Åqvist's Dyadic Deontic Logic E in HOL*

Christoph Benzmüller[1,2], Ali Farjami[1], and Xavier Parent[1]

[1] University of Luxembourg, Luxembourg
[2] Freie Universität Berlin, Germany
c.benzmueller@gmail.com
farjami110@gmail.com
xavier.parent@uni.lu

Abstract. We devise a shallow semantical embedding of Åqvist's dyadic deontic logic **E** in classical higher-order logic. This embedding is encoded in Isabelle/HOL, which turns this system into a proof assistant for deontic logic reasoning. The experiments with this environment provide evidence that this logic *implementation* fruitfully enables interactive and automated reasoning at the meta-level and the object-level.

Keywords: Dyadic deontic logic **E** · Preference models · Classical higher-order logic · Semantic embedding · Automated reasoning.

1 Introduction

Normative notions such as obligation and permission are the subject of deontic logics [18] and conditional obligations are addressed in so-called *dyadic deontic logic*. A landmark and historically important dyadic deontic logic has been proposed by B. Hansson [21] and Åqvist [3]. This dyadic deontic logic comes with a preference models semantics [24], in which a binary preference relation ranks the possible words in terms of betterness. The framework is immune to the well known paradoxes of *contrary-to-duty* (CTD) reasoning such as Chisholm [14]'s paradox. The class of all preference models, in which no specific constraints are put on the betterness relation, has a known axiomatic characterisation, given by Åqvist's system **E**. (See Parent [24].)

When applied as a meta-logical tool, *simple type theory* [15], aka classical higher-order logic (HOL), can help to better understand semantical issues (of embedded object logics). The syntax and semantics of HOL are well understood [7] and there exist automated proof tools for it; examples include Isabelle/HOL [23] and LEO-II [11]. As mentioned in the *Handbook of Deontic Logic and Normative Systems* [19], the development of computational techniques in deontic logic is still in its infancy.

In this paper we devise an *embedding* of **E** in HOL. This embedding utilizes the *shallow semantical embedding* approach that has been put forward by

* This work has been supported by the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690974. Benzmller received funding from the Volkswagen foundation.

Benzmüller (cf. [6] and references therein) as a pragmatical solution towards universal logic reasoning. This approach uses classical higher-order logic as (universal) meta-logic to specify, in a shallow way, the syntax and semantics of various object logics, in our case system **E**. The embedding has been encoded in Isabelle/HOL to enable experiments in deontic reasoning.

Related work [9] has developed an analogous shallow semantical embedding for the dyadic deontic logic proposed by Carmo and Jones [13]. A core difference concerns the notion of semantics employed in both works, which leads to different semantical embeddings. Instead of the semantics based on preference models as employed by Hansson and Åqvist, a neighborhood semantics is employed by Carmo and Jones. Based on the embeddings provided in [9] and here, further empirical studies are planned in order to compare these rival formalisations in particular regarding their practical reasoning performance and regarding their suitability to address relevant challenges (such as CTD reasoning) in practical applications.

Deep semantical embeddings of non-classical logics have been studied extensively in the related literature (examples are given in [17, 16]). The emphasis in these works typically is on interactive proofs of meta-logical properties. While meta-logical studies are also in reach for the methods presented here (see e.g. [8, 20]), the primary motivation, at least for this paper, is different. Our interest is in proof automation at object level, i.e. proof automation of Åqvist's system **E**. In other words, we are interested in practical normative reasoning applications of system **E** in which a high degree of automation (at object level) is required. Moreover, we are interested not only in the 'propositional' system **E**, but also in quantified extensions of it. For this, we plan to accordingly adapt the achievements of previous work (see e.g. [10, 4]). Making deep semantical embeddings scale for quantified non-classical logics, on the contrary, seems more challenging and less promising regarding proof automation.

The article is structured as follows: Sec. 2 describes system **E** and Sec. 3 introduces HOL. The semantical embedding of **E** in HOL is then devised and studied in Sec. 4. This section also shows the faithfulness (viz. soundness and completeness) of the embedding. Sec. 5 discusses the implementation in Isabelle/HOL [23]. Sec. 6 concludes the paper.

2 Dyadic Deontic Logic E

The language of **E** is obtained by adding the following operators to the syntax of propositional logic: $\Box$ (for necessity); $\Diamond$ (for possibility); and $\bigcirc(-/-)$ (for conditional obligation). $\bigcirc(\psi/\phi)$ is read "If ϕ, then ψ is obligatory". The set of well-formed formulas is defined in the straightforward way. Iteration of the modal and deontic operators is permitted, and so are "mixed" formulas, e.g., $\bigcirc(q/p)\wedge p$. We put $\top =_{df} \neg q \vee q$, for some propositional symbol q, and $\bot =_{df} \neg\top$.

A preference model is a structure $M = \langle S, \succeq, V\rangle$ where

- S is a non-empty set of items called possible worlds;

- $\succeq \subseteq S \times S$ (intuitively, $\succeq$ is a betterness or comparative goodness relation; "$s \succeq t$" can be read as "world s is at least as good as world t");
- V is a function assigning to each atomic sentence a set of worlds (i.e $V(q) \subseteq S$).

Given a preference model $M = \langle S, \succeq, V\rangle$ and a world $s \in W$, we define the satisfaction relation $M, s \vDash \varphi$ (read as "world s satisfies φ in model M") by induction on the structure of φ as described below. Standard deontic logic(SDL) is based on two class of states: good/bad (or green/red). In preference models we allow gradations between good and bad states. The closer a state is to ideality, the better. Intuitively, the evaluation rule for the dyadic obligation operator puts $\bigcirc(\psi/\phi)$ true whenever all the best ϕ-worlds are ψ-worlds. We define $V^M(\varphi) = \{s \in S \mid M, s \models \varphi\}$ and $\mathrm{opt}_{\succeq}(V(\varphi)) = \{s \in V(\varphi) \mid \forall t(t \vDash \varphi \rightarrow s \succeq t)\}$. Whenever the model M is obvious from context, we write $V(\varphi)$ instead of $V^M(\varphi)$.

