathsf{HOST}_i(\sigma') = \mathsf{HOST}_i((\sigma_1, \ldots, \sigma_m)k)$, hence $\mathsf{REL}_i((\sigma_1, \ldots, \sigma_m)k) : \mathsf{HOST}_i((\sigma_1, \ldots, \sigma_m)k) \Rightarrow \mathsf{bool}$, as desired.
The defined constant case for UNF_i (clause (U4)): We know that $c_\sigma \blacktriangleright_i t[\rho]$. Moreover, since $\sigma = \tau[\rho]$ and $t : \tau$, by Lemma 28 we have that $t[\rho] : \sigma$. By (IH), we have $\mathsf{UNF}(t[\rho]) : \mathsf{HOST}(\sigma)$; and since $\mathsf{UNF}_i(c_\sigma) = \mathsf{UNF}_i(t[\rho])$, we obtain $\mathsf{UNF}_i(c_\sigma) : \mathsf{HOST}(\sigma)$, as desired. □

Proof of Lemma 14. (1) and (2): Immediate by structural induction on σ.
(3): Immediate by structural induction on t. For the variable case, we use point (2) to obtain $\vdash_{\Sigma_0} \mathsf{UNF}_i(x_\sigma) = x_\sigma$ from the behavior of if-then-else.

(Since Σ_{init} has no defined or declared items, the recursive cases that deal with such items do not occur when applying the translations, and in particular REL_i does not depend recursively of UNF_i. This is why structural induction does the job, so there is no need for the more powerful well-founded induction.) □

Proof of Lemma 15. Immediate well-founded induction, using the property that definitions do not introduce free term variables or type variables. □

Proof of Lemma 16. By routine well-founded induction, using the properties of type substitution. For example: In the defined type cases for HOST_i and REL_i (clauses (H3) and (R4)), we use that, if $\mathsf{TV}(\sigma) \subseteq \{\alpha_1, \ldots, \alpha_m\}$ (as it is guaranteed by Def. 1), $\sigma[\sigma_1/\alpha_1, \ldots, \sigma_m/\alpha_m][\tau/\alpha] = \sigma[(\sigma_1[\tau/\alpha])/\alpha_1, \ldots, (\sigma_m[\tau/\alpha])/\alpha_m]$; in the defined constant case for UNF_i (clause (U4)), we use that $t[\rho][\tau/\alpha] = t[\rho \cdot (\tau/\alpha)]$. (Recall that $\cdot$ is the composition of substitutions.) □

Proof of Lemma 24. We will write $t_1 =_{\Delta^i} t_2$ instead of $\vdash_{\Delta^i} t_1 = t_2$. We define γ to map each $c_{\mathsf{HOST}_i(\sigma)}$ to $\mathsf{UNF}_i(c_\sigma)$. Thanks to Lemma 15(3), γ is indeed a constant-instance substitution. Now, points (1) and (2) follow by well-founded induction on the mutually recursive definitions of REL_i and UNF_i. The only interesting case is that of defined constants (clause (U4) for UNF_i). Assume $\sigma[\tau/\alpha] = \sigma'[\rho]$, such that $c_{\sigma'} \equiv t \in D_i$. We have two cases:

First, assume $\sigma \leq \sigma'$, say, $\sigma = \sigma'[\rho']$ for some ρ'. Then ρ and $\rho' \cdot (\tau/\alpha)$ are equal on $\mathsf{TV}(\sigma')$, a fortiori, on $\mathsf{TV}(t)$. Hence $t[\rho] = t[\rho' \cdot (\tau/\alpha)]$. i.e., $t[\rho] = t[\rho'][\tau/\alpha]$ (*). Both $\mathsf{UNF}_i(c_\sigma[\tau/\alpha])$ and $\mathsf{UNF}_i(c_\sigma)$ will unfold the definitions of their corresponding instances of c, allowing us to infer the desired fact from the induction hypothesis:

$$\begin{aligned}
&\mathsf{UNF}_i(c_\sigma[\tau/\alpha]) = \mathsf{UNF}_i(c_{\sigma[\tau/\alpha]}) = \mathsf{UNF}_i(c_{\sigma'[\rho]}) = (\text{by (U4)}) = \\
&\mathsf{UNF}_i(t[\rho]) = (\text{by (*)}) = \mathsf{UNF}_i(t[\rho'][\tau/\alpha]) =_{\Delta^i} \\
&(\text{by the induction hypothesis}) \\
&\mathsf{UNF}_i(t[\rho'])[\mathsf{HOST}_i(\tau)/\alpha][[\gamma]] = (\text{by (U4)}) = \\
&\mathsf{UNF}_i(c_{\sigma'[\rho']})[\mathsf{HOST}_i(\tau)/\alpha][[\gamma]] = \mathsf{UNF}_i(c_\sigma)[\mathsf{HOST}_i(\tau)/\alpha][[\gamma]]
\end{aligned}$$

Next, assume $\sigma \not\leq \sigma'$. Then only $\mathsf{UNF}_i(c_\sigma[\tau/\alpha])$ unfolds the definition of c'_σ, but γ repairs the mismatch. To ease readability, we will write _[•] instead of _[$\mathsf{HOST}_i(\tau)/\alpha$].

