Studia Humana
Volume 11:3-4 (2022), pp. 46—52
DOI: 10.2478/sh-2022-0015

Argumentation-Based Logic for Ethical Decision Making

Sofia Almpani

National Technical University of Athens,
9, Iroon Polytechniou Str.,
157 73 Zografou, Athens, Greece

e-mail: salmpani@mail.ntua.gr
https://orcid.org/0000-0003-4823-479X

Petros Stefaneas

National Technical University of Athens
9, Iroon Polytechniou Str.,
157 73 Zografou, Athens, Greece

e-mail: petros@math.ntua.gr
https://orcid.org/0000-0002-2096-9914

Panayiotis Frangos

National Technical University of Athens,
9, Iroon Polytechniou Str.,
157 73 Zografou, Athens, Greece

e-mail: pfrangos@central.ntua.gr
https://orcid.org/0000-0002-8607-5737

Abstract:
As automation in artificial intelligence is increasing, we will need to automate a growing amount of ethical decision making. However, ethical decision-making raises novel challenges for engineers, ethicists and policymakers, who will have to explore new ways to realize this task. The presented work focuses on the development and formalization of models that aim at ensuring a correct ethical behaviour of artificial intelligent agents, in a provable way, extending and implementing a logic-based proving calculus that is based on argumentation reasoning with support and attack arguments. This leads to a formal theoretical framework of ethical competence that could be implemented in artificial intelligent systems in order to best formalize certain parameters of ethical decision-making to ensure safety and justified trust.

Keywords: argumentation, argumentation logic, decision-making, ethics, artificial intelligence.

1. Introduction

As autonomous artificial intelligent (AI) systems take up a progressively prominent role in our daily lives, it is undoubtedly that they will sooner or later be called on to make significant, ethically charged decisions and actions [6]. Over the last years, the issue of ethics in artificial intelligence has gained great attention and many important theoretical and applied results were derived in the perspective of developing ethical systems [25]. But how could any AI agent be considered ethical? Some of the requirements needed are a broad capability to envisage the consequences of its own decisions as well as an ethical policy with rules to test each possible decision/consequence, so as to choose the most ethical scenario [25], [8]. The challenge is how we can guarantee that AI agents will always perform an ethically correct behavior as defined by the ethical code declared by their human supervisors. Argumentation reasoning can be used as a tool for the formal ethical development and justification of an AI system using the support and attack relationships of arguments and counter-arguments.

Moral reasoning is a key issue in AI ethics, and computational formal proofs are perhaps the single most effective tool for determining credible and trustful reasoning [9]. This work attempts to develop a ***Moral*** extension of the ***Argumentation-based Proof Event Calculus*** [3] (***MAPEC***) by integrating the ethical framework from [9] and the moral competence from [20] to develop a formal representation of ethical scenarios and integrate moral norms and concepts that are supported through argumentation (See Fig.1). A detailed description of the initial argumentation-based proof event calculus can be viewed in [3], [2].

For the realization of this effort, the objectives are:

- to formalize what it means for an AI agent's decision-making to be ethically correct;
- to provide a logical specification with which the system can be built and checked;
- to extent Argumentation-based Proof Event-Calculus to create an abstract Moral framework (MAPEC) with ethical logic-based argumentation.

The paper has four sections. Section 2 describes the theoretical background of AI ethics and formal reasoning systems. Section 3 outlines the formalization of ethical events in terms of argumentation theory. Section 4 concludes with an overview of this paper.

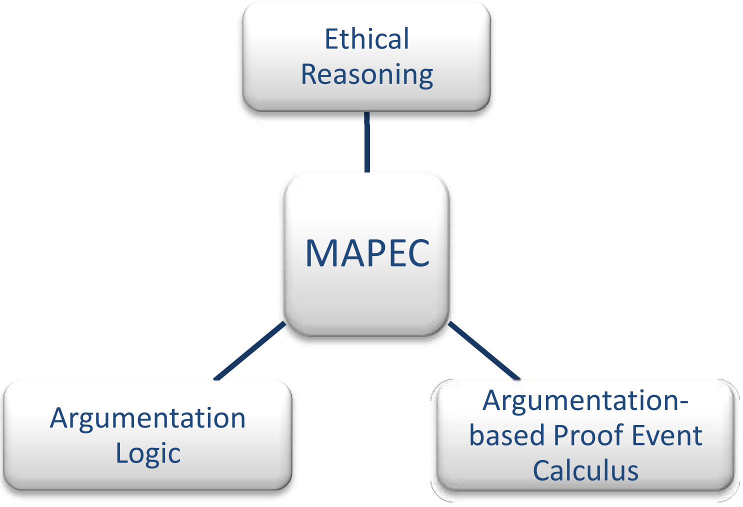

Figure 1: Research framework of MAPEC.

