Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories[*]

David Fuenmayor[1] and Christoph Benzmüller[1,2]

[1] Freie Universität Berlin, Germany
[2] University of Luxembourg, Luxembourg

Abstract. The computer-mechanization of an ambitious explicit ethical theory, Gewirth's Principle of Generic Consistency, is used to showcase an approach for representing and reasoning with ethical theories exhibiting complex logical features like alethic and deontic modalities, indexicals, higher-order quantification, among others. Harnessing the high expressive power of Church's type theory as a meta-logic to semantically embed a combination of quantified non-classical logics, our work pushes existing boundaries in knowledge representation and reasoning. We demonstrate that intuitive encodings of complex ethical theories and their automation on the computer are no longer antipodes.

1 Introduction

Hybrid architectures for ethical autonomous agents that integrate both bottom-up learning and top-down deliberation from upper principles are receiving increased attention; cf. (ed.); Dignum (2017); Scheutz (2017); Malle (2016); Dennis et al. (2016); Anderson and Anderson (2014); Wallach et al. (2008) and the references therein. Irrespective of the preferred direction, it is becoming increasingly evident that adequate explicit representations of ethical knowledge are beneficial, if not mandatory, to obtain satisfactory solutions. Bottom-up approaches may benefit from expressive languages to *explicitly* represent the learned ethical knowledge in an scrutable, communicable and transferable manner. Top-down approaches usually rely on expressive logic languages to enable an *intuitive and accurate representation and reasoning* with ethical theories. Unfortunately, however, very few approaches are currently available that enable adequate and realistic, explicit formal encodings of non-trivialized ethical theories, and that at the same time support intuitive interactive-automated reasoning with them.

In this paper *we demonstrate a methodology and implementation of such an ambitious ethical reasoning machinery*. Our approach is based on classical higher-order logic (HOL), aka Church's type theory (Benzmüller and Andrews, 2019), which we exploit as a meta-logic to encode combinations of non-classical logics for normative reasoning as suited for a given application context. The methodology and techniques we present, cf. also Benzmüller et al. (2019), can bring many benefits to the design of ethically-critical systems aiming at scrutability, verifiability, and the ability to provide justification for its decision-making. They are particularly relevant to the design of explicit ethical agents (Moor, 2009). In particular, this area faces tough philosophical

---

and practical challenges. No consensus is currently in sight, if possible at all, concerning the choice of upper moral values and principles that constitute a generally agreed normative ethics for intelligent autonomous agents. For example, utilitarianism and deontology have both been critically discussed in this context.

We exemplarily study another relevant and ambitious theory in normative ethics: Alan Gewirth's "Principle of Generic Consistency (PGC)" (Gewirth, 1981; Beyleveld, 1991), which has been proposed as an emendation of the *Golden Rule*. Our aim is not to defend or assess Gewirth's work in comparison to other approaches. We instead present a methodology and technique enabling the intuitive and accurate representation of ambitious ethical theories, and for this we take the PGC as a showcase and exemplarily assess its logical validity. Such an ambitious ethical theory has never before been assessed on the computer at such a level of detail (i.e. without trivializing it by abstraction).

Our method enables the reuse of modern interactive and automated higher-order theorem proving technology, and in this sense it establishes a *relevant bridge between different research communities*. On a practical level our work also addresses what we consider one of the biggest challenges in the area: to *represent complex ethical theories in both a machine and human interpretable manner and to carry out complex reasoning in real-time with incomplete and inconsistent information*. And finally, as a side-effect, we have *revealed and fixed some (minor) issues in Gewirth's PGC*.

Our choice of HOL at the meta-level is motivated by the goal of flexibly combining expressive non-classical logics as required for the formal encoding of complex ethical theories. Current theories in normative and machine ethics are, quite understandably, formulated predominantly in natural language. While this supports human deliberation and agreement about what kind of moral beings we want future intelligent agents to be, it also hampers their implementation in machines. Hence expressive formal languages are required, which enable flexible combinations of different types of non-classical logics. This is because ethical theories are usually challenged by complex linguistic expressions, including modalities (alethic, epistemic, temporal, etc.), counterfactual conditionals, generalized quantifiers, (un-)conditional obligations, among many others.

The meta-logical approach we exploit and demonstrate grounds on a technique known as *(shallow) semantical embedding*. The approach will be addressed in §2, where we present an extended embedding of a dyadic deontic logic (DDL) by Carmo and Jones (2002) in HOL and *combine, among others, conditional obligations with further modalities and quantifiers*. The combined logic is immune to known paradoxes in deontic logic, in particular, the so-called contrary-to-duty scenarios, in which a 'secondary' obligation must come into effect when a 'primary' obligation is violated (contradicted). Moreover, conditional (dyadic) obligations in DDL are of a defeasible and paraconsistent nature and thus lend themselves to normative reasoning with incomplete and inconsistent information. In §3 and §4 we will represent and formally assess Gewirth's PGC using this expressive logic combination. We also demonstrate how our technique has been utilized to reveal and fix some (minor) issues in Gewirth's work. Related work and short summary are presented in §5, and a formally-verified, unabridged version of our formal encoding of Gewirth's theory and argument is provided in Fuenmayor and Benzmüller (2018).