$\mathsf{UNF}_i(c_\sigma[\tau/\alpha]) =_{\Delta^i}$
(since $\vdash_{\Delta^i} \mathsf{REL}_i(\sigma[\tau/\alpha])\ \mathsf{UNF}_i(c_{\sigma[\tau/\alpha]})$ holds)
$\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma[\tau/\alpha])\ \mathsf{UNF}_i(c_{\sigma[\tau/\alpha]}))\ \ \mathsf{UNF}_i(c_{\sigma[\tau/\alpha]})\ \ (\varepsilon\ \mathsf{REL}_i(\sigma[\tau/\alpha])) =$
(by the definition of γ)
$\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma[\tau/\alpha])\ \gamma(c_{\mathsf{HOST}_i(\sigma[\tau/\alpha])}))\ \ \gamma(c_{\mathsf{HOST}_i(\sigma[\tau/\alpha])})\ \ (\varepsilon\ \mathsf{REL}_i(\sigma[\tau/\alpha])) =$
(by the definition of constant substitution)
$\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma[\tau/\alpha])\ c_{\mathsf{HOST}_i(\sigma[\tau/\alpha])}[[\gamma]])\ \ c_{\mathsf{HOST}_i(\sigma[\tau/\alpha])}[[\gamma]]\ \ (\varepsilon\ \mathsf{REL}_i(\sigma[\tau/\alpha])) =$
(by (IH) for REL_i and HOST_i's commutation with substutition)
$\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)[\bullet][[\gamma]]\ c_{\mathsf{HOST}_i(\sigma)}[\bullet][[\gamma]])\ \ c_{\mathsf{HOST}_i(\sigma)}[\bullet][[\gamma]]\ \ (\varepsilon[\bullet][[\gamma]]\ \mathsf{REL}_i(\sigma)[\bullet][[\gamma]]) =$
(by the definition of constant substitution)
$(\mathsf{if\_t\_e}\ (\mathsf{REL}_i(\sigma)\ c_{\mathsf{HOST}_i(\sigma)})\ \ c_{\mathsf{HOST}_i(\sigma)}\ \ (\varepsilon\ \mathsf{REL}_i(\sigma)))\ [\bullet][[\gamma]] =$
(by (U3))
$\mathsf{UNF}_i(c_\sigma)[\bullet][[\gamma]]$ □

Proof of Lemma 25. By Lemma 24(2), we have a constant-instance substitution γ such that $\mathsf{UNF}_i(\varphi[\sigma/\alpha]) = \mathsf{UNF}_i(\varphi)[\mathsf{HOST}_i(\sigma)/\alpha][[\gamma]]$. And since $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)[\mathsf{HOST}_i(\sigma)/\alpha]$ follows from $\vdash_{\Delta^i} \mathsf{UNF}_i(\varphi)$ by the type substitution rule (T-Inst), it would suffice to have the following: For all constant-instance substitutions γ and Δ^i-formulas φ, $\vdash_{\Delta^i} \varphi$ implies $\vdash_{\Delta^i} \varphi[[\gamma]]$. In words, if we substitute some (undefined) constant instances with terms of the same type we do not lose provability. This follows by routine induction on the definition of deduction. □

Proof of Lemma 26. If def_{i+1} is a type definition, then UNF_{i+1} and REL_{i+1} do extend UNF_i and REL_i, so the desired fact follows trivially. Now, assume def_{i+1} is a constant-instance definition $c_\sigma \equiv t$. Similarly to the proof of Lemma 24(2), we obtain a constant-instance substitution γ such that $\mathsf{UNF}_{i+1}(\varphi) = \mathsf{UNF}_i(\varphi)[[\gamma]]$, namely, γ maps each $d_{\mathsf{HOST}(\tau[\rho])}$ to $\mathsf{UNF}_{i+1}(s[\rho])$ where $d_\tau \equiv s$ are the constant definitions in D_{i+1}. (We need to do this replacement to all defined constant instances, not just c_σ, since other definitions from D_i may have already relied on c_σ.) And since, as we have seen, constant-instance substitution preserves deduction, we obtain our desired fact. □