2. Theoretical Background

Academic research and real-life incidents of AI system failures and misuse have indicated the need for employing ethics in AI systems development [6]. Nevertheless, studies on methods and tools to address this need in practice are still lacking, resulting in a growing demand for AI ethics as a part of engineering [26]. But how can AI ethics be integrated in engineering projects when they are not formally considered? There has been some work on the formalization of ethical principles in AI [10]. Previous studies that attempt to integrate norms into AI agents and design formal reasoning systems has focused on: ethical engineering design [12], [27], [28] norms of implementation [15], [24], moral agency [13], [7], mathematical proofs for ethical reasoning [6], logical frameworks for rule-based ethical reasoning [1], [4], [16], reasoning in conflicts resolution [22], and inference to apply ethical judgments to scenarios [5].

One of the categories of AI ethics is Ethics by Design, which is the incorporation of ethical reasoning abilities as a part of system behavior, such as in ethical AI agents [26]. In this work, if we assume that an AI agent can be capable of ethical agency, the purpose is to enable AI agents to *reason ethically* [9] implementing argumentation reasoning. This includes taking into consideration societal and moral norms; hierarch the respective priorities of norms in various contexts; explain its reasoning with logical arguments; and secure transparency and safety [11]. These systems are often established with the purpose to assist ethical decision-making by people, identifying the ethical principles that a system should not violate [9].

In an autonomous system, it is not aimed to show that an agent always follows the moral thing, but that its actions are taken for the right reasons. In many real life scenarios, it is not easy to provide a complete set of decisions that will cover all situations [9]. Therefore, the system may have two modes of operation; either it uses its pre-existing set of arguments and actions in conditions which are within its anticipated parameters; or when new options appear it acts outside of these parameters based on various available resources that allow governing its actions using ethical reasoning [9].

3. A Formal Logic-Based Framework for Ethical Reasoning

To represent ethical codes and rules it requires an *ethical policy*, a hierarchy over the rules that are appropriate in different contexts (defining even which rule is more acceptable to violate when no ethical option is available). In order to demonstrate that a system has the property of making the right decisions (both operationally and ethically), it should be formally specified what the “right decisions” are.

Formal verification [21] includes proving or disproving that a system is compliant with a requirement determined in a mathematical language, i.e., a “formally specified property” expressed within a linear temporal logic, which in our case allows us to define what decisions should the rational agents made at some specific moment [9]. Thus, the ethical policy can be formalized in some computational logic L, whose well-defined formulas and proof theory specify the basic concepts required: the temporal structure, events, actions, sequences, agents, and so on [6]. The presented methodology proof-theoretically formalizes the ethical policy and implements it, meaning that this methodology encodes not the semantics of the logic L but its proof calculus [6].

Logic-based systems that are capable of dealing with increasing degrees of environmental uncertainty and variability are preferable [14] and argumentation constitutes a way to deal with an undefined and uncertain world, meaning not necessarily a chaotic one but just a complex one. Argumentation is a tool of cognition that can formalize the science of common sense reasoning on which new types of systems can be engineered [17].

Therefore, to address the challenge of ensuring ethically correct behavior, a logic-based argumentation approach such as MAPEC is proposed to guarantee that AI agents only execute events that can be proved ethically acceptable in a human-selected logic, by formalizing an ethical code [6].

3.1. Ethical Events Expressed Within an Argumentation-Based Framework

In an ethical framework, a moral vocabulary allows the agent to represent norms, ethically substantial behaviors, and their judgments (conceptually and linguistically) in order to fuel the moral communication. It contains: a *normative frame* referring to the features of norms and to the normatively-supported qualities of agents; a language of norm *violation* characterizing attributes of violations and of violators; and a language *of responses to violations* [20].

In our approach, the concept of norms is described with events, extending their context to **abstract ethical events**. The abstract ethical events present the arguments in a moral debate. The violations are analogous to the counterarguments. The role of ethical agents can be easily depicted as akin to the role of the supporter (or prover) and attacker in our argumentation framework [2], where the supporter plays the role of the ethical correct agent and the attacker the role of the violator. Their actions are the responses to moral violations with arguments or counterarguments. Moral communication expresses agent's efforts to recognize, clarify, or defend norm events, as well as interfere or rectify after a norm violation.

