Thesis Proposal

Title

Testing real time compliance of artificial intelligent agent actions with ethical rules of conduct

Summary

The thesis will take place in the tri-lateral (French-German-Japanese) RECOMP research action that aims at ensuring the artificial intelligent agents compliance with legal and ethical norms. More specifically, the proposed research is dedicated to computational ethics i.e. to model ethical concepts and ethical reasoning with AI tools in the context of the RECOMP project. The practical goal will be to design real time ethical supervisors for artificial agents, which are pieces of software that have to govern artificial agents, e.g. autonomous vehicles, robots or healthcare artificial agents, in order to ensure that they don't infringe moral rules, except in specific situations where ethical norms are conflicting. This doesn't mean that the agents that would result be, properly speaking, ethical, which would be ridiculously anthropocentric, but that ***such an ethical supervisor makes the agents' behaviors to conform as far as possible to moral rules of conducts****. The candidate will have to ensure the compatibility of the representation and the mechanism he will use with the representations used by the other partners of the project, which are German and Japanese. The work will be done in ACASA-LIP6 team, on the Pierre and Marie Curie Campus, under the supervision of Jean-Gabriel Ganascia and Gauvain Bourgne who are both Professor at Sorbonne University (Paris – France).*

Contact: Jean-Gabriel.Ganascia@lip6.fr and Gauvain.Bourgne@lip6.fr

Description of the project

The thesis is dedicated to computational ethics i.e. to model ethical concepts and ethical reasoning with AI tools. The practical goal will be to design real time ethical supervisors for artificial agents, which are pieces of software that have to govern artificial agents, e.g. autonomous vehicles, robots or healthcare artificial agents, in order to ensure that they don't infringe moral rules, except in specific situations where ethical norms are conflicting. This doesn't mean that the agents that would result be, properly speaking, ethical, which would be ridiculously anthropocentric, but that ***such an ethical supervisor makes the agents' behaviors to conform as far as possible to moral rules of conducts****.*

This research will have to build artificial agents able to overcome in ***real time*** *difficulties that arise from* ***conflicting norms****, which generate logical contradictions between legal and ethical duties that have all to be obeyed. The logic-based AI allows, using non-monotonic formalisms as default logics or stable models, to overcome these contradictions. In addition, the consequences of an artificial agent's action have to be evaluated using causal models or action languages. Lastly, both ethics and legal reasoning deals with obligations, permissions and prohibitions, i.e. with deontic modalities, which may require the use of specific modal logics or of other relevant formalisms. As a consequence, it would be suitable to* ***combine non-monotonic formalisms, causal models*** *and* ***a formalism supporting obligations and permissions*** *to model both ethical and legal reasoning. That is the challenge to which this thesis will have to respond supporting with many existing works (cf. [Tolmeijer and al., 2020]).*

In the past, the ACASA-LIP6 team has used a non-monotonic formalism based on stable models, that is, Answer Set Programming to model ethical reasoning (cf. [Ganascia 2007; 2015]). However, it lacks a causal models and deontic modalities. Some works have made use of deontic logics (e.g. [Arkoudas and al., 2005]) and even of defeasible deontic logics [Horty, 1997], but they hardly manage with non-monotonicity and they do not include causal models. Others have tried to include causal models (cf. [Halpern and al., 2018]) or Action Languages (cf. [Berreby and al. 2018]) but they do not really deal with ethical conflicts nor make use of deontic modalities. This thesis will have to proceed with our preliminary work with Answer Set Programming and with Action Languages, also implemented in the ACASA-LIP6 research team, to try to solve the above-mentioned triple constraint, i.e. evaluating consequences, deal with deontic modalities and overcome conflicts of norms. This work will be done under the supervision of Jean-Gabriel Ganascia and Gauvain Bourgne who are both Professor at Sorbonne University and who work in the ACASA-LIP6 team.

This thesis will take place in the RECOMP research action that aims at ensuring the artificial intelligent agents compliance with legal and ethical norms. The goal of the project is to develop real time mechanism checking compliance and revision of agent behavior. Three groups involved in the project: a German group, headed by Professor Adrian Paschke, in the department of computer science of Freie Universität Berlin, focuses on representation of norms. The Japanese group headed by Professor Ken Satoh in the National Institute of Informatics, will simulate legal reasoning using laws that are viewed as hard constraints, i.e. constraints that cannot be violated. The French partner that is the ACASA-LIP6 group, headed by Professor Jean-Gabriel Ganascia, have to manage ethical rules viewed as soft constraints, i.e. constraints that could be infringed in some specific situation. The candidate will have to ensure the compatibility of the representation he will use with the representations used by the other partners of the project.