2   Combining Expressive Logics in HOL

We utilize the *shallow semantical embeddings* (SSE) approach to combining logics. SSE exploits HOL as a meta-logic in order to embed the syntax and semantics of some target logics, thereby turning theorem proving systems for HOL into universal reasoning engines (Benzmüller, 2019). Moreover, an approach drawing upon SSE has been proposed as the foundation for a flexible deontic logic reasoning infrastructure (Benzmüller et al., 2019). We thus assess, in some sense, the promises of this framework at hand of a non-trivial, concrete example.

In the following, we present *an extract* of the embedding of (extended) DDL in HOL. Our work thereby extends previous work by Benzmüller et al. (2018): Besides adding higher-order quantification, we also extend this embedding to a two-dimensional semantics (Schroeter, 2017) by additionally adding contextual information; for this we use Kaplanian *contexts of use*, cf. Kaplan (1989a,b). The system platform used to implement this ambitious logic combination is the Isabelle proof assistant (Nipkow et al., 2002). In what follows, we are using Isabelle/HOL syntax to render axioms, theorems and definitions (providing the appropriate indications when needed).[3]

2.1   Definition of Types

The type w corresponds to the original type for possible worlds/situations in DDL, cf. Benzmüller et al. (2018). We draw in this work upon David Kaplan's *logic of indexicals/demonstratives* as originally presented in Kaplan (1989a). In Kaplan's logical theory, entities of the aforementioned type w would correspond to his so-called "circumstances of evaluation". Moreover, Kaplan introduces an additional dimension c, so-called "contexts of use", which allow for the modelling of particular context-dependent linguistic expressions, i.e. *indexicals* (see section 2.4). We additionally introduce some type aliases: wo for intensions (also called "contents" or "propositions" in Kaplan's work), which are identified with their truth-sets i.e. the set of worlds at which the proposition is true, and cwo (aliased m) for sentence meanings (also called "characters" in Kaplan's theory), which are modelled as functions from contexts to intensions. Moreover, a type e for individuals is introduced to e.g. enable quantification over individuals.

> **typedecl** w — Type for possible worlds (Kaplan's "circumstances of evaluation")
> **typedecl** c — Type for Kaplan's "contexts of use"
> **typedecl** e — Type for individuals
> **type-synonym** wo = w⇒bool — Type for contents/propositions
> **type-synonym** cwo = c⇒wo — Type for sentence meanings (Kaplan's "characters")
> **type-synonym** m = cwo   — Type alias 'm' for characters

2.2   Embedding of DDL Modal and Deontic Operators

The semantics of DDL draws on Kripke semantics for its (normal) alethic modal operators and on a neighbourhood semantics[4] for its (non-normal) deontic operators. In

---

[3] The formal content of this paper has been generated directly by Isabelle from our source files. A benefit is the prevention of typos. As a side contribution we showcase the usability of modern proof assistants for the non-initiated in order to foster their application.

[4] Neighbourhood semantics is a generalisation of Kripke semantics, developed independently by Dana Scott and Richard Montague. Whereas a Kripke frame features an accessibility relation

order to embed those, we need to introduce the operators av and pv (which can be seen as accessibility relations between worlds), and ob (denoting a neighborhood function operating on sets of worlds) at the meta-logical level. Several axioms, not shown here, adequately constraint the interpretations of av, pv and ob (e.g. $av(w)$ is always a subset of $pv(w)$). See Carmo and Jones (2002) and Benzmüller et al. (2018) for further details.

The following Isabelle/HOL commands illustrate the way logical operators in the target logic (enhanced DDL) can be defined as metalogical predicates using lambda expressions of the appropriate arity/type. The two definitions below, introduced using Isabelle's keyword "abbreviation", realize the embedding of the different modal box and diamond operators (shown here only for $\Box_a$ and $\Diamond_a$). Each of them is embedded as a function from sentence meanings to sentence meanings (type "$m{\Rightarrow}m$"), and they employ (restricted) quantification over possible worlds, following a Kripke semantics.[5]

abbreviation cjboxa :: m⇒m ($\Box_a$-) **where** $\Box_a\varphi \equiv \lambda$c w. $\forall$ v. (av w) v $\longrightarrow$ (φ c v)
abbreviation cjdiaa :: m⇒m ($\Diamond_a$-) **where** $\Diamond_a\varphi \equiv \lambda$c w. $\exists$ v. (av w) v $\wedge$ (φ c v)

The following definitions correspond to the semantical embedding of DDL deontic operators in Isabelle/HOL. The first one represents conditional obligations of the form "φ must be the case given σ" and is embedded as a dyadic relation (type "$m{\Rightarrow}m{\Rightarrow}m$"). The second and third represent the so-called "actual" and "ideal" obligations.