$M, s \models p$ if and only if $s \in V(p)$
$M, s \models \neg\varphi$ if and only if $M, s \not\models \varphi$ (that is, not $M, s \models \varphi$)
$M, s \models \varphi \vee \psi$ if and only if $M, s \models \varphi$ or $M, s \models \psi$
$M, s \models \Box\varphi$ if and only if $V(\varphi) = S$
$M, s \models \bigcirc(\psi/\varphi)$ if and only if $\mathrm{opt}_{\succeq}(V(\varphi)) \subseteq V(\psi)$

As usual, a formula φ is valid in a preference model $M = \langle S, \succeq, V\rangle$ (notation: $M \models \varphi$) if and only if, for all worlds $s \in S$, $M, s \models \varphi$. A formula φ is valid (notation: $\models^{\mathbf{E}} \varphi$) if and only if it is valid in every preference model. The notions of semantic consequence and satisfiability in a model are defined as usual.

System **E** is defined by the following axioms and rules:

All truth functional tautologies	(PL)
S5-schemata for $\Box$ and $\Diamond$	(S5)
$\bigcirc(\psi_1 \rightarrow \psi_2/\varphi) \rightarrow (\bigcirc(\psi_1/\varphi) \rightarrow \bigcirc(\psi_2/\varphi))$	(COK)
$\bigcirc(\psi/\varphi) \rightarrow \Box\bigcirc(\psi/\varphi)$	(Abs)
$\Box\psi \rightarrow \bigcirc(\varphi/\psi)$	(Nec)
$\Box(\varphi_1 \leftrightarrow \varphi_2) \rightarrow (\bigcirc(\psi/\varphi_1) \leftrightarrow \bigcirc(\psi/\varphi_2))$	(Ext)
$\bigcirc(\varphi/\varphi)$	(Id)
$\bigcirc(\psi/\varphi_1 \wedge \varphi_2) \rightarrow \bigcirc(\varphi_2 \rightarrow \psi/\varphi_1)$	(Sh)
If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ then $\vdash \psi$	(MP)
If $\vdash \varphi$ then $\vdash \Box\varphi$	(N)

The notions of theoremhood, deducibility and consistency are defined as usual.

Theorem 1. *System* **E** *is (strongly) sound and complete with respect to:*

- *the class of all preference models*
- *the class of preference models in which $\succeq$ is reflexive*

– *the class of preference models in which* $\succeq$ *is total (for all* $s, t \in S$, $s \succeq t$ *or* $t \succeq s$*)*

Proof. See Parent [24].

3 Classical Higher-order Logic

In this section we introduce classical higher-order logic (HOL). The presentation, which has partly been adapted from [5], is rather detailed in order to keep the article sufficiently self-contained.

3.1 Syntax of HOL

To define the syntax of HOL, we first introduce the set T of *simple types*. We assume that T is freely generated from a set of *basic types* $BT \supseteq \{o, i\}$ using the function type constructor $\to$. Type o denotes the (bivalent) set of Booleans, and i a non-empty set of individuals.

For the definition of HOL, we start out with a family of denumerable sets of typed constant symbols $(C_\alpha)_{\alpha \in T}$, called the HOL *signature*, and a family of denumerable sets of typed variable symbols $(V_\alpha)_{\alpha \in T}$.[3] We employ Church-style typing, where each term t_α explicitly encodes its type information in subscript α.

The *language of HOL* is given as the smallest set of terms obeying the following conditions.

– Every typed constant symbol $c_\alpha \in C_\alpha$ is a HOL term of type α.
– Every typed variable symbol $X_\alpha \in V_\alpha$ is a HOL term of type α.
– If $s_{\alpha \to \beta}$ and t_α are HOL terms of types $\alpha \to \beta$ and α, respectively, then $(s_{\alpha \to \beta}\, t_\alpha)_\beta$, called *application*, is an HOL term of type β.
– If $X_\alpha \in V_\alpha$ is a typed variable symbol and s_β is an HOL term of type β, then $(\lambda X_\alpha s_\beta)_{\alpha \to \beta}$, called *abstraction*, is an HOL term of type $\alpha \to \beta$.

The above definition encompasses the simply typed λ-calculus. In order to extend this base framework into logic HOL we simply ensure that the signature $(C_\alpha)_{\alpha \in T}$ provides a sufficient selection of primitive logical connectives. Without loss of generality, we here assume the following *primitive logical connectives* to be part of the signature: $\neg_{o \to o} \in C_{o \to o}$, $\vee_{o \to o \to o} \in C_{o \to o \to o}$, $\Pi_{(\alpha \to o) \to o} \in C_{(\alpha \to o) \to o}$ and $=_{\alpha \to \alpha \to o} \in C_{\alpha \to \alpha \to o}$, abbreviated as $=^\alpha$. The symbols $\Pi_{(\alpha \to o) \to o}$ and $=_{\alpha \to \alpha \to o}$ are generally assumed for each type $\alpha \in T$. The denotation of the primitive logical connectives is fixed below according to their intended meaning. *Binder notation* $\forall X_\alpha\, s_o$ is used as an abbreviation for $(\Pi_{(\alpha \to o) \to o}(\lambda X_\alpha s_o))$. Universal quantification in HOL is thus modeled with the help of the logical constants $\Pi_{(\alpha \to o) \to o}$ to be used in combination with lambda-abstraction. That is, the only binding mechanism provided in HOL is lambda-abstraction.

[3] For example in Sec. 4 we will assume constant symbol r, with type $i \to i \to o$ as part of the signature.

HOL is a logic of terms in the sense that the *formulas of HOL* are given as the terms of type o. In addition to the primitive logical connectives selected above, we could assume *choice operators* $\epsilon_{(\alpha\to o)\to\alpha} \in C_{(\alpha\to o)\to\alpha}$ (for each type α) in the signature. We are not pursuing this here.

Type information as well as brackets may be omitted if obvious from the context, and we may also use infix notation to improve readability. For example, we may write $(s \vee t)$ instead of $((\vee_{o\to o\to o} s_o)t_o)$.

From the selected set of primitive connectives, other logical connectives can be introduced as abbreviations.[4] For example, we may define $s \wedge t := \neg(\neg s \vee \neg t)$, $s \to t := \neg s \vee t$, $s \longleftrightarrow t := (s \to t) \wedge (t \to s)$, $\top := (\lambda X_i X) = (\lambda X_i X)$, $\bot := \neg\top$ and $\exists X_\alpha s := \neg\forall X_\alpha \neg s$.

