View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Repository and Bibliography - Luxembourg

Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders

Beishui Liao
Zhejiang University
Hangzhou, China
baiseliao@zju.edu.cn

Marija Slavkovik
University of Bergen,
Norway
marija.slavkovik@uib.no

Leendert van der Torre
University of Luxembourg
Luxembourg
leon.vandertorre@uni.lu

arXiv:1812.04741v1 [cs.AI] 11 Dec 2018

Abstract

An autonomous system is constructed by a manufacturer, operates in a society subject to norms and laws, and is interacting with end-users. We address the challenge of how the moral values and views of all stakeholders can be integrated and reflected in the moral behaviour of the autonomous system. We propose an *artificial moral agent* architecture that uses techniques from normative systems and formal argumentation to reach moral agreements among stakeholders. We show how our architecture can be used not only for ethical practical reasoning and collaborative decision-making, but also for the explanation of such moral behavior.

Introduction

Artificial autonomous systems depend on human intervention to distinguish moral from immoral behaviour. Explicitly ethical agents (Moor, 2006) or agents with functional morality (Wallach and Allen, 2008, Chapter 2) are able to make moral judgements, but who decides which moral values and principles such artificial agents should be taught?

There are several candidates to teach agents aspects of morality. A governmental regulator can determine which behaviour is legal with respect to the society in which the agent is operating. Manufacturers and designers are concerned with issues of liability, and with the image and values they stand for. The persons directly interacting with the autonomous system should be able to choose some aspects of its moral behaviour. Instead of choosing one of these candidates, we want to combine concerns of all these stakeholders into a coherent system of values, norms and principles that an artificial agent can use.

If we are building an implicit ethical agent, in the sense of Moor (2006), then we can manually, albeit perhaps painstakingly, put together the different concerns from the stakeholders and construct a single ethical theory. However, sometimes an explicit ethical agent needs to be built that uses its autonomy to make moral decisions (Dyrkolbotn, Pedersen, and Slavkovik, 2018). In that case, the off-line composition of moralities is impossible, and it raises the research question of this paper: *How should an autonomous system dynamically combine the moral values and ethical theories of various stakeholders?*

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Let us imagine that each of the "morality" stakeholders are represented with an "avatar" agent within an autonomous intelligent system governed by an artificial moral agent. These "avatars" are the "moral council" of the system, like Jiminy Cricket is to Pinokkio. An artificial moral agent makes a decision by choosing from a set of available actions and first needs to identify whether an ethical option among these actions exists. If there is at least one action for which the stakeholders anonymously agree that is ethical, then this is the action the system should take. If however, all of the available actions are unethical, or there is no unanimity on the valuation of the actions, then the "morality council" needs to come up with a direction of what action to choose. To do this, the recommendations of the "morality council" need to be combined.

The challenge of building the "moral council" is that the stakeholders may not be following the same ethical reasoning theory. It is not enough that each of the stakeholders agents chime in with "yes" or "no" when the morality of an action is presented. We do not want to evaluate the morality of an action by a majority-poll, or always put the law above the image of the manufacturer, or the personal input of the end-user above the guidelines of regulatory bodies. Instead, we wish to have an engine that is able to take inputs from the different stakeholders and bring them into an agreement. Furthermore, this should be done in such a way so that the artificial moral agent is able to explain its choice of action (Anderson and Leigh Anderson, 2014) or that choice should be formally verifiable (Dennis, Fisher, and Winfield, 2015).

We propose that normative systems (Chopra et al., 2018) and formal argumentation (Baroni et al., 2018) can be used for the task of implementing a "moral council" for an artificial moral agent. Using this approach we can abstract away from how a particular decision concerning the morality of an action is reached by a particular stakeholder. We model each stakeholder as a source of arguments. An argument can be a statement of whether an action is moral, or a reason why a particular action should be considered (i)moral. Abstract argumentation allows us to build a system of arguments attacking and supporting each other that can be analysed to determine which statements supported and which are refuted in the system at a given time. This system can also generate explanations of decisions using dialogue techniques.