abbreviation cjod :: m⇒m⇒m ($\mathbf{O}\langle$-|-$\rangle$) **where** $\mathbf{O}\langle\varphi|\sigma\rangle \equiv \lambda$c w. ob ($\sigma$ c) (φ c)
abbreviation cjoa :: m⇒m ($\mathbf{O}_a$-) **where**
  $\mathbf{O}_a\varphi \equiv \lambda$c w. (ob (av w)) ($\varphi$ c) $\wedge$ ($\exists$ x. (av w) x $\wedge$ $\neg$(φ c x))
abbreviation cjop :: m⇒m ($\mathbf{O}_i$-) **where**
  $\mathbf{O}_i\varphi \equiv \lambda$c w. (ob (pv w)) ($\varphi$ c) $\wedge$ ($\exists$ x. (pv w) x $\wedge$ $\neg$(φ c x))

2.3 Logical Validity (Classical)

The SSE technique also allows us to embed different notions of logical validity: context-dependent modal validity and general validity (modal validity in each context).

abbreviation modvalidctx :: m⇒c⇒bool ($\lfloor$-$\rfloor^M$) **where** $\lfloor\varphi\rfloor^M \equiv \lambda$c. $\forall$ w. φ c w
abbreviation modvalid :: m⇒bool ($\lfloor$-$\rfloor$) **where** $\lfloor\varphi\rfloor \equiv \forall$ c. $\lfloor\varphi\rfloor^M$ c

2.4 Kaplan's Context Features

Kaplan's theory, originally named "Logic of Demonstratives (LD)" (Kaplan, 1989a,b), aims at modeling the behavior of certain context-sensitive linguistic expressions like the pronouns 'I', 'my', 'it', the demonstrative pronouns 'that', 'this', the adverbs 'here', 'now', 'tomorrow', the adjectives 'actual', 'present', and others. Such expressions are known as *indexicals* and so Kaplan's logical system, among others, is usually referred to as a "logic of indexicals".

$R : W{\to}2^W$ indicating which worlds are alternatives to (or, accessible from) others, a neighborhood frame $N : W{\to}2^{2^W}$ (or, as in our case, $N : 2^W{\to}2^{2^W}$) features a neighbourhood function assigning to each world (or set of worlds) a set of sets of worlds.

[5] Note that in addition to the ASCII name "cjboxa", Isabelle/HOL supports graphical notation "($\Box_a$-)". This is essential for obtaining intuitive mathematical representations.

It is characteristic of an indexical that its content varies with context, i.e. they have a context-sensitive character. Non-indexicals have a fixed character. LD models context-sensitivity by representing contexts as quadruples of features: $\langle Agent(c), Position(c), World(c), Time(c) \rangle$. The agent and the position of context c can be seen as the actual speaker and place of the utterance respectively, while c's world and time stand for the circumstances of evaluation of the expression's content and allow for the interaction of indexicals with alethic and tense modalities respectively. To keep things simple, we restrict ourselves to representing a context c as the pair: $\langle Agent(c), World(c) \rangle$ and model the functional concepts "Agent" and "World" as uninterpreted logical constants. An extension of our work to operate on Kaplan's context quadruples is straightforward.

> **consts** Agent::c$\Rightarrow$e — function retrieving the agent corresponding to context c
> **consts** World::c$\Rightarrow$w — function retrieving the world corresponding to context c

2.5 Indexical Validity

Kaplan's notion of (context-dependent) logical truth for a sentence corresponds to its context-sensitive formula (of type "m", i.e. "$c \Rightarrow w \Rightarrow bool$") being true in the given context and at its corresponding world. Kaplan's notion of logical validity for a sentence requires its truth in all contexts. This notion is also known as indexical validity.

> **abbreviation** ldtruectx::m$\Rightarrow$c$\Rightarrow$bool ($\lfloor$-$\rfloor$-) **where** $\lfloor \varphi \rfloor_c \equiv \varphi$ c (World c)
> **abbreviation** ldvalid::m$\Rightarrow$bool ($\lfloor$-$\rfloor^D$) **where** $\lfloor \varphi \rfloor^D \equiv \forall$c. $\lfloor \varphi \rfloor_c$

The following lemmas show that indexical validity is indeed weaker than its classical modal counterpart (truth at all worlds for all contexts).

> **lemma** $\lfloor A \rfloor \Longrightarrow \lfloor A \rfloor^D$ **by** simp — proven using Isabelle's term-rewriting engine (simp)
> **lemma** $\lfloor A \rfloor^D \Longrightarrow \lfloor A \rfloor$ **nitpick oops** — countermodel

The *countermodel* computed by the model finder *Nitpick* (Blanchette and Nipkow, 2010) for the latter lemma consists of one context c_1 and two worlds w_1 and w_2; where World(c_1) = w_1 and where A holds for c_1 and w_1, but not for c_1 and w_2 (*Nitpick* returns further insightful details which we omit here). Below we use *Nitpick* to show that the interplay between indexical validity and the DDL modal and deontic operators does not result in *modal collapse*. Moreover, we show that the necessitation rule does not work for indexical validity (in contrast to classical modal validity as defined for DDL).