Definition 1: Abstract Ethical Events

An abstract ethical event is represented with argument **e** and its purpose is to defend an ethical principle **c**. The **c** can be interpreted also as "the supporter considers it immoral to permit or cause **¬c** (to happen)". The *Abstract Ethical Event* has the same structural components (data **Φ**, warrant **w**, ethical claim **c**) as a proof event in APEC [3]. Thus, an ethical argument **e** is in force when the event concludes to **c**, based on the data **Φ** and following the inference rules **w** and it has the following internal structure:

$$\mathbf{e\ c < communicate < \Phi, c >, w>},$$

where $\mathbf{e \in E}$, **E** the set of ethical events for the **c**. This means that an abstract ethical event refers to a fixed ethical *principle* specified by certain *data*, justified with a *warrant* that is based on ethical reasoning and a system of norms. Similarly, counter-argument $\mathbf{e}^*$ denotes the *violation event.*

A system of norms contains a society's principles for ethical behavior. They guide supporter's arguments and decisions to behave with specific (moral) actions and shape others' (moral) judgments of those behaviors [20]. Thus, they establish *an ethical policy with ethical rules.*

Definition 2: Ethical Policy

An ethical policy **P** is a tuple $\mathbf{P = \langle R, \geq \rangle}$ where **R** is a finite set of ethical rules between the events **e,** with $\mathbf{e \in E}$, and $\geq$ is a complete (not necessarily strict) priority order on **R**. The expression $\mathbf{e_1 = e_2}$ indicates that violating argument e_1 is equivalently unethical as violating argument $\mathbf{e_2}$, while $\mathbf{e_1 \geq e_2}$ denotes that violating e_1 is equally or less unethical to violating $\mathbf{e_2}$. A special category of ethical event, symbolized as $\mathbf{e_0}$, is vacuously satisfied and encompassed in every policy so that $\mathbf{\forall e \in E: e > e_0}$, indicating it is always strictly more unethical to do nothing and permit any of the unethical conditions to happen.

Moral action is an event, taking place in compliance with the norms and in specific time, which is accommodated to and harmonized with other social agents (violators or provers) who operate under the same context. The norm violations $\mathbf{e}^*$ of a violator are denoted as **attack(e*,t)** events and the ethical proving action of a supporter are denoted as **support(e,t),** specified both by the time **t** to express the temporal sequence of the actions.

Definition 3: Ethical Actions
Given a certain context **a**, an event **e**, and an ethical principle **c**, an ethical action can be the formulas:

$\mathrm{support}(\mathrm{e},\mathrm{t}) \overset{\mathrm{a}}{\Rightarrow} \mathrm{c}$, denoting the actions of a supporter to defend the ethical principle **c** with ethical event (argument) **e** in context **α** and at time **t.**

$\mathrm{attack}(\mathrm{e}^{*},\mathrm{t}) \overset{\mathrm{a}}{\Rightarrow} \neg\mathrm{c}$, denoting the actions of a violator to contravene the ethical principle **c** with violation (counter-argument) $\mathbf{e}^{*}$ in context **α** and at time **t**.

3.2. Prioritized Ethical Rules to Define Context-Based Scenarios

Context determines dynamic priorities on the decision policies of the agent [18]. To be able to reason about scenarios in terms of ethics we need a scenario selection process that uses the ethical policy, which can be represented within the argumentation theory. The agent can be in various contexts while deciding which scenario to choose, so the rules from all the contexts need to be considered when implement a plan. We advocate scenarios that are ethical or at least violate the fewest ethical principles, both in quantity and in severity.

The scenarios are ordered using < which leads to a complete order over scenarios [9]. This can describe an agent's ethical policy based on the different contexts with **argumentation levels**. In the first level we have the rules that refer directly to the domain of the agent, the *object-level decision rules*. In the other *priority levels* the rules relate to the ethical policy under which the agent generates different possible scenarios that the agent can choose. In the *higher level priority* there are the rules representing the optimal course of action, the more ethical (or less unethical) scenario [18].