The notions of *free variables*, *α-conversion*, *$\beta\eta$-equality* (denoted as $=_{\beta\eta}$) and *substitution* of a term s_α for a variable X_α in a term t_β (denoted as $[s/X]t$) are defined as usual.

3.2 Semantics of HOL

The semantics of HOL is well understood and thoroughly documented. The introduction provided next focuses on the aspects as needed for this article. For more details we refer to the previously mentioned literature [7].

The semantics of choice for the remainder is Henkin semantics, i.e., we work with Henkin's general models [22]. Henkin models (and standard models) are introduced next. We start out with introducing frame structures.

A *frame* D is a collection $\{D_\alpha\}_{\alpha\in \mathrm{T}}$ of nonempty sets D_α, such that $D_o = \{T, F\}$ (for truth and falsehood). The $D_{\alpha\to\beta}$ are collections of functions mapping D_α into D_β.

A *model* for HOL is a tuple $M = \langle D, I\rangle$, where D is a frame, and I is a family of typed interpretation functions mapping constant symbols $p_\alpha \in C_\alpha$ to appropriate elements of D_α, called the *denotation of* p_α. The logical connectives $\neg$, $\vee$, Π and $=$ are always given their expected, standard denotations:[5]

- $I(\neg_{o\to o}) = not \in D_{o\to o}$ such that $not(T) = F$ and $not(F) = T$.

[4] As demonstrated by Andrews [2], we could in fact start out with only primitive equality in the signature (for all types α) and introduce all other logical connectives as abbreviations based on it. Alternatively, we could remove primitive equality from the above signature, since equality can be defined in HOL from these other logical connectives by exploiting Leibniz' principle, expressing that two objects are equal if they share the same properties. *Leibniz equality* $\doteq^\alpha$ at type α is thus defined as $s_\alpha \doteq^\alpha t_\alpha := \forall P_{\alpha\to o}(Ps \longleftrightarrow Pt)$. The motivation for the redundant signature as selected here is to stay close to the the choices taken in implemented theorem provers such as LEO-II and Leo-III and also to theory paper [7], which is recommended for further details.

[5] Since $=_{\alpha\to\alpha\to o}$ (for all types α) is in the signature, it is ensured that the domains $D_{\alpha\to\alpha\to o}$ contain the respective identity relations. This addresses an issue discovered by Andrews [1]: if such identity relations did not exist in the $D_{\alpha\to\alpha\to o}$, then Leibniz equality in Henkin semantics might not denote as intended.

- $I(\vee_{o\to o\to o}) = or \in D_{o\to o\to o}$ such that $or(a,b) = T$ iff $(a = T$ or $b = T)$.
- $I(=_{\alpha\to\alpha\to o}) = id \in D_{\alpha\to\alpha\to o}$ such that for all $a, b \in D_\alpha$, $id(a,b) = T$ iff a is identical to b.
- $I(\Pi_{(\alpha\to o)\to o}) = all \in D_{(\alpha\to o)\to o}$ such that for all $s \in D_{\alpha\to o}$, $all(s) = T$ iff $s(a) = T$ for all $a \in D_\alpha$; i.e., s is the set of all objects of type α.

Variable assignments are a technical aid for the subsequent definition of an interpretation function $\|.\|^{M,g}$ for HOL terms. This interpretation function is parametric over a model M and a variable assignment g.

A *variable assignment* g maps variables X_α to elements in D_α. $g[d/W]$ denotes the assignment that is identical to g, except for variable W, which is now mapped to d.

The *denotation* $\|s_\alpha\|^{M,g}$ of an HOL term s_α on a model $M = \langle D, I\rangle$ under assignment g is an element $d \in D_\alpha$ defined in the following way:

$$\begin{aligned}
\|p_\alpha\|^{M,g} &= I(p_\alpha)\\
\|X_\alpha\|^{M,g} &= g(X_\alpha)\\
\|(s_{\alpha\to\beta}\, t_\alpha)_\beta\|^{M,g} &= \|s_{\alpha\to\beta}\|^{M,g}(\|t_\alpha\|^{M,g})\\
\|(\lambda X_\alpha s_\beta)_{\alpha\to\beta}\|^{M,g} &= \text{the function } f \text{ from } D_\alpha \text{ to } D_\beta \text{ such that}\\
&\qquad f(d) = \|s_\beta\|^{M,g[d/X_\alpha]} \text{ for all } d \in D_\alpha
\end{aligned}$$

A model $M = \langle D, I\rangle$ is called a *standard model* if and only if for all $\alpha, \beta \in T$ we have $D_{\alpha\to\beta} = \{f \mid f : D_\alpha \longrightarrow D_\beta\}$. In a *Henkin model (general model)* function spaces are not necessarily full. Instead it is only required that for all $\alpha, \beta \in T$, $D_{\alpha\to\beta} \subseteq \{f \mid f : D_\alpha \longrightarrow D_\beta\}$. However, it is required that the valuation function $\|\cdot\|^{M,g}$ from above is total, so that every term denotes. Note that this requirement, which is called *Denotatpflicht*, ensures that the function domains $D_{\alpha\to\beta}$ never become too sparse, that is, the denotations of the lambda-abstractions as devised above are always contained in them.

Corollary 1. *For any Henkin model $M = \langle D, I\rangle$ and variable assignment g:*

1. $\|(\neg_{o\to o}\, s_o)_o\|^{M,g} = T$ *iff* $\|s_o\|^{M,g} = F$.
2. $\|((\vee_{o\to o\to o}\, s_o)\, t_o)_o\|^{M,g} = T$ *iff* $\|s_o\|^{M,g} = T$ *or* $\|t_o\|^{M,g} = T$.
3. $\|((\wedge_{o\to o\to o}\, s_o)\, t_o)_o\|^{M,g} = T$ *iff* $\|s_o\|^{M,g} = T$ *and* $\|t_o\|^{M,g} = T$.
4. $\|((\to_{o\to o\to o}\, s_o)\, t_o)_o\|^{M,g} = T$ *iff* *(if* $\|s_o\|^{M,g} = T$ *then* $\|t_o\|^{M,g} = T$*).*
5. $\|((\longleftrightarrow_{o\to o\to o}\, s_o)\, t_o)_o\|^{M,g} = T$ *iff* *(*$\|s_o\|^{M,g} = T$ *iff* $\|t_o\|^{M,g} = T$*).*
6. $\|\top\|^{M,g} = T$.
7. $\|\bot\|^{M,g} = F$.
8. $\|(\forall X_\alpha s_o)_o\|^{M,g} = T$ *iff* *for all* $d \in D_\alpha$ *we have* $\|s_o\|^{M,g[d/X_\alpha]} = T$.
9. $\|(\exists X_\alpha s_o)_o\|^{M,g} = T$ *iff* *there exists* $d \in D_\alpha$ *such that* $\|s_o\|^{M,g[d/X_\alpha]} = T$.