> **lemma** $\lfloor P \rightarrow O_a P \rfloor^D$ **nitpick oops** — countermodel for deontic modal collapse found
> **lemma** $\lfloor P \rightarrow \Box_a P \rfloor^D$ **nitpick oops** — countermodel for alethic modal collapse found
> **lemma** $\lfloor A \rfloor^D \Longrightarrow \lfloor \Box_a A \rfloor^D$ **nitpick oops** — countermodel for necessitation rule found

Below we introduce a kind of "a priori necessity" operator (to be contrasted to the more traditional alethic necessity). This operator satisfies the necessitation rule for indexical validity.[6] In Kaplan's framework, a sentence being logically (i.e. indexically) valid means its being true *a priori*: It is guaranteed to be true in every possible context

---

[6] Note that $\Box^D$ is not part of Kaplan's original system. It has been added by us in order to better highlight some semantic features of our formalization of Gewirth's theory in the next section and for enabling the use of the necessitation rule for drawing inferences.

in which it is uttered, even though it may express distinct propositions (i.e. contents or intensions) in different contexts. This correlation between indexical validity and *a prioricity* has also been claimed in other two-dimensional semantic frameworks (Schroeter, 2017).

> **abbreviation** ldvalidbox :: m⇒m ($\Box^D$-) **where** $\Box^D \varphi \equiv \lambda$c w. $\lfloor \varphi \rfloor^D$
> **lemma** NecLD: $\lfloor A \rfloor^D \Longrightarrow \lfloor \Box^D A \rfloor^D$ **by** simp — necessitation rule proven (term-rewriting)

2.6 Quantification

By utilizing Isabelle/HOL's parameterized types (rank-1 polymorphism), we can easily enrich our logic with (first-order and higher-order) quantifiers.

> **abbreviation** mforall::($'$t⇒m)⇒m ($\forall$) **where** $\forall \Phi \equiv \lambda$c w.$\forall$x. ($\Phi$ x c w)
> **abbreviation** mexists::($'$t⇒m)⇒m ($\exists$) **where** $\exists \Phi \equiv \lambda$c w.$\exists$x. ($\Phi$ x c w)

This definition of embedded parametric quantifiers (which reuses λ-abstraction to avoid the explicit introduction of a new binding mechanism) follows earlier work (Benzmüller and Paulson, 2013). However, it is defined here for Kaplan's sentence meanings and in this sense constitutes another relevant extension of previous work.

3   Representing Gewirth's Ethical Theory

In this section we encode and mechanize Gewirth's (1981) ethical theory —respectively, ethical argument— which aims at justifying an upper moral principle called the "Principle of Generic Consistency" (PGC). In a nutshell, according to this principle, any intelligent agent (by virtue of its self-understanding as an agent) is rationally committed to asserting that (i) it has rights to freedom and well-being, and (ii) all other agents have those same rights. The argument used by Gewirth to derive the PGC (presented in detail in Gewirth (1981); Beyleveld (1991)) is by no means trivial and has stirred much controversy in legal and moral philosophy during the last decades. It has also been discussed in political philosophy as an argument for the *a priori* necessity of human rights (Beyleveld, 2012). Perhaps more relevant for us, the PGC has lately been proposed as a means to bound the impact of artificial general intelligence (AGI) by Kornai (2014).

Kornai draws on Gewirth's PGC as the paradigmatic principle which, assuming it can reliably be represented in a machine, will enable the design of a safety mechanism of a mathematical nature that ensures that an AGI will always respect basic human's rights over all other things. This is based on the assumption that such an intelligent agent is able to recognize itself, as well as humans, as agents acting voluntarily on self-chosen purposes, i.e. as what Gewirth calls: prospective purposive agents (PPA). Every agent designed to follow the PGC will thus be deductively committed, on pain of self-contradiction, to acting in accord with the *generic* rights (i.e. to freedom and well-being) of all agents.[7]

---

[7] Our work constitutes a most relevant first step for further assessment of Kornai's claim. E.g. we plan to embody our encoding of Gewirth's theory in virtual agents and devise and conduct

3.1 Gewirth's Ethical Theory

Gewirth's meta-ethical position is known as moral (or ethical) rationalism. According to it, moral principles are knowable *a priori*, by reason alone. Immanuel Kant is the most famous figure who has defended such a position. He argued for the existence of upper moral principles (e.g. his "categorical imperative") from which we can reason in a top-down fashion to deduce and evaluate other more concrete maxims and actions. In contrast to Kant, Gewirth derives such upper moral principles by starting from purely logical (i.e. non-moral) considerations alone. The argument for the PGC employs what Gewirth calls "the dialectically necessary method" within the "internal viewpoint" of an agent. Although the logical inferences leading to the PGC are drawn relative to the reasoning agent, Gewirth (1981) further argues that *"the dialectically necessary method propounds the contents of this relativity as necessary ones, since the statements it presents reflect judgements all agents necessarily make on the basis of what is necessarily involved in their actions ... The statements the method attributes to the agent are set forth as necessary ones in that they reflect what is conceptually necessary to being an agent who voluntarily or freely acts for purposes he wants to attain."* In other words, the "dialectical necessity" of the assertions and inferences made in the argument comes from the definitional features (i.e. conceptual analysis) of the involved notions of agency, purposeful action, obligation, rights, etc. In order to adequately represent this informal notion of *a priori* dialectical/analytic necessity, we resorted to the formal notion of *indexical validity* as developed in David Kaplan's logical framework LD (Kaplan, 1989a,b).