The developed models will be implemented and validated on real cases of health technologies and social robotics. We are particularly interested in health technologies, since ethical issues in health are crucial for many reasons, in particular because there are multiple conflicting requirements that are on the one hand protection of privacy and security and on the other hand, efficiency of the decision and responsibility of the physician. Our contact on the one hand with physicians in Sorbonne University (Pitié-Salpêtrière hospital) and on the other hand with an industrial partner, the Berger-Levrault company that is a software and digital solution publisher particularly concerned on the questions of confidence and privacy protection, but also on health technologies and robotics, will provide use-cases on which we shall be able to test the applicability of our models. Lastly, in collaboration with the RECOMP partners, we shall model GDPR — General Data Protection Regulation — with our formalism and automatically check compliance of artificial agents with this GDPR legal norms and legal norms and privacy-related ethical rules.

Required Knowledge and Skills

We require applicants to have training in either symbolic artificial intelligence or mathematical logic. In addition, applicants must show an interest in philosophy and, in particular, in ethical issues. Finally, computer skills are also required to be able to actively contribute to the programming of mechanisms checking compliance of both legal norms and ethical rules.

Context

The researcher will be based in the ACASA-LIP6 team, on the Pierre et Marie Curie Campus (Sorbonne University), in Paris, under the supervision of Professors Jean-Gabriel Ganascia and Gauvain Bourgne. The thesis will be funded by the RECOMP project that is a French-German-Japanese research action. It means that the postulant will have relationships with all the RECOMP partners that are not only French, but also German and Japanese,

and will have to take part to the different meetings and to work on the different RECOMP working packages attributed to the ACASA-LIP6 team.

The thesis will begin the 1st April 2021

References

[Arkoudas and al., 2005] Arkoudas, K., Bringsjord, S. & Bello P. (2005) "Toward Ethical Robots via Mechanized Deontic Logic.", Machine Ethics: Papers from the AAAI Fall Symposium, Tech. Report FS–05–06, pp. 17–23.

[Awad and al., 2018] Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I., The Moral Machine experiment, Nature, 563, 59–64, 2018.

[Berreby and al. 2018] Berreby F., Bourgne G., Ganascia J.-G.: Event-Based and Scenario-Based Causality for Computational Ethics. AAMAS 2018: 147-155

[Berreby and al., 2017] Berreby, F., Bourgne, G., and Ganascia, J. (2017). A declarative modular framework for representing and applying ethical principles. In AAMAS 2017 proceeedings, pages 96–104.

[Berreby and al., 2015] Berreby F., Bourgne G., Ganascia J.-G., "Modelling Moral Reasoning and Ethical Responsibility With Logical Programming", Logic for Programming, Artificial Intelligence, and Reasoning, Volume 9450 of the series Lecture Notes in Computer Science pp 532-548, November 2015

[Ganascia, 2015] Ganascia J.-G., "Non-monotonic Resolution of Conflicts for Ethical Reasoning", in A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, ed. R. Trappl, 2015*

[Ganascia and al. 2018] Ganascia, J-G, Tessier C and Powers T., « On the Autonomy and Threat of "Killer Robots" », APA Newsletter on Philosophy and Computers, vol. 17, n° 2, summer 2018, p. 3-9 (https://c.ymcdn.com/).

[Ganascia, 2007] Ganascia J.-G.: "Modeling Ethical Rules of Lying with Answer Set Programming", Ethics and Information Technology, vol. 9, pp. 39-47 (ISBN : 1388-1957) (2007)276.

[Halpern and al., 2018] Halpern, J. Y., Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (pp. 1853–1860).

[Horty, 1997] Horty J. Defeasible Deontic Logic, chapter Nonmonotonic foundations for deontic logic, pages 17–44. Kluwer Academic Publishers, 1997.

[Lorini and al., 2014] Lorini, E., Longin, D. and Mayor, E. 2014. "A logical analysis of responsibility attribution: emotions, individuals and collectives". Journal of Logic and Computation 24(6):1313– 1339.

[Tolmeijer and al., 2020] Tolmeijer S., Kneer M., Sarasua C., Christen M., Bernstein A., "Implementations in Machine Ethics: A Survey", 2020, preprint arXiv, https://arxiv.org/abs/2001.07573

[Tufis and Ganascia, 2015] Tufis M., Ganascia J.-G., "Grafting Norms onto the BDI Agent Model", in A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, ed. R. Trappl, 2015.*