Proof. We leave the proof as an exercise to the reader.

An HOL formula s_o is *true* in a Henkin model M under assignment g if and only if $\|s_o\|^{M,g} = T$; this is also expressed by writing that $M, g \models^{\text{HOL}} s_o$. An HOL formula s_o is called *valid* in M, which is expressed by writing that $M \models^{\text{HOL}} s_o$, if and only if $M, g \models^{\text{HOL}} s_o$ for all assignments g. Moreover, a formula s_o is called *valid*, expressed by writing that $\models^{\text{HOL}} s_o$, if and only if s_o is valid in all Henkin models M.

4 Embedding E into HOL

This section, the core contribution of this article, presents a shallow semantical embedding of system **E** in HOL and proves its soundness and completeness. In contrast to a deep logical embedding, in which the syntactical structure and the semantics of logic L would be formalized in full detail (using e.g. structural induction and recursion), only the core differences in the semantics of both system **E** and meta-logic HOL are explicitly encoded here. In a certain sense we show, that system **E** can in fact in be identified and handled as a natural fragment of HOL.

4.1 Semantical Embedding

The formulas of **E** are identified in our semantical embedding with certain HOL terms (predicates) of type $i \to o$. They can be applied to terms of type i, which are assumed to denote possible worlds. That is, the HOL type i is now identified with a (non-empty) set of worlds. Type $i \to o$ is abbreviated as τ in the remainder. The HOL signature is assumed to contain the constant symbol $r_{i\to\tau}$. Moreover, for each propositional symbol p^j of **E**, the HOL signature must contain the corresponding constant symbol p^j_τ. Without loss of generality, we assume that besides those symbols and the primitive logical connectives of HOL, no other constant symbols are given in the signature of HOL.

The mapping $\lfloor\cdot\rfloor$ translates a formula φ of **E** into a formula $\lfloor\varphi\rfloor$ of HOL of type τ. The mapping is defined recursively:

$$\begin{array}{ll}
\lfloor p^j \rfloor & = p^j_\tau \\
\lfloor \neg s \rfloor & = \neg_\tau \lfloor s \rfloor \\
\lfloor s \vee t \rfloor & = \vee_{\tau\to\tau\to\tau} \lfloor s \rfloor \lfloor t \rfloor \\
\lfloor \Box s \rfloor & = \Box_{\tau\to\tau} \lfloor s \rfloor \\
\lfloor \bigcirc(t/s) \rfloor & = \bigcirc_{\tau\to\tau\to\tau} \lfloor s \rfloor \lfloor t \rfloor
\end{array}$$

$\neg_\tau$, $\vee_{\tau\to\tau\to\tau}$, $\Box_{\tau\to\tau}$, $\bigcirc_{\tau\to\tau\to\tau}$ abbreviate the following formulas of HOL:

$$\begin{array}{ll}
\neg_{\tau\to\tau} & = \lambda A_\tau \lambda X_i \neg (A\, X) \\
\vee_{\tau\to\tau\to\tau} & = \lambda A_\tau \lambda B_\tau \lambda X_i (A\, X \vee B\, X) \\
\Box_{\tau\to\tau} & = \lambda A_\tau \lambda X_i \forall Y_i (A\, Y) \\
\bigcirc_{\tau\to\tau\to\tau} & = \lambda A_\tau \lambda B_\tau \lambda X_i \forall W_i (\,(\lambda V_i (A\, V \wedge (\forall Y_i (A\, Y \to r_{i\to\tau} V\, Y))))\, W \to B\, W)^6
\end{array}$$

Analyzing the truth of a translated formula $\lfloor s \rfloor$ in a world represented by term w_i corresponds to evaluating the application $(\lfloor s \rfloor\, w_i)$. In line with previous work [10], we define $\mathrm{vld}_{\tau\to o} = \lambda A_\tau \forall S_i (A\, S)$. With this definition, validity of a formula s in **E** corresponds to the validity of the formula $(\mathrm{vld}\, \lfloor s \rfloor)$ in HOL, and vice versa.

[6] If $\mathrm{opt}_{\succeq}(A)$ is taken as a abbreviation for $\lambda V_i (AV \wedge (\forall Y_i (AY \to r_{i\to\tau} V\, Y)))$, then this can be simplified to $\bigcirc_{\tau\to\tau\to\tau} = \lambda A_\tau \lambda B_\tau \lambda X_i (\mathrm{opt}_{\succeq}(A) \subseteq B)$.

4.2 Soundness and Completeness

To prove the soundness and completeness, that is, faithfulness, of the above embedding, a mapping from preference models into Henkin models is employed.