The cogency of Gewirth's theory will be put to the test in Section 4 by using it to reconstruct his argument (with minor fixes) for the PGC as logically valid. However, we first need to introduce the basic theory itself. To get some inspiration we study the main steps of Gewirth's argument (with original numbering from Beyleveld (1991)):

(1) [Premise] I act voluntarily for some (freely chosen) purpose E —equivalent by definition to: I am a prospective purposive agent (PPA).
(2) E is (subjectively) good —i.e. I value E proactively.
(3) My freedom and well-being (FWB) are generically necessary conditions of my agency —i.e. I need them to achieve any purpose whatsoever.
(4) My FWB are necessary goods (at least for me).
(5) I have (maybe nobody else does) a claim right to my FWB.
(13) [Conclusion] Every PPA has a claim right to their FWB.

In his informal proof, Gewirth claims that the latter generalization step (from "I" to all agents) is done on purely logical grounds and does not presuppose any kind of universal moral principle, and his result is meant to hold with some kind of necessity.[8]

---

respective empirical studies. The merits of the work presented here are however not tied to the validity of Kornai's claim. We illustrate that representation and reasoning with complex ethical theories is meanwhile feasible to an extent unmatched before; and this is highly relevant for implementing explicit ethical intelligent systems. In the following, we will present some commented extracts of our formal encoding of Gewirth's theory and of the computer-supported verification of the argument leading to the PGC.

[8] We were indeed able to formally verify Gewirth's claim, on condition of committing to an alternative notion of (logical) necessity: Kaplan's "indexical validity".

In this respect, Deryck Beyleveld, author of an authoritative book on Gewirth's theory (1991), comments on its first page: *"[Gewirth's] argument purports to establish the PGC as a rationally necessary proposition with an apodictic status* for any PPA *equivalent to that enjoyed by the logical principle of noncontradiction itself."*

In what follows, we provide some *meaning postulates*[9] for the core ethical concepts used to articulate both the PGC and the argument leading to it (as outlined above). We illustrate how to exploit the expressivity of our embedded object logic (DDL enhanced with quantifiers and contexts) to *intuitively* represent and mechanize such a complex ethical theory for the first time in a computer. We also illustrate the utilization of interactive proof assistants (Isabelle/HOL) to assess the argument and to reason with Gewirth's theory.

3.2   Agency

Since Isabelle/HOL is a based on a Church's functional type theory, we need to assign all terms a type. We give "purposes" the same type as sentence meanings (type '$c{\Rightarrow}w{\Rightarrow}bool$' aliased 'm'), so that "acting on a purpose" is represented analogously to having a certain propositional attitude (like "desiring that so and so ..."). The terms "ActsOnPurpose" and "NeedsForPurpose" obtain functional types, and thus expressions like "(ActsOnPurpose A E)" and "(NeedsForPurpose A P E)" are read as "agent A acts on purpose E" and "agent A needs to have property P in order to reach purpose E". We also define a type alias p for properties (functions mapping individuals to characters).

> **type-synonym** p = e⇒m — function from individuals to sentence meanings (characters)
> **consts** ActsOnPurpose:: e⇒m⇒m
> **consts** NeedsForPurpose:: e⇒p⇒m⇒m

In Gewirth's argument, an individual with agency (i.e. capable of purposive action) is said to be a PPA (prospective purposive agent). This definition is supplemented with a meaning postulate stating that being a PPA is an essential (i.e. identity-constitutive) property of an individual. Quite interestingly, this postulate entails a kind of ability for a PPA to recognize other PPAs.[10] For instance, if some individual holds itself as a PPA (seen from its own perspective/context 'd') then this individual 'Agent(d)' is considered a PPA from any other agent's perspective/context 'c'.

> **definition** PPA:: p **where** — Definition of PPA
> **axiomatization where** essentialPPA: $\lfloor \forall$ a. PPA a $\rightarrow \square^D$ (PPA a)$\rfloor^D$
> **lemma** recognizeOtherPPA: $\forall$ c d. $\lfloor$PPA (Agent d)$\rfloor_d \longrightarrow \lfloor$PPA (Agent d)$\rfloor_c$
>   **using** essentialPPA **by** blast — proven using Isabelle blast tactic (tableaux)

---

[9] Definitions and axiomatized conceptual interrelations framing the inferential role of terms. We also refer to them as "explications". Meaning postulates were introduced in Carnap (1952).

[10] Lemma "recognizeOtherPPA" below is indeed inferred from axiom "essentialPPA" using Isabelle's *blast* tactic (a tableaux prover).

