Abstract. We discuss ongoing work on reusing existing (higher-order) automated reasoning infrastructure for seamlessly combining and reasoning with different non-classical logics (modal, deontic, epistemic, paraconsistent, etc.), particularly suited for normative reasoning. Our work illustrates, in particular, the utilisation of the Isabelle/HOL proof assistant for the representation and formal assessment of linguistically complex ethical arguments. Our work pushes existing boundaries in knowledge representation and reasoning. We demonstrate that intuitive, formal encodings of complex ethical theories and their automation on the computer are no longer antipodes.

1 Motivation

Hybrid architectures for ethical autonomous agents that integrate both bottom-up learning and top-down deliberation from upper principles are receiving increased attention; cf. [13, 12, 25, 21, 11, 1, 27] and the references therein. Irrespective of the preferred direction, it is becoming evident that adequate explicit representations of ethical knowledge are beneficial, if not mandatory, to obtain satisfactory solutions. Bottom-up approaches benefit from expressive languages to explicitly represent the learned ethical knowledge in a scrutable, communicable and transferable manner. Top-down approaches have to rely on expressive logic languages to enable an intuitive and accurate representation and reasoning with ethical theories. Unfortunately, however, only few approaches are currently available that enable adequate and realistic, explicit formal encodings of non-trivial ethical theories, and that at the same time support intuitive interactive-automated reasoning with them.

2 Framework

Our framework relies on the utilisation of (higher-order) automated reasoning infrastructure for seamlessly combining and reasoning with different non-classical logics (modal, deontic, epistemic, paraconsistent, etc.) as suited for a given application context. Our approach to combining logics is based on a technique called shallow semantical embeddings (SSE) [2, 6]. SSEs harness the high expressive power of classical higher-order logic (HOL), aka. Church’s type theory [3], as a meta-language in order to embed the syntax and semantics of (combinations of) object logics.

A SSE for an object logic corresponds to adding a set of axioms and definitions to the expressive meta-logic (HOL) in such a way as to encode the connectives of the object logic as meta-logical expressions. This has interesting practical implications; for example, the semantically embedded (combinations of) object logics can easily be varied by adding or removing (meta-logical) sentences, thereby enabling their rapid prototyping and formal verification. Moreover, the approach scales for quantified object logics and, due to the expressivity of HOL, it is possible to directly encode bridge rules, or, as an alternative, their corresponding semantic counterparts [2, 15].

The framework and techniques we present, cf. [14] and [5], 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 [22].

3 Ethical Theories and NL Arguments

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, generalised quantifiers, (conditional) obligations, among several others.

In previous work [14, 15] we have introduced and justified (by examples) a logical normative reasoning system requiring the extension and combination of a dyadic deontic logic (DDL) [9] with higher-order quantification and a 2D-Semantics [26] drawing on Kaplan’s logic of indexicals [20]. The logic DDL has been encoded for the first time in the Isabelle/HOL proof assistant [23] shortly before [4]. Further extensions of DDL (including context sensitivity, quantification, etc.) have subsequently been implemented, and the extended DDL has been shown stable against different versions of Chisholm’s (aka. contrary-to-duty) paradox [10] as intended [15].

Regarding the combined logics, conditional obligations in DDL are of a defeasible and paraconsistent nature, and thus lend themselves to reasoning with incomplete and inconsistent knowledge. Kaplan’s logic of indexicals aims at modelling the behaviour of certain context-sensitive linguistic expressions known as indexicals (such as pronouns, demonstrative pronouns, and some adverbs and adjectives). It is characteristic of an indexical that its content varies with context, i.e., they have a context-sensitive character. We have modelled Kaplanian contexts by introducing a new type of object (context) and by modelling sentence meanings as so-called “char-

---

1 Freie Universität Berlin, Germany, email: david.fuenmayor@fu-berlin.de
2 Freie Universität Berlin and Université du Luxembourg, Luxembourg, email: c.benzmueller@fu-berlin.de
acers” [20], i.e., functions from contexts to sets of possible worlds (following a Kripke semantics). For simplicity of exposition, we have omitted tenses in our treatment of Kaplan’s logical theory.

This way, we have illustrated how a non-trivial combination of logics can be stepwise developed and formally assessed [15]. In particular, we demonstrated the utilisation of the SSE approach within the Isabelle/HOL proof assistant for the representation and assessment of complex linguistic phenomena in normative arguments and theories and also motivated applications of the combined logic for the encoding of challenging ethical theories.

Utilising a similar logic combination as above, an ambitious ethical theory: Alan Gewirth’s “Principle of Generic Consistency” (PGC) [17], has been exemplarily encoded and Gewirth’s justifying argument has been reconstructed and assessed on the computer [14]. We showed how our approach 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.

4 Related Work and Summary

Such a rich and heterogeneous combination of expressive logics as utilised in our work [14, 15] has not been automated before. By allowing higher-order quantification (e.g. as required by Gewirth’s argument for the PGC) and being immune, among others, to contrary-to-duty paradoxes, the mechanisation of this particular logic combination also constitutes an improvement over related work on automated deontic reasoning (e.g., [8, 16]): (i) Due the use of enriched DDL (enabled by our higher-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 non-trivial ethical theories (e.g. Gewirth’s PGC [14]); (iii) we overcome uninitiative, 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 (e.g., [18, 19, 7, 24]), which is not applicable to complex theories such as Gewirth’s PGC without considering significant simplifications and abstractions, which may lead to potentially dangerous behaviour, e.g., in the case of contrary-to-duty paradoxes.

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.

REFERENCES