Definition 1 (Preference model ⇒ Henkin model). *Let $M = \langle S, \succeq, V\rangle$ be a preference model. Let $p^1, ..., p^m \in PV$, for $m \geq 1$ be propositional symbols and $\lfloor p^j \rfloor = p^j_\tau$ for $j = 1, ..., m$. A Henkin model $H^M = \langle \{D_\alpha\}_{\alpha\in T}, I\rangle$ for M is defined as follows: D_i is chosen as the set of possible worlds S and all other sets $D_{\alpha\to\beta}$ are chosen as (not necessarily full) sets of functions from D_α to D_β. For all $D_{\alpha\to\beta}$ the rule that every term $t_{\alpha\to\beta}$ must have a denotation in $D_{\alpha\to\beta}$ must be obeyed, in particular, it is required that D_τ and $D_{i\to\tau}$ contain the elements Ip^j_τ and $Ir_{i\to\tau}$. Interpretation I is constructed as follows:*

1. *For $1 \leq i \leq m$, $Ip^j_\tau \in D_\tau$ is chosen such that $Ip^j_\tau(s) = T$ if $s \in V(p^j)$ in M and $Ip^j_\tau(s) = F$ otherwise.*
2. *$Ir_{i\to\tau} \in D_{i\to\tau}$ is chosen such that $Ir_{i\to\tau}(s, u) = T$ if $s \succeq u$ in M and $Ir_{i\to\tau}(s, u) = F$ otherwise.*

Since we assume that there are no other symbols (besides the r, the p^j and the primitive logical connectives) in the signature of HOL, I is a total function. Moreover, the above construction guarantees that H^M is a Henkin model: $\langle D, I\rangle$ is a frame, and the choice of I in combination with the Denotatpflicht ensures that for arbitrary assignments g, $\|.\|^{H^M,g}$ is a total evaluation function.

Lemma 1. *Let H^M be a Henkin model for a preference model M. For all formula δ of $\mathbf{E}$, all assignment g and world s it holds:*

$$M, s \models \delta \text{ if and only if } \|\lfloor\delta\rfloor\, S_i\|^{H^M, g[s/S_i]} = T$$

Proof. See appendix.

Lemma 2 (Henkin model ⇒ Preference model). *For every Henkin model $H = \langle \{D_\alpha\}_{\alpha\in T}, I\rangle$ there exists a corresponding preference model M. Corresponding here means that for all formula δ of $\mathbf{E}$ and for all assignment g and world s, $\|\lfloor\delta\rfloor S_i\|^{H,g[s/S_i]} = T$ if and only if $M, s \models \delta$.*

Proof. Suppose that $H = \langle \{D_\alpha\}_{\alpha\in T}, I\rangle$ is a Henkin model. Without loss of generality, we can assume that the domains of H are denumerable [22]. We construct the corresponding preference model M as follows:

- $S = D_i$.
- $s \succeq u$ for $s, u \in S$ iff $Ir_{i\to\tau}(s, u) = T$.
- $s \in V(p^j_\tau)$ iff $Ip^j_\tau(s) = T$ for all p^j.

Moreover, the above construction ensures that H is a Henkin model for M. Hence, Lemma 1 applies. This ensures that for all formulas δ of $\mathbf{E}$, for all assignment g and all world s we have $\|\lfloor\delta\rfloor S_i\|^{H,g[s/S_i]} = T$ if and only if $M, s \models \delta$.

Theorem 2 (Soundness and Completeness of the Embedding).

$$\models^{\mathbf{E}} \varphi \textit{ if and only if } \models^{HOL} \textit{vld}\lfloor\varphi\rfloor$$

Proof. (Soundness, $\leftarrow$) The proof is by contraposition. Assume $\not\models^{\mathbf{E}} \varphi$, i.e, there is a preference model $M = \langle S, \succeq, V\rangle$, and a world $s \in S$, such that $M, s \not\models \varphi$. By Lemma 1 for an arbitrary assignment g it holds that $\|\lfloor\varphi\rfloor\, S_i\|^{H^M, g[s/S_i]} = F$ in Henkin model $H^M = \langle\{D_\alpha\}_{\alpha\in T}, I\rangle$. Thus, by definition of $\|.\|$, it holds that $\|\forall S_i(\lfloor\varphi\rfloor\, S_i)\|^{H^M,g} = \|\text{vld}\,\lfloor\varphi\rfloor\|^{H^M,g} = F$. Hence, $H^M \not\models^{\text{HOL}} \text{vld}\,\lfloor\varphi\rfloor$. By definition $\not\models^{\text{HOL}} \text{vld}\,\lfloor\varphi\rfloor$.

(Completeness, $\rightarrow$) The proof is again by contraposition. Assume $\not\models^{\text{HOL}} \text{vld}\,\lfloor\varphi\rfloor$, i.e., there is a Henkin model $H = \langle\{D_\alpha\}_{\alpha\in T}, I\rangle$ and an assignment g such that $\|\text{vld}\,\lfloor\varphi\rfloor\|^{H,g} = F$. By Lemma 2, there is a preference model M such that $M \not\models \varphi$. Hence, $\not\models^{\mathbf{E}} \varphi$.

5 Implementation in Isabelle/HOL

The semantical embedding as devised in Sec. 4 has been implemented in the higher-order proof assistant Isabelle/HOL [23]. Figure 1 displays the respective encoding. Figure 2 applies this encoding to Chisholm's paradox (cf. [14]), which involves the following four statements:

1. It ought to be that a certain man go to help his neighbors;
2. It ought to be that if he goes he tells them he is coming;
3. If he does not go, he ought not to tell them he is coming;
4. He does not go.

These statements can be given a consistent formalisation in DDL see Fig. 2. This is confirmed by the model finder Nitpick [12] integrated with Isabelle/HOL. Nitpick computes an intuitive, small model for the scenario consisting of one possible world i_1. The actual world is i_1 also. Preference relation r is interpreted in this model as $r = \emptyset$.

In the actual world the man doesn't go to help his neighbors and doesn't tell them that he is coming. That is, $V(\neg go) = V(\neg tell) = \{i_1\}$. Also, we have $op(V(\top)) = \emptyset$. So, $i_1 \models \bigcirc(go/\top)$ by the evaluation rule for $\bigcirc$. Similarly, $op(V(go)) = op(V(\neg go)) = \emptyset$ implies $i_1 \models \bigcirc(tell/go)$ and $i_1 \models \bigcirc(\neg tell/\neg go)$.

6 Conclusion

A shallow semantical embedding of Åqvist's dyadic deontic logic **E** in classical higher-order logic has been presented, and shown to be faithful (sound an complete). The works presented here and in [9] provide the theoretical foundation for the implementation and automation of dyadic deontic logic within existing theorem provers and proof assistants for HOL. We do not provide new logics. Instead, we provide an empirical infrastructure for assessing practical aspects of an ambitious, state-of-the-art deontic logics; this has not been done

before. An interesting and relevant aspect of the approach is that (based on the ideas of previous work [10, 4]) quantified extensions of system **E** and [9] can easily be implemented and studied in the framework and experimentally assessed. There is much room for future work. For example, experiments could investigate whether the provided implementation already supports non-trivial applications in practical normative reasoning, or whether further improvements are required. Moreover, we could employ our implementation to systematically study some meta-logical properties of dyadic deontic logic system **E** within Isabelle/HOL.