3.3 Goodness

Gewirth's concept of (subjective) goodness applies to purposes and is relative to some agent. It is thus modeled as a binary relation relating an individual (of type 'e') with a purpose (of type 'm'). The axioms below are meaning postulates interrelating the concept of goodness with agency and are given as indexically valid sentences (in Kaplan's sense).[11] In particular, we have noticed the need to postulate a further axiom (*explGoodness3*), which represents the intuitive notion of "seeking the good" by asserting that, from an agent's perspective, necessarily good purposes are not only action motivating, but also entail an instrumental obligation to their realization (but only where possible).

> **consts** Good::e⇒m⇒m
> **axiomatization where**
>   explGoodness1: $\lfloor \forall$ a P. ActsOnPurpose a P $\rightarrow$ Good a P $\rfloor^D$
>   explGoodness2: $\lfloor \forall$ P M a. Good a P $\wedge$ NeedsForPurpose a M P $\rightarrow$ Good a (M a) $\rfloor^D$
>   explGoodness3: $\lfloor \forall \varphi$ a. $\Diamond_p \varphi \rightarrow \mathbf{O}\langle \varphi \mid \Box^D$ Good a $\varphi \rangle \rfloor^D$

3.4 Freedom and Well-Being

According to Gewirth, enjoying freedom and well-being (which we take together as the predicate "FWB") is the *contingent* property which represents the "necessary conditions" or "generic features" of agency (i.e. FWB is *always* required in order to be able to act on *any* purpose whatsoever). As before, we take this as an *a priori* characteristic of FWB and therefore axiomatize it as an indexically valid sentence. The last two axioms postulate that FWB is a contingent property.

> **consts** FWB::p — FWB is a property (has type $e{\Rightarrow}m$)
> **axiomatization where**
>   explicationFWB1: $\lfloor \forall$ P a. NeedsForPurpose a FWB P $\rfloor^D$
>   explicationFWB2: $\lfloor \forall$ a. $\Diamond_p$ FWB a $\rfloor^D$
>   explicationFWB3: $\lfloor \forall$ a. $\Diamond_p \neg$FWB a $\rfloor^D$

3.5 Obligation and Interference

Kant's Law ("ought implies can") plays an important role in Gewirth's argument.[12] We have noticed the need to slightly amend it in order to render the argument as logically valid. The new variant reads as: "ought implies *ought to* can". Our variation is indeed closer to Gewirth's (1981, p. 91-95) textual description, that having an obligation to do X implies that *"I ought (in the same sense and the same criterion) to be free to do X, that I ought not to be prevented from doing X, that my capacity to do X ought not to be interfered with."*

> **lemma** $\lfloor \mathbf{O}_i \varphi \rightarrow \Diamond_p \varphi \rfloor$ **using** sem-5ab **by** simp [13]

---

[11] Their higher-order and modal nature well illustrates the need for expressive knowledge representation and reasoning techniques.

[12] This theorem is indeed derivable directly in DDL from the definition of obligations: If φ oughts to obtain then φ is possible.

[13] Here we use Isabelle's *simp* tool to prove that Kant's lemma follows from one of the DDL semantic conditions (not shown here).

axiomatization where OIOAC: $\lfloor \mathbf{O}_i\varphi \to \mathbf{O}_i(\Diamond_a\varphi) \rfloor^D$

Concerning the concept of interference, we have noticed the need to presume that the existence of an individual *b* (successfully) interfering with some state of affairs φ implies that φ cannot possibly be obtained in any of the actually possible situations (and the other way round). This axiom implies that if someone (successfully) interferes with agent *a* having FWB, then *a* can no longer possibly enjoy its FWB (and the converse).

> **consts** InterferesWith::e⇒m⇒m
> **axiomatization where** explicationInterference: $\lfloor (\exists \, b. \, \text{InterferesWith b } \varphi) \leftrightarrow \neg\Diamond_a\varphi \rfloor$
> **lemma** InterferenceWithFWB: $\lfloor \forall \, a. \, (\exists \, b. \, \text{InterferesWith b (FWB a)}) \leftrightarrow \neg\Diamond_a(\text{FWB a}) \rfloor$
>   **using** explicationInterference **by** blast

3.6   Rights and Other-Directed Obligations

Gewirth (1981, p. 66) points out the existence of a correlation between an agent's own claim rights and other-referring obligations. A claim right is a right which entails duties or obligations for other agents regarding the right-holder (so-called Hohfeldian claim rights in legal theory). We model this concept of claim rights in such a way that an individual *a* has a (claim) right to having some property φ if and only if it is obligatory that every (other) individual *b* does not interfere with the state of affairs $(\varphi \, a)$. Since there is no particular individual to whom this directive is addressed, this obligation has been referred to by Gewirth as being "other-directed" (aka. "other-referring") in contrast to "other-directing" obligations which entail a moral obligation for some particular subject (Beyleveld, 1991, p. 41, 51). This latter distinction is essential to Gewirth's argument.

> **definition** RightTo::e⇒(e⇒m)⇒m **where** RightTo a $\varphi \equiv \mathbf{O}_i(\forall \, b. \, \neg\text{InterferesWith b } (\varphi \, a))$

Now that all axioms of the theory are in place, we need to show that they are indeed logically consistent. For this we use Isabelle's model finder *Nitpick* to compute a corresponding model (not shown here) having one context, one individual and two worlds.

