The Jiminy Advisor: an extended abstract

Beishui Liao[1] , **Pere Pardo**[2] , **Marija Slavkovi**[3] , **Leendert van der Torre**[1,2]
Zheijang University, Hangzhou, China
University of Luxembourg, Luxembourg
University of Bergen, Norway
baiseliao@zju.edu.cn, pere.pardo@uni.lu, marija.slavkovik@uib.no, leon.vandertorre@uni.lu

Abstract

This work considers the assumption that multiple-stakeholders would need to inform the ethical behaviour of an artificial agent. We study how to combine the normative views of these stakeholders, resolving possible dilemmas on different levels, in a transparent and tractable way that lends itself to building explanations of the agent's decisions to the stakeholder.

1 Introduction

Artificial autonomous agents depend on human intervention to distinguish moral from immoral behavior. The multiple persons and institutions that are affected by the moral behavior of an artificial agent should be given the opportunity to indicate their moral requirements for that system's behavior (Baum 2020). Let us assume that people can have multiple roles when interacting with an autonomous system, and that all the stakeholders' moral instructions should be included when deciding the moral behavior of an autonomous system. The problem that we address in the article is: *how should an autonomous system follow the ethical input of various stakeholders?*

For this extended abstract we assume the reader is familiar with formal argumentation theory.

2 Main contributions of the paper

In our approach we formally represent each "morality" stakeholder that participates in the "moral council" of an artificial agent. We propose that normative systems (Chopra et al. 2018) and formal argumentation (Baroni et al. 2018) can be used to implement this "moral council". We call this moral council Jiminy Advisor, as it should play, whimsically, the same role that Jiminy cricket played for Pinocchio in the children's story. The advantage of using normative systems for stakeholder modelling accounts for the possibility that different stakeholders may not use the same moral theory or reasoning approach to inform their points of view.

A norm is a formal description of desirable behaviour, action or outcome of an action. We call detachment the relation between norms and consequent obligations they impose, permission they allow, or institutional fact they establish. We define a moral dilemma as a situation where it is not possible to satisfy all norms, i.e., at least one detached obligation must be violated. A contribution of the paper is to carefully and formally define several different conflicts and inconsistencies that can occur when making moral decisions.

The normative systems representing the stakeholders are used by the Jiminy advisor to calculate what moral decision to recommend to the guided artificial agent. Given a set of normative systems and a moral decision problem, Jiminy formulates an ASPIC-style argumentation system for checking and resolving conflicts. Specifically, arguments are constructed from an argumentation theory that consists of a normative system and a knowledge base of brute facts shared by all the stakeholders.

Arguments are constructed from the given normative systems that are associated with one stakeholder or more stakeholders. Priorities among arguments are also constructed based on the nature of the argument: we discern between institutional facts, obligations and permissions. Given the argumentation theory of each stakeholder, and a decision problem, we distinguish four ways (i.e., four collections of extensions) and check whether there is a dilemma and, if so, whether it can be solved at the next level:

1. We consider the normative system of each stakeholder independently, compute the extensions of the corresponding argumentation frameworks, and check whether there is a dilemma among the extensions of the stakeholders.
2. We consider the arguments of all the stakeholders together to construct a single argumentation framework and check whether it contains dilemmas. Each argument still consists of norms from one single stakeholder.
3. We put all the normative systems together into a unified argumentation theory and check whether it contains dilemmas. Arguments now combine norms from different stakeholders.
4. We use the Jiminy defaults to decide which stakeholders are the most competent for the context and dilemma at hand.[1]

At any of these four levels, if no dilemma is found, the Jiminy submits as its moral recommendation the set of obligations occurring in at least one of the semantic extensions (and as facts or permissions only those that occur in all of the semantic extensions).

[1]The source of the Jiminy's priorities is domain specific. We assume that the set of norms in the Jiminy is given.

How we integrate a Jiminy advisor with an artificial agent depends on what type of moral agent we need to construct, or rather whether the agent itself has any moral reasoning capabilities apart from the Jiminy advisor. Following the work of (Moor 2006), an *implicit ethical agent* does make ethically sensitive decisions or operates in an ethically sensitive context, but, the agent's actions are constrained so that unethical outcomes are avoided. An *explicit ethical agent* also makes ethically sensitive decisions or operates in an ethically sensitive context, but, the explicit ethical agent is able to use its own autonomy and reasoning abilities to distinguish ethical from unethical outcomes and actions.