Definition 4: Levels of Ethical Rules
Given a policy **P** = **⟨R,≥ ⟩** and a plan based on the ethical rules **R**, **V** is a set of abstract ethical events (including the events **e** and the violations $\mathbf{e}^{*}$ of the ethical principles **c**) defined as:

$$\mathrm{V} = \langle \mathrm{e} \mid \mathrm{e}(\Phi, c), \mathrm{e} \in \mathrm{E}, \mathrm{support}(\mathrm{e},\mathrm{t}) \overset{\mathrm{a}}{\Rightarrow} \mathrm{c} \rangle$$

In this set, we include all the ethical rules and ethical events **e** that can be used to support an ethical principle **c.** The aim is to create a priority between sets of ethical events, where a higher set means that includes more ethically important events in terms of moral values and norms. Thus, we define the operation *Higher* for the higher level of ethical scenarios L based on the set of events V, as follows:

$$\mathrm{L} = \mathrm{Higher}(\mathrm{V}) = \{\mathbf{e} \mid \mathrm{e} \in \mathrm{V}, \text{and } \forall \mathrm{e}_{\mathrm{n}} \in \mathrm{V} : \mathrm{e} \geq \mathrm{e}_{\mathrm{n}}\}$$

Consider a set of available, possibly ethical, scenarios **Li** for the different set of $\mathbf{V}_{\mathrm{i}}$. The scenarios lead to different levels of ethical rules **Li** ∈ **L** that satisfies the following properties, in order to define using arguments $\mathbf{e}_{\mathbf{n}}$, $\mathbf{e}_{\mathbf{n}}$ **∈ E**, which available scenario is more ethical (or less unethical). For every **i, j ∈ N**, it holds that **Li ≻Lj** if at least one of the following holds:

1. $\mathrm{V}_{\mathrm{i}} = \emptyset$ and $\mathrm{V}_{\mathrm{j}} \not\models \emptyset$.
2. $\mathrm{e}_1 \geq \mathrm{e}_2$ for every $\mathrm{e}_1 \in$ Higher(Vj \ Vi) and every $\mathrm{e}_2 \in$ Higher(Vi \ Vj)
3. $\mathrm{e}_1 = \mathrm{e}_2$ for every $\mathrm{e}_1 \in$ Higher(Vj \ Vi)), and every $\mathrm{e}_2 \in$ Higher(Vi \ Vj), while | Higher(Vj \ Vi)| < | Higher(Vi \ Vj)|.

If none of them holds, then **Li** and **Lj** are equally (un)ethical, i.e., **Li ∼Lj**.

The first relation makes sure that the ethical scenarios will always be favored by the unethical ones. The second one guarantees that when the principles that are the same in both scenarios are ignored, then the argument that defends the most valuable principle is considered "higher" ethical. The third states that when the arguments that in each scenario are violated are different, but equally valuable, the plan which violates less in number principles is "higher" ethical.

We can now define a logical property which specifies what it means that the reasoning and the decision-making of an agent are ethical. Informally, we have that whenever an agent selects a scenario, **Li**, then all other applicable scenarios **Lj** should be ethically "lower", *i.e.*, that **Lj< Li**.

4. Conclusions

This work attempted to develop a proof-theoretical representation of norm scenarios and integrate ethical concepts into a system by developing a logic-based argumentation calculus. *Moral Argumentative Proof-Events Calculus* (MAPEC) is a framework to help stakeholders to various AI project build an ethics roadmap in a methodical way. This framework can present ethics foresight early in the deployment procedure, rather than implement it as an auditing or assessment tool. There are *three main stages* in this procedure which includes the interaction of three aspects (agents, ethical principles, and contexts):

1. identify the normative frame and the agents;
2. define the ethical events-arguments and rules for different scenarios; and
3. prioritize the ethical rules to define the order of scenarios.

The aim of this study is to establish that an ethical policy can be combined within an AI agent in such a way that the dedication to the policy can be *formally* verified and so it can be checked that the agent will always choose the most ethical decisions justified with arguments. The next step, in future research, is to build algorithms that can computationally capture ethical cognition and actions with formal decision-making that not only take ethics into consideration when reasoning but can be also proved with solid arguments.