> **lemma** True **nitpick**[satisfy, card c = 1, card e = 1, card w = 2] **oops** — model found

4   Reasoning with Gewirth's Ethical Theory

The PGC can be seen as a particular variant (or emendation) of the *golden rule*: treating others as one's self would wish to be treated. A self-acknowledged agent (i.e. a PPA) would read the PGC as a moral commandment: "I ought to act in accord with the generic rights of my recipients as well as of myself" (Gewirth, 1981, p. 153). Urging a fellow human being to obey such a principle without having explained its deeper rationale will presumably at best elicit an absent-minded, cursory acknowledgment. The difficulty here lies not only in the lack of understanding or agreement of what the given words mean (what is a "generic right"?), but also in the addressee's lack of 'immersion' in the underlying conceptual framework and the inferential practices behind such a principle (an unaware addressee would not be able to infer a third-party obligation from a right

claim). In short, any moral principle *qua sentence* makes best sense in the context of the background theory from which it is obtained as a well-founded part; this has been argued e.g. by the philosopher W. V. O. Quine in his holistic view of meaning (cf. 1960).

This situation is not much different for machines. In order to correctly interpret and apply an ethical principle, we need to (i) determine the meaning of its constituent concepts (action/agency, right, freedom and well-being, etc.); and (ii) determine the meaning of other relevant concepts (goodness, necessity, interference, obligation, etc.) playing a role in its articulation (and justification) within the underlying theory. Talk of meanings can be obscure, so let us put it in model-theoretical terms: The set of models of the logical theory has to be constrained to properly fit the target conceptualization (i.e. to only entail intended models). These constraints are set by meaning postulates, i.e. axioms and definitions. Their adequacy can be assessed by studying the extent to which they enable the validation (or invalidation) of candidate theorems (or non-theorems). As is already known, the main theorem we aim at validating here is the PGC, suitably paraphrased as: *Every PPA has a claim right to its freedom and well-being.* The reconstructed proof in Isabelle/HOL of the theorem below is shown in Fig. 1.

theorem PGC: **shows** $\forall$ C. $\lfloor$PPA (Agent C) $\rightarrow$ (RightTo (Agent C) FWB)$\rfloor_C$

In Sections 2 and 3, besides from formally articulating Gewirth's theory, we have used some of Isabelle's proof methods (simp, blast, etc.) and the *Nitpick* model finder to verify some relevant inferences and to guarantee consistency, thus the theory's adequacy has already partly been assessed. In addition, we have used a combination of interactive and automated theorem proving to reconstruct Gewirth's argument for the PGC as logically valid by formally proving it within the complex logical framework built so far. We thus contribute an exemplary case study illustrating how to reason with highly-expressive formal representations of complex, natural-language ethical theories by harnessing the power of higher-order theorem provers (drawing on the SSE approach). In the argument's reconstruction as displayed in Fig. 1, some of the intermediate inference steps leading to the main conclusion (PGC) have indeed been hinted at by automated tools; cf. Fuenmayor and Benzmüller (2019, 2018) for further details. In particular, some missing implicit premises (not considered in Gewirth's original argument) have been uncovered, namely the explications of the concepts of *goodness* and *interference* and the amendment to Kant's Law: "ought implies *ought to* can". Note that the mechanized argument matches the granularity-level as can also be found in human constructed informal arguments, and all the sub-arguments (sub-proofs) can automatically be found by automated theorem proving technology. Moreover, the whole proof as presented can be automatically verified using a standard laptop in under a second.

5   Related Work and Summary

We achieve several improvements over related work such as Bringsjord et al. (2006) and Furbach and Schon (2015): (i) Due the use of enriched DDL (enabled by our hi¡gher-order meta-logic) we are not suffering from contrary-to-duty issues; (ii) we make use of truly higher-order encodings as required for the adequate modeling of the PGC; (iii)

Fig. 1: Gewirth's proof encoded in the Isabelle/HOL proof assistant.

we overcome unintuitive, machine-oriented formula representations; and (iv) we do not stop with supporting proof automation, but combine it with intuitive user interaction. Combinations of (i)–(iv) also apply to more recent related work by Govindarajulu and Bringsjord (2017), Hooker and Kim (2018) and Pereira and Saptawijaya (2016), which are not applicable to complex theories like Gewirth's PGC without considering significant simplifications (accepting e.g. contrary-to-duty issues is potentially dangerous).

Utilizing a semantical embedding of a suitable combination of expressive non-classical logics in meta-logic HOL, an ambitious ethical theory, Gewirth's PGC, has exemplarily been encoded and mechanized on the computer. Our methodology supports both highly intuitive representation of and interactive-automated reasoning with the encoded theory. Automated theorem provers have even helped to reveal some hidden issues in Gewirth's argument. The presented methodology is motivating research in different, albeit related, directions: (i) for conducting analogous formal assessments of further ambitious ethical theories, and (ii) for progressing with the implantation of explicit ethical reasoning competencies in future intelligent autonomous systems *by adapting state-of-the-art theorem proving technology and by combining the expertise of different research communities*.

