24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Normative Reasoning with Expressive Logic Combinations

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

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

acters" [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[3] 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 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 (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

[1] Michael Anderson and Susan Leigh Anderson, 'Geneth: a general ethical dilemma analyzer', *Paladyn*, **9**(1), 337–357, (2018).
[2] Christoph Benzmüller, 'Universal (meta-)logical reasoning: Recent successes', *Science of Computer Programming*, **172**, 48–62, (2019).
[3] Christoph Benzmüller and Peter Andrews, 'Church's type theory', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, Metaphysics Research Lab, Stanford University, summer 2019 edn., (2019).
[4] Christoph Benzmüller, Ali Farjami, and Xavier Parent, 'A dyadic deontic logic in HOL', in *Deontic Logic and Normative Systems — 14th International Conference, DEON*, eds., J. Broersen, C. Condoravdi, S. Nair, and G. Pigozzi, pp. 33–50. College Publications, (2018).

[3] The SSE approach has also been illustrated by formalising the "Wise Men Puzzle" (a riddle in multi-agent epistemic reasoning) [2].

[5] 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).
[6] Christoph Benzmüller and Lawrence Paulson, 'Quantified multimodal logics in simple type theory', *Logica Universalis*, **7**(1), 7–20, (2013).
[7] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia, 'A declarative modular framework for representing and applying ethical principles', in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, eds., K. Larson, M. Winikoff, S. Das, and E.H. Durfee, pp. 96–104. ACM, (2017).
[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).
[9] José Carmo and Andrew J.I. Jones, 'Deontic logic and contrary-to-duties', in *Handbook of Philosophical Logic*, ed., Guenthner F. Gabbay D.M., volume 8, 265–343, Springer, (2002).
[10] Roderick M. Chisholm, 'Contrary-to-duty imperatives and deontic logic', *Analysis*, **24**, 33–36, (1963).
[11] 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).
[12] Virginia Dignum, 'Responsible autonomy', in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, ed., C. Sierra, pp. 4698–4704. ijcai.org, (2017).
[13] Virginia Dignum (ed.), 'Special issue: Ethics and artificial intelligence', *Ethics and Information Technology*, **20**(1), (2018).
[14] David Fuenmayor and Christoph Benzmüller, 'Harnessing higher-order (meta-)logic to represent and reason with complex ethical theories', in *PRICAI 2019: Trends in Artificial Intelligence*, eds., A. Nayak and A. Sharma, volume 11670 of *LNAI*, pp. 418–432. Springer, (2019).
[15] David Fuenmayor and Christoph Benzmüller, 'Mechanised assessment of complex natural-language arguments using expressive logic combinations', in *Frontiers of Combining Systems, 12th International Symposium, FroCoS*, eds., A. Herzig and A. Popescu, volume 11715 of *LNAI*, pp. 112–128. Springer, (2019).
[16] Ulrich Furbach and Claudia Schon, 'Deontic logic for human reasoning', in *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation - Essays Dedicated to Gerhard Brewka on the Occasion of His 60th Birthday*, eds., T. Eiter, H. Strass, M. Truszczynski, and S. Woltran, volume 9060 of *LNCS*, pp. 63–80. Springer, (2015).
[17] Alan Gewirth, *Reason and morality*, University of Chicago Press, 1981.
[18] 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*, ed., C. Sierra, pp. 4722–4730. ijcai.org, (2017).
[19] John N. Hooker and Tae Wan 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, AIES*, eds., J. Furman, G. E. Marchant, H. Price, and F. Rossi, pp. 130–136. ACM, (2018).
[20] David Kaplan, 'Demonstratives', in *Themes from Kaplan*, eds., J. Almog, J. Perry, and H. Wettstein, 481–563, Oxford University Press, (1989).
[21] 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).
[22] James Moor, 'Four kinds of ethical robots', *Philosophy Now*, **72**, 12–14, (2009).
[23] Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel, *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*, volume 2283 of *LNCS*, Springer, 2002.
[24] Luís Moniz Pereira and Ari Saptawijaya, *Programming machine ethics*, Springer, 2016.
[25] Matthias Scheutz, 'The case for explicit ethical agents.', *AI Magazine*, **38**(4), 57–64, (2017).
[26] Laura Schroeter, 'Two-dimensional semantics', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, Metaphysics Research Lab, Stanford University, summer 2017 edn., (2017).
[27] 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).