By coupling a Jiminy component with an agent that has no ethical reasoning abilities, we can create an implicit ethical agent. In such an integration, the Jiminy serves as an "external labeler" of decisions or actions for the purpose of avoiding unethical outcomes. Effectively, the Jiminy acts as an ethical governor.

Explicit ethical agents are able to engage in ethical reasoning and possibly also develop their own moral theories. However, for some agents, it would be important to make sure that certain ethically sensitive situations are not left entirely to the autonomous decision-making capabilities of the agent. This is where a Jiminy can be used in the role of ethical advisor, interfacing not directly with the agent's planner, knowledge base and possibly sensors but with the agent's ethical reasoning engine. Having a Jiminy as an advisor does not change the resulting behavior of the agent in the sense that the agent remains an explicit ethical agent.

The Jiminy advisor can operate in morally sensitive scenarios that are pre-identified, which allows us also to specify ultimate defaults that the Jiminy component can use when there are persistent dilemmas at the fourth level. If the scenarios in which Jiminy can operate why don't we just resolve the issues before the artificial agent deployment and not bother with constructing the Jiminy advisor? Because stakeholders can change over the life-time of an artificial agent. By dynamically using the Jiminy component we have tractability and transparency of which stakeholder supported which decision. We also can remove stakeholders and/or add stakeholders, as well as expand or contract the morally sensitive scenarios of consideration.

The main contributions of this article are the following:

1. Within the field of machine ethics, we describe the first computational model that combines the ethical theories of multiple stakeholders in ethical decision making;
2. Within the field of structured argumentation, we describe the first model that resolves moral dilemmas arising from multiple normative systems.

3 Relevance of the paper to KR

This paper bridges KR to other areas of AI specifically machine ethics. Machine ethics is concerned with the moral behavior of machines towards humans and other machines (Anderson and Anderson 2007). Within computer science this concern involves developing methods for automated moral reasoning that can be used by artificial agents of varying reasoning capacity.

Knowledge engineering is used to specify the normative systems that represent the stakeholders who are informing the moral behaviour of the artificial agent. We are using the notion of argument as defined in the terms of (Pigozzi and van der Torre 2018): we assume that all norms are defeasible, and that all arguments constructed from our normative systems are defeasible. Therefore to calculate the extensions Jiminy needs access to a non-monotonic reasoner, like for example an answer set programming engine.

This paper bridges KR to areas that make use of KR, specifically multi-agent systems and explainable AI. The Jiminy advisor is a reasoning component to an artificial agent that exists in a multi-agent system. Although we do not go into details regarding the explainability abilities of the Jiminy advisor, we discuss them in the paper.

4 Significance of the results

Machine ethics is a nascent AI field. It offers many challenges for computer science, software engineering and philosophy, but also specifically to KR. We need to understand how to best specify the moral preferences, points of view or simply basic requirements of people who interact with devices that exhibit intelligent behaviour. Since one device is always serving many proverbial masters, computationally resolving conflicts among different moral inputs is necessary to advance AI supported technology. Consider for example the current state with social media content moderation: there is more content to evaluate for suitability and harm than people can, or should, be exposed to[2].

We propose one conceptual framework that is explicitly designed to resolve moral related reasoning conflicts. We are among the first to do so, but we hope that there will be many different approaches proposed in the future. It is, at this point, difficult to evaluate the shortcomings of any moral conflict reasoning approach beyond the evident engineering implementation challenges. But we can identify and refine the requirements with each proposed approach.

References

Anderson, M., and Anderson, S. L. 2007. The status of machine ethics: a report from the aaai symposium. *Minds and Machines* 17(1):1–10.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baum, S. D. 2020. Social choice ethics in artificial intelligence. *AI & SOCIETY* 35(1):165–176.

Chopra, A.; van der Torre, L.; Verhagen, H.; and Villata, S. 2018. *Handbook of normative multiagent systems*. College Publications.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Pigozzi, G., and van der Torre, L. 2018. Arguing about constitutive and regulative norms. *Journal of Applied Non-Classical Logics* 28(2-3):189–217.

[2]https://time.com/6247678/openai-chatgpt-kenya-workers/