Bibliography

Michael Anderson and Susan Leigh Anderson. Geneth: a general ethical dilemma analyzer. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Christoph Benzmüller. Universal (meta-)logical reasoning: Recent successes. *Science of Computer Programming*, 172:48–62, 2019. https://doi.org/10.1016/j.scico.2018.10.008. Url (preprint): http://doi.org/10.13140/RG.2.2.11039.61609/2.

Christoph Benzmüller and Peter Andrews. Church's type theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. URL https://plato.stanford.edu/entries/type-theory-church/.

Christoph Benzmüller and Lawrence Paulson. Quantified multimodal logics in simple type theory. *Logica Universalis (Special Issue on Multimodal Logics)*, 7(1):7–20, 2013. https://doi.org/10.1007/s11787-012-0052-y.

Christoph Benzmüller, Ali Farjami, and Xavier Parent. A dyadic deontic logic in HOL. In Jan Broersen, Cleo Condoravdi, Shyam Nair, and Gabriella Pigozzi, editors, *Deontic Logic and Normative Systems — 14th International Conference, DEON 2018, Utrecht, The Netherlands, 3-6 July, 2018*, pages 33–50. College Publications, 2018. ISBN 978-1-84890-278-7. John-Jules Meyer Best Paper Award.

Christoph Benzmüller, Xavier Parent, and Leendert W. N. van der Torre. Designing normative theories of ethical reasoning: Formal framework, methodology, and tool support. *CoRR*, abs/1903.10187, 2019. URL http://arxiv.org/abs/1903.10187.

Deryck Beyleveld. *The dialectical necessity of morality: An analysis and defense of Alan Gewirth's argument to the principle of generic consistency*. Univ. of Chicago Press, 1991.

Deryck Beyleveld. The principle of generic consistency as the supreme principle of human rights. *Human Rights Review*, 13(1):1–18, 2012. ISSN 1874-6306.

Jasmin C. Blanchette and Tobias Nipkow. Nitpick: A counterexample generator for higher-order logic based on a relational model finder. In *Proc. of ITP 2010*, volume 6172 of *LNCS*, pages 131–146. Springer, Lecture Notes in Computer Science, 2010. ISBN 978-3-642-14051-8.

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intellig. Systems*, 21(4):38–44, 2006.

José Carmo and Andrew J.I. Jones. Deontic logic and contrary-to-duties. In *Handbook of Philosophical Logic*, pages 265–343. Springer, 2002.

Rudolf Carnap. Meaning postulates. *Philosophical studies*, 3(5):65–73, 1952.

Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016. URL https://doi.org/10.1016/j.robot.2015.11.012.

Virginia Dignum. Responsible autonomy. In *IJCAI-17*, pages 4698–4704, 2017.

Virginia Dignum (ed.). Special issue: Ethics and artificial intelligence. *Ethics and Information Technology*, 20(1), 2018.

David Fuenmayor and Christoph Benzmüller. Formalisation and evaluation of Alan Gewirth's proof for the principle of generic consistency in Isabelle/HOL. *Archive of Formal Proofs*, 2018. URL https://www.isa-afp.org/entries/GewirthPGCProof.html.

David Fuenmayor and Christoph Benzmüller. Isabelle/HOL sources associated with this PRICAI-2019 paper. Online available at http://bit.ly/Appendix-PRICAI-19, 2019.

Ulrich Furbach and Claudia Schon. Deontic logic for human reasoning. In *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation*, volume 9060 of *LNCS*, pages 63–80. Springer, 2015.

Alan Gewirth. *Reason and morality*. University of Chicago Press, 1981.

Naveen Sundar Govindarajulu and Selmer Bringsjord. On automating the doctrine of double effect. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4722–4730, 2017. URL https://doi.org/10.24963/ijcai.2017/658.

John N Hooker and Tae Wan N Kim. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 130–136. ACM, 2018.

David Kaplan. Demonstratives. In Joseph Almog, John Perry, and Howard Wettstein, editors, *Themes from Kaplan*, pages 481–563. Oxford University Press, 1989a.

David Kaplan. Afterthoughts. In Joseph Almog, John Perry, and Howard Wettstein, editors, *Themes from Kaplan*, pages 565–612. Oxford University Press, 1989b.

András Kornai. Bounding the impact of AGI. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):417–438, 2014.

Bertram F Malle. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18(4):243–256, 2016.

James Moor. Four kinds of ethical robots. *Philosophy Now*, 72:12–14, 2009.

Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*, volume 2283 of *LNCS*. Springer, Lecture Notes in Computer Science, 2002.

Luís Moniz Pereira and Ari Saptawijaya. *Programming machine ethics*. Springer, 2016.

Willard Van Orman Quine. *Word and Object*. MIT press, 1960.

Matthias Scheutz. The case for explicit ethical agents. *AI Magazine*, 38(4):57–64, 2017.

Laura Schroeter. Two-dimensional semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017.

Wendell Wallach, Colin Allen, and Iva Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22(4):565–582, 2008.