Explicit Normative Reasoning and Machine Ethics*

Christoph Benzmüller[1] and Geoff Sutcliffe[2]

[1] FU Berlin, Germany & University of Luxembourg, Luxembourg, c.benzmueller@gmail.com
[2] University of Miami, USA, geoff@cs.miami.edu

Abstract

The proposed Explicit Normative Reasoning and Machine Ethics (ENoRME) project will contribute foundational technology to foster and enable the development of ethical and legal governors for intelligent autonomous systems. These governors will provide symbolic, deductive means of control that operate orthogonally to and in addition to mechanisms at the less transparent, less accountable, and less trustworthy level of subsymbolic reasoning. In particular, ENoRME will develop a much needed, powerful and flexible, on-demand universal reasoning workbench, with a particular emphasis on normative reasoning. It will explore and demonstrate the use of the workbench in selected case studies and experiments. ENoRME will contribute essential input to the development of "Trustworthy AI".

1 Introduction

Intelligent autonomous systems (IASs) are rapidly entering applications in industry, military, finance, governance, administration, healthcare, etc., leading to a transition period with unprecedented dynamics of innovation and change, and with unpredictable outcomes [9]. Legislatures, regulatory bodies, intergovernmental organizations, etc., indeed society as a whole, are challenged not only with keeping pace with these potentially disruptive developments, but also with staying ahead and wisely guiding the transition. Since there is much at stake, preemptive investments in the exploration, development, and implementation of appropriate means of controlling IASs are absolutely justified. A balanced approach to AI is needed, fostering positive impacts while preventing negative side effects. This vision is shared in particular by the European Commissions (EC) High-Level Expert Group (HLEG) on Artificial Intelligence (AI) in their recently published "Ethics Guidelines for Thrustworthy AI" [15].

Many AI researchers, developers, and users are concerned with challenges such as "How to stay ahead ...", "How to keep up with ...", and "How to close the gap to ..." the rapid developments in AI. A global technology race is the result. A much smaller community is concerned with researching and developing means of wisely regulating and controlling future IASs. The proposed **Explicit Normative Reasoning and Machine Ethics** (ENoRME) project addresses these issues. The primary focus is on risk prevention, more precisely, on

*The ENoRME project we outline in this document has its roots in collaborative prior research activities, and the proposal received significant input, among others, from David Fuenmayor, Xavier Parent, Raúl Rojas, Alexander Steen and Leon van der Torre. Benzmüller received funding from VolkswagenStiftung under grant CRAP (Consistent Rational Argumentation in Politics).

explicit ethical governance of IASs [1]. We see it as a societal mandate to invest in risk prevention technology; such a sensitive issue should not be driven by commercial interests alone. Despite of this focus on "Trustworthy AI", the range of AI applications that will be enabled by ENoRME is wide and colourful.

2 The State of the Art

The questions of how transparency, explainability, and verifiability can best be achieved in future IASs, and whether bottom-up or top-down architectures should be preferred, are discussed in various papers, e.g., [7, 17, 8, 20]. Dennis et al. [7] make a compelling case for the use of formal verification – a well-established technique for proving correctness of computer systems – in the area of machine ethics. Earlier proposals towards explicit ethical governance of intelligent machines include [2] and [14]. ENoRME will develop a reasoning infrastructure for such normative reasoning. Such an infrastructure is much needed, and related work is thin.

Research and concrete results for the automation of reasoning for quantified deontic logics, and their combinations with other non-classical logics (epistemic, temporal, defeasible, probabilistic, etc.) as required in realistic applications of normative reasoning, is sparse. A connection-based ATP covering among others, first-order standard deontic logic, has been developed by Otten [16]. A tableaux-based propositional reasoner is employed in the work of Furbach and Schon [11]. First-order resolution methods for propositional modal logics have been contributed by Schmidt and Hustadt [18]. Steen and Benzmüller developed Leo-III [19], which has been adapted to the automation of more than 120 different quantified (multi-)modal logics [12], and also for an initial range of quantified deontic logics. Benzmüller, Parent, and van der Torre have collaborated on the automation of propositional and quantified deontic logics, on the development of an overall methodology for the automation of normative reasoning, and on the use of those in some small, selected case studies [4, 3].