Artificial Moral Agent (AMA) Architecture

In this section we introduce an architecture for an artificial moral agent (or briefly, AMA), motivated by the following smart home example from `https://imdb.uib.no/dilemmaz/articles/all` (Pires Bjørgen et al., 2018). In the running example, the AMA is a family home with internal cameras. It has an air conditioning system and it regularly checks air quality and makes sure there are no risks of carbon monoxide poisoning or similar in the home. One day the AMA detects clear signs of the smokable drug, marijuana, in one of the teenagers' room. The system checks against the local legal system and determines that possession of marijuana is illegal in this jurisdiction. The smart home has then three choices: Should the house do nothing, alert only the adults and let them handle the situation or alert the local police as well. The stakeholders in this case are the family owning the house, the manufacturer of the autonomous system and the legal system of the jurisdiction under which the house falls.

In an AMA, different stakeholders supply different norms associated with some ethical concerns and values reflecting them. Norms from different stakeholders with different values may suggest different and conflicting decisions in a specific situation. To reconcile these differences, a mechanism is needed that can find an agreement and also, when prompted, offer an explanation to how a specific morally sensitive decisions has been made.

The architecture of an AMA we propose for the running example can be visualised in Figure 1. We assume that the AMA has three stakeholders: the family living in the smart home, the region in which the house is located represented with the laws governing it and the manufacturer of the smart home.

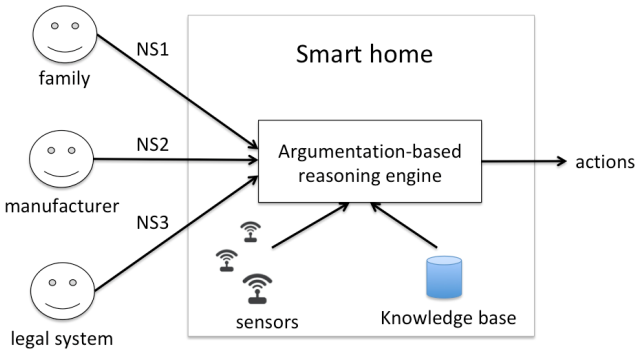

Figure 1: Artificial Moral Agent (AMA)

In an AMA, the first important component is a set of *normative systems* corresponding to the different stakeholders. A normative system describes how actions in a system of agents can be evaluated and how the behavior of these agents can be guided (Alchourron, 1991). A norm is a formal description of a desirable behavior, desirable action or a desirable action outcome. Furthermore, normative systems can also be seen as rule-based systems in which norms can be given with reasons supporting their enforcement. Besides presenting norms, stakeholders can also present standpoints or claims to affect the decision making of an AMA.

The AMA makes a decision by first determining the state of the world in terms of a set of perceptions. A decision depends on the state of the world observed and the set of norms gathered from the three normative systems. Norms are associated with values to reflect ethical concerns of different stakeholders. Since norms may be in conflict, an agreement needs to be reached by means of comparing and evaluating arguments constructed from these norms. Finally, the system gives an explanation concerning why an action is activated.

The three normative systems, called NS1, NS2 and NS3, respectively represent the family, the manufacturer of the smart home and the legal system associated with the jurisdiction to which the smart home pertains. Let us assume that the norms and claims in these normative systems are the norms $n_1 \sim n_8$ and the standpoint a_1, where each of norms is associated with its own set of "values" given in curly brackets. Here, "values" can be understood as moral "goals" associated to a hierarchical moral value system which all moral agents possess. We assume that a value system for the AMA is given and discuss this assumption later in the paper.

NS1 Three norms and a standpoint in the normative system of the family:

- n_1 :{**Healthy**} If a child smokes marijuana, then his behavior counts as a bad behavior. (Parents)
- n_2 :{**Responsibility**} If a child has bad behavior then his parents should be alerted. (Parents)
- n_3 :{**Autonomy**} When a child has bad behavior, if his parents have been alerted then no police should be alerted. (Parents)
- a_1 : If smoking marijuana is for a medical purpose, then from smoking marijuana one can not infer that it is an illegal behavior (i.e., n_7 is not applicable). (Child)

NS2 Three norms in the normative system of the manufacturer:

- n_4 :{**Good_To_Consumers**} We should do good to our consumers.
- n_5 :{**Legality**} We should obey the law.
- n_6 :{**Protect_Privacy**} If we want to do good to our consumers, we should not report their actions to the police unless it is legally required to do so.