References

1. Andrews, P.B.: General models and extensionality. Journal of Symbolic Logic **37**(2), 395–397 (1972)
2. Andrews, P.B.: Church's type theory. In: Zalta, E.N. editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2014 edition, (2014)
3. Åqvist, L.: Deontic logic. In: Handbook of philosophical logic, pp. 147–264. Springer, Dordrecht (2002)
4. Benzmüller, C.: Automating quantified conditional logics in HOL. In: Rossi, F. (eds.) 23rd International Joint Conference on Artificial Intelligence, IJCAI-13, Beijing, China, AAAI Press, pp. 746–753, (2013)
5. Benzmüller, C.: Cut-elimination for quantified conditional logic. Journal of Philosophical Logic, **46**(3), 333–353, (2017)
6. Benzmüller, C.: Recent successes with a meta-logical approach to universal logical reasoning (extended abstract). In: da Costa Cavalheiro, S.A., Fiadeiro, J.L. (eds.) Formal Methods: Foundations and Applications, volume 10623 of Lecture Notes in Computer Science, pp. 7–11. Springer, (2017)
7. Benzmüller, C., Brown, C., Kohlhase, M.: Higher-order semantics and extensionality. Journal of Symbolic Logic, **69**(4), 1027–1088, (2004)
8. Benzmüller, C., Claus, M., Sultana, N.: Systematic verification of the modal logic cube in Isabelle/HOL. In: Kaliszyk, C., Paskevich, A. (eds.) PxTP 2015, Berlin, Germany, EPTCS, vol. 186, pp. 24–41 (2015). https://doi.org/10.4204/EPTCS.186.5
9. Benzmüller, C., Farjami, A., Parent., X.: A dyadic deontic logic in HOL. In: Broersen, J., Condoravdi, C., Nair, S., Pigozzi, G. (eds.) Deontic Logic and Normative Systems — 14th International Conference, DEON 2018, Utrecht, The Netherlands, 3-6 July, 2018, pp. 33–50, College Publications, UK, (2018)
10. Benzmüller, C., Paulson, L.C.: Quantified multimodal logics in simple type theory. Logica Universalis (Special Issue on Multimodal Logics), **7**(1), 7–20, (2013)
11. Benzmüller, C., Sultana, N., Paulson, L. C., Theiß, F.: The higher-order prover LEO-II. Journal of Automated Reasoning, **55**(4), 389–404, (2015)
12. Blanchette, J.C., Nipkow, T.: Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In: ITP 2010, number 6172 in Lecture Notes in Computer Science, pp. 131–146. Springer, (2010)
13. Carmo, J. M. C. L. M., Jones, A. J. I.: Completeness and decidability results for a logic of contrary-to-duty conditionals. Journal of Logic and Computation **23**(3), 585–626 (2013)
14. Chisholm, R.M.: Contrary-to-duty imperatives and deontic logic. Analysis, **24**(2), 33–36 (1963)

15. Church, A.: A formulation of the simple theory of types. Journal of Symbolic Logic, **5**(2), 56–68, (1940)
16. Doczkal, C., Bard, J.: Completeness and decidability of converse PDL in the constructive type theory of Coq. In Proceedings of the 7th ACM SIGPLAN International Conference on Certified Programs and Proofs, pp. 42–52, ACM, New York, USA (2018)
17. Doczkal, C., Smolka., G.: Completeness and decidability results for CTL in constructive type theory. Journal of Automated Reasoning **56**(32), 343–365 (2016)
18. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L.: Handbook of deontic logic and normative systems. Volume 1. College Publications, UK, (2013)
19. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L.: Handbook of deontic logic and normative systems. Volume 2. College Publications, UK, (2018)
20. Kirchner, D., Benzmüller, C., Zalta, E.: Mechanizing principia logico-metaphysica in functional type theory. CoRR `https://arxiv.org/abs/1711.06542`, (2017)
21. Hansson, B.: An analysis of some deontic logics. Nous, 373–398 (1969)
22. Henkin, L.: Completeness in the theory of types. Journal of Symbolic Logic, **5**(2), 81–91, (1950)
23. Nipkow, T., Paulson, L.C., Wenzel., M.: Isabelle/HOL — A proof assistant for higher-Order logic, volume 2283 of Lecture Notes in Computer Science. Springer, (2002)
24. Parent, X.: Completeness of Åqvists systems E and F. The Review of Symbolic Logic, **8**(1), 164–177 (2015)

Appendix (Proof for Lemma 1)

In the proof we implicitly employ curring and uncurring, and we associate sets with their characteristic functions. Throughout the proof whenever possible we omit types in order to avoid making the notation too cumbersome. The proof of lemma 1 is by induction on the structure of δ. We start with the case where δ is p^j. We have $\|\lfloor p^j \rfloor S\|^{H^M, g[s/S_i]} = T$

$\Leftrightarrow \quad \|p^j_\tau S\|^{H^M, g[s/S_i]} = T$
$\Leftrightarrow \quad Ip^j_\tau(s) = T$
$\Leftrightarrow \quad s \in V(p^j) \quad$ (by definition of H^M)
$\Leftrightarrow \quad M, s \vDash p^j$

For the inductive cases we make the hypothesis that the claim holds for sentences δ' shorther than δ:

> **Inductive hypothesis:** For all assignment g and state s, $\|\lfloor \delta' \rfloor S\|^{H^M, g[s/S_i]} = T$ if and only if $M, s \vDash \delta'$

We consider each inductive case in turn:
(a) $\delta = \varphi \vee \psi$. In this case:

$\|\lfloor \varphi \vee \psi \rfloor S\|^{H^M, g[s/S_i]} = T$
$\Leftrightarrow \quad \|(\lfloor \varphi \rfloor \vee_{\tau\to\tau\to\tau} \lfloor \psi \rfloor) S\|^{H^M, g[s/S_i]} = T$
$\Leftrightarrow \quad \|(\lfloor \varphi \rfloor S) \vee (\lfloor \psi \rfloor S)\|^{H^M, g[s/S_i]} = T \quad (((\lfloor \varphi \rfloor \vee_{\tau\to\tau\to\tau} \lfloor \psi \rfloor) S) =_{\beta\eta} ((\lfloor \varphi \rfloor S) \vee (\lfloor \psi \rfloor S)))$

$\Leftrightarrow \quad \|\lfloor\varphi\rfloor S\|^{H^M,g[s/S_i]} = T$ or $\|\lfloor\psi\rfloor S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad M, s \vDash \varphi$ or $M, s \vDash \psi$ (by induction hypothesis)
$\Leftrightarrow \quad M, s \vDash \varphi \vee \psi$

(b) $\delta = \neg\varphi$. In this case:

$\|\lfloor\neg\varphi\rfloor S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|(\neg_{\tau\to\tau}\lfloor\varphi\rfloor)S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|\neg(\lfloor\varphi\rfloor)S)\|^{H^M,g[s/S_i]} = T \quad ((\neg_{\tau\to\tau}\lfloor\varphi\rfloor)S =_{\beta\eta} \neg(\lfloor\varphi\rfloor S))$
$\Leftrightarrow \quad \|\lfloor\varphi\rfloor S\|^{H^M,g[s/S_i]} = F$
$\Leftrightarrow \quad M, s \nvDash \varphi$ (by induction hypothesis)
$\Leftrightarrow \quad M, s \vDash \neg\varphi$

(c) $\delta = \Box\varphi$. We have the following chain of equivalences:

$\|\lfloor\Box\varphi\rfloor S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|(\lambda X\forall Y(\lfloor\varphi\rfloor Y))S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|\forall Y(\lfloor\varphi\rfloor Y)\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad$ For all $a \in D_i$ we have $\|\lfloor\varphi\rfloor Y\|^{H^M,g[s/S_i][a/Y_i]} = T$
$\Leftrightarrow \quad$ For all $a \in D_i$ we have $\|\lfloor\varphi\rfloor Y\|^{H^M,g[a/Y_i]} = T \quad (S \notin free(\lfloor\varphi\rfloor))$
$\Leftrightarrow \quad$ For all $a \in S$ we have $M, a \models \varphi$ (by induction hypothesis)
$\Leftrightarrow \quad M, s \models \Box\varphi$

(d) $\delta = \bigcirc(\psi/\varphi)$. We have the following chain of equivalences:

$\|\lfloor\bigcirc(\psi/\varphi)\rfloor S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|(\lambda X\forall W(\,(\lambda V(\lfloor\varphi\rfloor\, V \wedge (\forall Y(\lfloor\varphi\rfloor\, Y \to r\, V\, Y))))\, W \to \lfloor\psi\rfloor\, W))S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad \|\forall W(\,(\lambda V(\lfloor\varphi\rfloor\, V \wedge (\forall Y(\lfloor\varphi\rfloor\, Y \to r\, V\, Y))))\, W \to \lfloor\psi\rfloor\, W)\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \quad$ For all $u \in D_i$ we have:
$\|(\lambda V(\lfloor\varphi\rfloor\, V \wedge (\forall Y(\lfloor\varphi\rfloor\, Y \to r\, V\, Y))))\, W \to \lfloor\psi\rfloor\, W\|^{H^M,g[s/S_i][u/W_i]} = T$
$\Leftrightarrow \quad$ For all $u \in D_i$ we have:
If $\|(\lambda V(\lfloor\varphi\rfloor\, V \wedge (\forall Y(\lfloor\varphi\rfloor\, Y \to r\, V\, Y))))\, W\|^{H^M,g[s/S_i][u/W_i]} = T$,
then $\|\lfloor\psi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$
$\Leftrightarrow \quad$ For all $u \in D_i$ we have:
If $\|\lfloor\varphi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$ and
$\|\forall Y(\lfloor\varphi\rfloor\, Y \to r\, W\, Y)\|^{H^M,g[s/S_i][u/W_i]} = T$,
then $\|\lfloor\psi\rfloor V\|^{H^M,g[s/S_i][u/W_i]} = T$
$\Leftrightarrow \quad$ For all $u \in D_i$ we have:
If $\|\lfloor\varphi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$ and
for all $t \in D_i$ we have $\|\forall Y(\lfloor\varphi\rfloor\, Y \to r\, W\, Y)\|^{H^M,g[s/S_i][u/W_i][t/Y_i]} = T$,
then $\|\lfloor\psi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$
$\Leftrightarrow \quad$ For all $u \in D_i$ we have:
If $\|\lfloor\varphi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$ and
for all $t \in D_i$ we have $\|\lfloor\varphi\rfloor Y\|^{H^M,g[s/S_i][u/W_i][t/Y_i]} = T$ implies $Ir_{i\to\tau}(u,t) = T$,
then $\|\lfloor\psi\rfloor W\|^{H^M,g[s/S_i][u/W_i]} = T$

$\Leftrightarrow$ For all $u \in D_i$ we have:
If $u \in V(\varphi)$ and
for all $t \in D_i$ we have $t \in V(\varphi)$ implies $u \succeq t$,
then $u \in V(\psi)$ (**see the justification ***)
$\Leftrightarrow$ $\mathrm{opt}_{\succeq}(V(\varphi)) \subseteq V(\psi)$
$\Leftrightarrow$ $M, s \models \bigcirc(\psi/\varphi)$

Justification *: What we need to show is: $\|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}$ is identified with $V(\varphi)$ (analogously ψ). By induction hypothesis, for all assignment g and state s, we have $\|\lfloor\varphi\rfloor S\|^{H^M,g[s/S_i]} = T$ if and only if $M, s \vDash \varphi$. Expanding the details of this equivalence we have: for all assignment g and state s

$\Leftrightarrow s \in \|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}$ (functions to type o are associated with sets)
$\Leftrightarrow \|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}(s) = T$
$\Leftrightarrow \|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}\|S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow \|\lfloor\varphi\rfloor S\|^{H^M,g[s/S_i]} = T$
$\Leftrightarrow M, s \vDash \varphi$
$\Leftrightarrow s \in V(\varphi)$

Hence, $s \in \|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}$ if and only if $s \in V(\varphi)$.
By extensionality we thus know that $\|\lfloor\varphi\rfloor\|^{H^M,g[s/S_i]}$ is identified with $V(\varphi)$.
Moreover, since H^M obeys the Denotatpflicht we know that $V(\varphi) \in D_\tau$.