There are some first-steps in the development of logical architectures for ethical reasoning that are particularly relevant for the planned work in ENoRME. For example, Carnielli and Bueno-Soler are leading experts in paraconsistent reasoning and logic combinations [5, 6]. Slavkovik, Liao, and van der Torre have collaborated on a methodological level to develop an artificial moral agent architecture that uses techniques from normative systems and formal argumentation to reach moral agreements among stakeholders [13]. Dennis and Fisher have studied ethically critical machine reasoning, and have proposed practical architectures and reasoning tools [7]. Fuenmayor and Benzmüller have formalised, using the ENoRME approach, some first ambitious ethical theory [10].

3 ENoRMEous Plans

Progress in Machine Learning (ML), in particular deep learning, has enabled impressive recent success stories in the development and deployment of IASs. ENoRME will target pressing challenges that these successes have generated. A major concern is that IASs that rely exclusively on non-symbolic technologies increasingly lack transparency, explainability, verifiability, and ethical behaviour – the essential requirements for "Trustworthy AI" [15].[1] A particular

[1] We acknowledge the valuable efforts being made inside the deep learning research community towards the extraction of higher-level representations from modularised neural networks. However, there are still many open research questions. Regarding interpretability of outcomes (cf. "explainable AI"), such mechanisms are rather laborious and only useful in a kind of 'forensic' way. AI systems need explicit 'on-line' interpretability.

challenge concerns the development of mechanisms of ethical and legal control for future IASs, such as the ability to construct and reason with high-level conceptual representations of ethical and legal concepts. A convincing solution must fruitfully integrate non-symbolic with symbolic techniques, based on explicit knowledge representation and reasoning.

The success of deep learning is based on the availability of huge amounts of data, and relies on the impressive growth in hardware capabilities that allow for massively parallel data processing. The area of automated reasoning in expressive logics is lagging behind in this regard. A building block that is missing on the symbolic side is powerful automated reasoning technology for expressive non-classical logics and their combinations, as required for realising adequate forms of automated and semi-automated normative reasoning. The development of competitive automated reasoning technology for such ambitious logics typically requires substantial resource investment and expertise, which is typically not available to small research teams or small-to-medium sized enterprises (SMEs). It is clear that any strategy aimed at democratizing access to AI, in particular for SMEs and educational institutions, must facilitate their access to such critical hardware resources. Developing, sharing, and standardising such technology to stimulate research and deployment is one important mission of ENoRME.

The ENoRME vision has been specifically geared towards simultaneously addressing these two concerns: developing mechanisms of ethical and legal control, and provding the necessary reasoning infrstructure for that reasoning. In addition to boosting knowledge transfer of research outcomes with the scientific community, industry, and the public sector, ENoRME will address the real need of SMEs and educational institutions for experimenting with AI *at scale*, with little risk and low cost. A primary focus thereby is on the provision of technology for automating expressive normative reasoning. This is beneficial for AI research in a wide sense, not only for the legal and ethical governing of IASs. ENoRME will develop, as its core deliverable, an integrated *cloud-based, on demand, Universal Reasoning Workbench (URW) with a specific focus on normative reasoning*. This URW will support modelling and automated experimentation with explicit theories for use in ethical governors within IASs. The URW will also find application in other areas, e.g., rational argumentation, natural language processing, computational metaphysics, etc. Moreover, all of ENoRME's solutions will be applicable both independently and in combination with alternative means based on ML and other non-symbolic approaches. The overall objective is to contribute a missing *explicit reasoning* component in the development of secure, explainable, and trustworthy IASs.

At the same time as providing pragmatic solutions, ENoRME will engage in theoretical and experimental research in explicit ethical governing mechanisms for IASs. In this sense ENoRME will become its own first customer. The idea is to create an immediate feedback loop between (i) the URW, (ii) the ethical governing architectures, and (iii) their experimental application in selected case studies with our industrial partners. This feedback loop will then particularly inform our work on the URW.

The specific deliverables of ENoRME are as follows: • A cloud-based, on demand, URW, including an associated experimentation platform; • Standardisation and benchmarking in normative reasoning; • Theory and implementation of novel logic combinations; • Examples of mechanised/automated ethical and legal theories; • Specialist ATP and model finding technology for expressive normative reasoning; • An agent-based simulation environment enabling experiments with ethical and legal theories; • Case studies conducted within this agent-based simulation environment to assess different ethical and legal theories; • Exemplary implementation, experimentation, and assessment of ethical governor architectures equipped with ethical and legal theories; • Conferences, workshops, and tutorials on the topics of ENoRME; • Lecture courses, seminars, eLearning resources, PhD and student projects, all on the topic of ENoRME;

• Outreach: public debates, media appearances, science slams, website, Facebook, Twitter, etc.