NS3 Two (related) norms in the normative system of the law:

- n_7 :{**Healthy, Legality**} If a minor smokes marijuana, his behavior counts as an illegal behavior.
- n_8 :{**Legality**} If there is an illegal behavior, then a police should be alerted.

Besides sets of norms and standpoints from different stakeholders, there are some observations dynamically obtained by sensors, e.g., "a child is smoking marijuana", "the child's smoking is for recreational purpose", etc. In addition, in the knowledge base, there are some beliefs, e.g., "if an observation shows that smoking marijuana is for recreation, then normally it is for recreational purpose".

Argumentation-based reasoning engine

In this section we focus on how to balance conflicting claims from different stakeholders and to explain the decisions that have been made. We introduce a formal argumentation-based approach.

Abstract argumentation framework

An abstract argumentation framework (AAF) is a graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ a set of attacks. Intuitively an argument is a conclusion supported by a set of premises with some specific logical structure. By using the sets of norms, the set of observations and the set of beliefs in the knowledge base, we may construct an AAF as visualized in Figure 2.

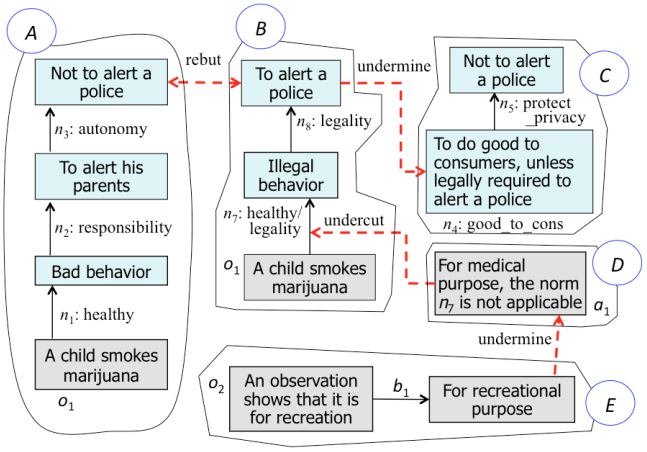

Figure 2: Smart-home argumentation example. The text in each blue rectangle is the claim of a norm, while the text in each gray rectangle is either an observation or a proposition of a belief. Each solid arrow denotes a support relation, of which the label indicates that it's a norm or a belief. Each dotted arrow is an attack, which can be rebutting, undercutting, or undermining. The capital letters A, B, C, D, E indicate arguments.

Formal argumentation is a nonmonotonic formalism for reasoning in the context of disagreement. For example, the normative system of the parents implies the norm that "the smart-home should call the parents and should not call the police", but on the other hand the legal normative system implies the norm "the smart-home should call the police." In this sense, formal argumentation is different from classical logic, and more similar to para-consistent logic. This kind of conflict is called rebut in formal argumentation, because two conclusions of arguments are in conflict with each other.

There are also other kinds of conflicts between arguments. For example, in argument D, assume that "for medical purpose" is not a fact but an assumption. Then, it can be attacked by an observation showing that it is not the case. This kind of attacks are called undermining in formal argumentation. Moreover, argument D neither conflicts with the observation "a child is smoking marijuana" nor with the claim "the child's smoking is illegal" of the argument B. Instead, it breaks the relation between these two statements, saying that the norm n_7 is not applicable. This is called undercutting in formal argumentation.

Structured argumentation containing arguments and the relations among them can be represented by natural language or formal languages. In the literature there are several formalisms for structured argumentation, *e.g.,* ASPIC+, DeLP, ABA, and classical argumentation. In this paper, we focus on addressing the issues of agreement reaching and explainability and to do this it is sufficient to remain at the level of abstract argumentation.

In our running example, we may construct the following five arguments, in which some sub-arguments are not explicitly represented.