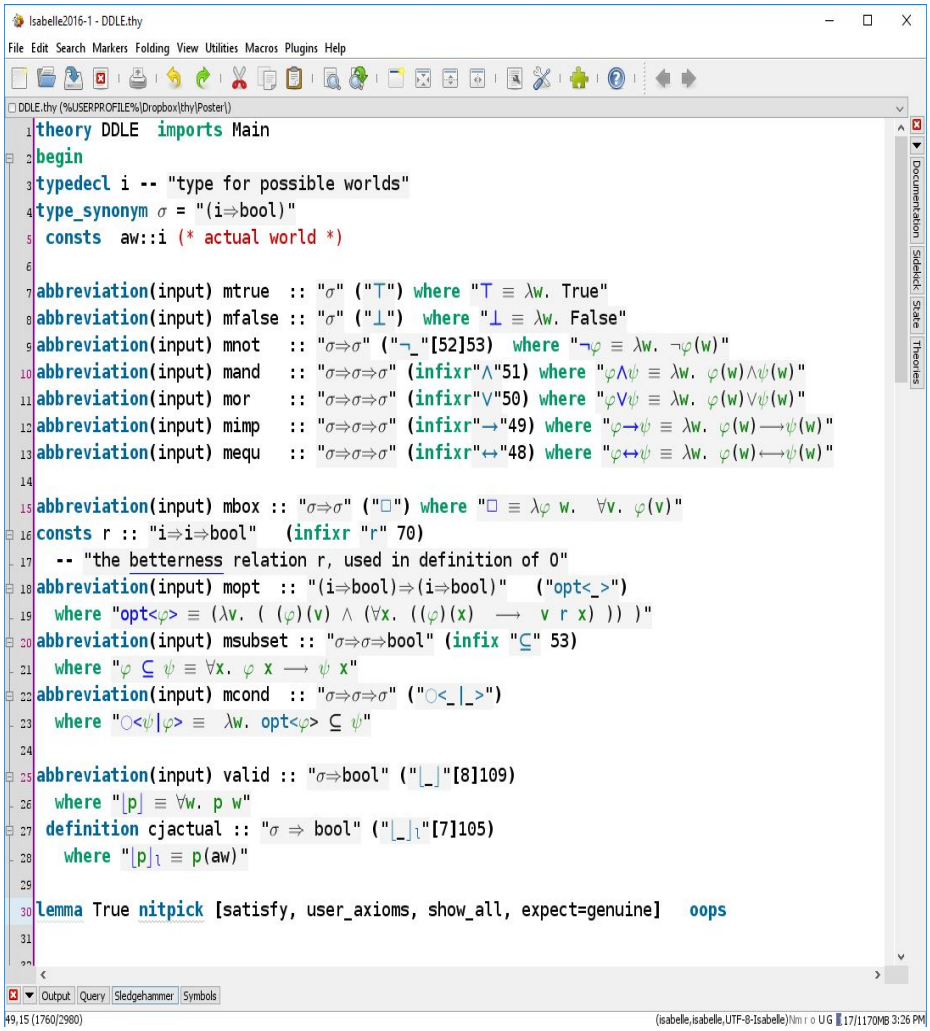

Fig. 1. Shallow semantical embedding of **E** in Isabelle/HOL.

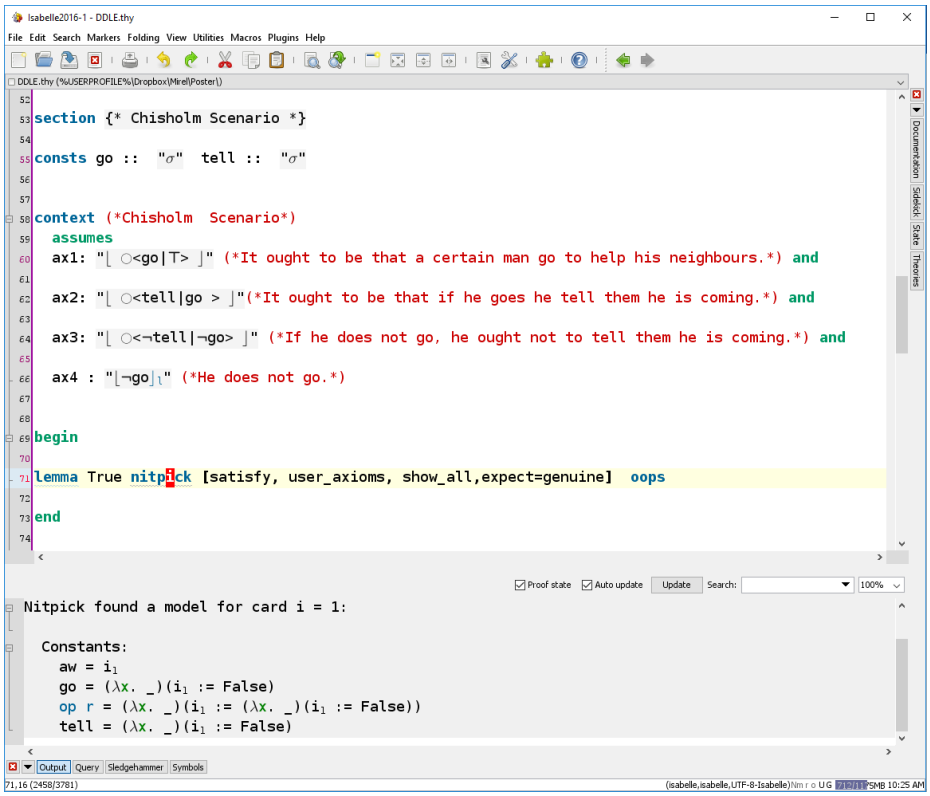

Fig. 2. The Chisholm paradox scenario encoded in **E** (the shallow semantical embedding of **E** in Isabelle/HOL as displayed in Fig. 1 is imported here). Nitpick confirms consistency the encoded statements.