References

[1] R. Arkin, P. Ulam, and B. Duncan. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical Report Technical Report GIT-GVU-09-02, Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology, Atlanta, USA, 2009.

[2] K. Arkoudas, S. Bringsjord, and P. Bello. Toward Ethical Robots via Mechanized Deontic Logic. In M. Anderson, S.L. Anderson, and C. Armen, editors, *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*, number FS-05-06 in Technical Report, pages 17–23. AAAI Press, 2006.

[3] C. Benzmüller, X. Parent, and L. van der Torre. A Deontic Logic Reasoning Infrastructure. In F. Manea, R.G. Miller, and D. Nowotka, editors, *Proceedings of the 14th Conference on Computability in Europe*, number 10936 in Lecture Notes in Computer Science, pages 60–69, 2018.

[4] C. Benzmüller, X. Parent, and L. van der Torre. Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support. arXiv:1903.10187, 2019.

[5] J. Bueno-Soler and W. Carnielli. Paraconsistent Probabilities: Consistency, Contradictions and Bayes' Theorem. *Entropy*, 18(9):325, 2016.

[6] W. Carnielli, M. Coniglio, D. Gabbay, P. Gouveia, and C. Sernadas. *Analysis and Synthesis of Logics - How to Cut and Paste Reasoning Systems*. Number 35 in Applied Logic Series. Springer Verlag, 2008.

[7] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[8] V. Dignum, editor. *Special Issue: Ethics and Artificial Intelligence*, volume 20, 2018.

[9] L. Floridi. What the Near Future of Artificial Intelligence Could Be. *Philosophy & Technology*, 32(1):1–15, 2019.

[10] D. Fuenmayor and C. Benzmüller. Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories. In A. Nayak and M. Khan, editors, *Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence*, Lecture Notes in Artificial Intelligence, page To Appear. Springer-Verlag, 2019.

[11] U. Furbach and C. Schon. Deontic Logic for Human Reasoning. In T. Eiter, H. Strass, M. Truszczyński, and S. Woltran, editors, *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation*, number 9060 in Lecture Notes in Computer Science, pages 63–80. Springer-Verlag, 2015.

[12] T. Gleissner, A. Steen, and C. Benzmüller. Theorem Provers for Every Normal Modal Logic. In T. Eiter and D. Sands, editors, *Proceedings of the 21st International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, number 46 in EPiC Series in Computing, pages 14–30. EasyChair Publications, 2017.

[13] B. Liao, M. Slavkovik, and L. van der Torre. Building Jiminy Cricket: An Architecture for Moral Agreements among Stakeholders. In V. Conitzer, G. Hadfield, A. McAfee, and S. Vallor, editors, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, page To appear, 2019.

[14] G-J. Lokhorst. Computational Meta-Ethics. *Minds and Machines*, 21(2):261–274, 2011.

[15] European Commission Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Thrustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation, 2019.

[16] J. Otten. Non-clausal Connection Calculi for Non-classical Logics. In C. Nalon and R. Schmidt, editors, *Proceedings of the 26th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*, number 10501 in Lecture Notes in Artificial Intelligence, pages 209–227. Springer-Verlag, 2017.

[17] M. Scheutz. The Case for Explicit Ethical Agents. *AI Magazine*, 38(4):57–64, 2017.

[18] R. Schmidt and U. Hustadt. First-Order Resolution Methods for Modal Logics. In A. Voronkov and C. Weidenbach, editors, *Programming Logics, Essays in Memory of Harald Ganzinger*, number 7797 in Lecture Notes in Computer Science, pages 345–391. Springer-Verlag, 2013.

[19] A. Steen and C. Benzmüller. The Higher-Order Prover Leo-III. In D. Galmiche, S. Schulz, and R. Sebastiani, editors, *Proceedings of the 8th International Joint Conference on Automated Reasoning*, number 10900 in Lecture Notes in Artificial Intelligence, pages 108–116, 2018.

[20] A. Winfield and M. Jirotka. Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180085, 2018.