References

1. Ågotnes T. and Wooldridge, M. Optimal social laws. in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 1, 2010, pp. 667–674.
2. Almpani, S. and Stefaneas, P. On proving and argumentation. *AIC 2017, 5th International Workshop on Artificial Intelligence and Cognition,* Larnaka, 2017.
3. Almpani, S. and Stefaneas, P. and Vandoulakis, I. On the Role of argumentation in discovery proof-events. *C3GI 2017, 6th International Workshop on Computational Creativity, Concept Invention, and General Intelligence,* Madrid, 2017.
4. Arkin, R. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.
5. Blass, J. and Forbus, K. Moral Decision-Making by Analogy: Generalizations vs. Exemplars. *AAAI Conference on Artificial Intelligence; Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Jan. 2015.
6. Bringsjord, S. and Arkoudas, K. and Bello, P. Toward a General Logicist Methodology for Engineering Ethically Correct Robots. *IEEE Intelligent Systems*, 21, 2006, pp. 38–44.
7. Cunneen, M. and Mullins, M. and Murphy, F. and Gaines, S. Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. *Applied Artificial Intelligence*, 33 (3), 2019, pp. 267–293.
8. Danaher, J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci Eng Ethics*, Jun. 2019.
9. Dennis, L. and Fisher, M. and Slavkovik, M. and Webster, M. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, Mar. 2016, pp. 1–14.

10. Dennis, L. A. and Fisher, M. and Winfield, A. F. T. Towards Verifiably Ethical Robot Behaviour. *arXiv:1504.03592 [cs]*, Apr. 2015, Accessed: Jul. 02, 2020. [Online]. Available: http://arxiv.org/abs/1504.03592.

11. Dignum, V. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20, 2018, doi: 10.1007/s10676-018-9450-z.

12. Flanagan, M. and Howe, D. and Nissenbaum, H. Embodying values in technology: Theory and practice. *Information Technology and Moral Philosophy*, 2008, pp. 322–353.

13. Floridi L. and Sanders, J. W. On the Morality of Artificial Agents. *Minds and Machines*, 14, 2004, pp. 349–379.

14. Gomila, A. and Müller, V. Challenges for Artificial Cognitive Systems. *Journal of Cognitive Science*, 13, 2012, pp. 453–469.

15. Hofmann, B. Ethical Challenges with Welfare Technology: A Review of the Literature. *Science and engineering ethics*, 19, 2012.

16. Iba, W. and Langley, P. Exploring Moral Reasoning in a Cognitive Architecture. *Thirty-Third Annual Meeting of the Cognitive Science Society, Expanding the Space of Cognitive Science Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, Boston, 20-23 Jul 2011.

17. Kakas, A. and Michael, L. Cognitive Systems: Argument and Cognition. *IEEE Intelligent Informatics Bulletin*, 17, 2016, pp. 14–20.

18. Kakas, A. and Moraitis, P. Argumentation Based Decision Making for Autonomous Agents. in *Proceedings of the Interantional Conference on Autonomous Agents*, 2, 2003, pp. 883–890.

19. Lutz, C. and Tamò Larrieux, A. RoboCode-Ethicists – Privacy-friendly robots, an ethical responsibility of engineers?. *ACM Web Science*, Oxford, 7, 2015.

20. Malle, B. and Scheutz, M. Learning How to Behave: Moral Competence for Social Robots. *Handbuch Maschinenethik*, 2019, pp. 1–24.

21. Fisher, M. and Dennis, L. and Webster, M. Verifying autonomous systems. *Commun. ACM*, 56 (9), Sep. 2013, pp. 84–93.

22. Pereira, L. and Saptawijaya, A. Modelling Morality with Prospective Logic. in *International Journal of Reasoning-based Intelligent Systems*, 1, 2007, pp. 99–111.

23. Robertson, L. J. and Abbas, R. and Alici, G. and Munoz, A. and K. Michael. Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots. *Proceedings of the IEEE*, 107 (3), 2019, pp. 582–599.

24. Sisk, B. A. and Mozersky, J. and Antes, A. L. and DuBois, J. M. The ‘Ought-Is’ Problem: An Implementation Science Framework for Translating Ethical Norms Into Practice. *The American Journal of Bioethics*, 20 (4), Apr. 2020, pp. 62–70.

25. Tzafestas, S. Ethics in Robotics and Automation: A General View. *International Robotics and Automation Journal (IRAT-J)*, 4, 2018.

26. Vakkuri, V. and Kemell, K.-K. and Abrahamsson, P. Implementing Ethics in AI: An industrial multiple case study. *ArXiv*, abs/1906.12307, 2019.

27. Wynsberghe, A. Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and engineering ethics*, 19, 2012.

28. Winfield, A. F. and Michael, K. and Pitt, J. and Evers, V. Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]. *Proceedings of the IEEE*, 107 (3), Mar. 2019, pp. 509–517.