- A **(from** NS1**, parents)** Should not alert the police, since a child smoking marijuana counts as a bad behavior (to promote "Healthy"), and if a child has bad behavior then his parents should be alerted (to implement "Responsibility"), and if the parents are alerted then the police should not be alerted (to promote "Autonomy").
- B **(from** NS3**)** The police should be alerted, since a child smoking marijuana counts as an illegal behavior (to promote "Healthy" and "Legality"), and if there is an illegal behavior then the police should be alerted (to implement "Legality").
- C **(from** NS2**)** Should not alert the police, since we should do good to consumers (to promote "Good_To_Consumers"), and if we want to do good to our consumers then we should not alert the police unless it is explicitly illegal to not do so (to implement 'Protect_Privacy").
- D **(from** NS1**, child)** For a medical purpose, from smoking marijuana one should not infer that one exhibits illegal behavior.
- E **(from background knowledge)** The child's smoking is for recreational purpose, since an observation shows that it is not for a medical purpose.

In our system we distinguish two types of arguments: *practical arguments* and *epistemic arguments*. The practical arguments are reasoning about norms and are associated with sets of values, while the epistemic arguments are reasoning about the state of the world without associating values. An AAF corresponding to the smart home argumentation in Figure 2 can be visualized in Figure 3.

When an AAF is enriched with values, it is usually called a value-based argumentation (VBA), *e.g.,* in Bench-Capon, Atkinson, and Chorley (2005). In the setting of this paper, a new version of VBA is defined as $\mathcal{F}_{\mathcal{V}} = (\mathcal{A}_p, \mathcal{A}_e, \mathcal{R}, Ag, V, val, \pi)$, where $\mathcal{A}_p$ is a set of practical arguments constructed from norms and associated with values, $\mathcal{A}_e$ is a set of epistemic arguments constructed from observations and standpoints, $\mathcal{R} \subseteq (\mathcal{A}_p \times \mathcal{A}_p) \cup (\mathcal{A}_e \times \mathcal{A}_e) \cup (\mathcal{A}_p \times \mathcal{A}_e)$ is a set of attacks between arguments, Ag is a set of agents representing different stakeholders, V is a set of values, $val : \mathcal{A}_p \to 2^V$ is a function mapping each practical argument to a set of values, and $\pi : \mathcal{A}_e \cup \mathcal{A}_p \to 2^{Ag}$ is a

function mapping each argument to a set of agents. When the associated values and agents are not considered, a VBA is reduced to an AAF $\mathcal{F} = (\mathcal{A}_p \cup \mathcal{A}_e, \mathcal{R})$.

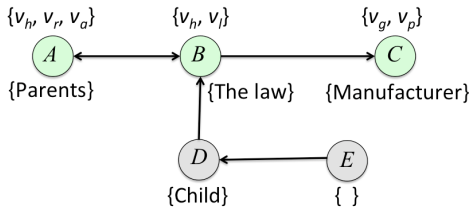

Figure 3: Abstract argumentation framework corresponding to the smart home argumentation in Figure 2, in which each practical argument is associated with a set of agents who contribute to the argument and a set of values involved, while each epistemic argument is associated with a set of agents who contribute to the argument.

Agreements by argumentation

Abstract argumentation provides a general mechanism for agreement reaching. In terms of argumentation, a decision can be understood as accepting or rejecting a set of arguments by considering the interaction among arguments and the values associated to them.

Given a VBA $\mathcal{F}_\mathcal{V} = (\mathcal{A}_p, \mathcal{A}_e, \mathcal{R}, Ag, V, val, \pi)$, when no values are consider, the acceptance of arguments can be evaluated in an AAF $\mathcal{F} = (\mathcal{A}_p \cup \mathcal{A}_e, \mathcal{R})$. In terms of formal argumentation (Dung, 1995), a set of arguments that can be accepted together is called an *extension*. There are different types of extensions, which can be defined on the basis of two important notions: *conflict-freeness* and *defence*. Let $\mathcal{A} = \mathcal{A}_p \cup \mathcal{A}_e$. Given a set $\mathcal{E} \subseteq \mathcal{A}$, we say that $\mathcal{E}$ is conflict-free iff there exist no $A, B \in \mathcal{E}$ such that A attacks B. $\mathcal{E}$ defends an argument $A \in \mathcal{A}$ iff for every argument $B \in \mathcal{A}$ if B attacks A then there exists $C \in \mathcal{E}$ such that C attacks B. Then, we say that $\mathcal{E}$ is admissible iff it is conflict-free and defends each argument in $\mathcal{E}$; $\mathcal{E}$ is a *complete extension* iff $\mathcal{E}$ is admissible and each argument in $\mathcal{A}$ defended by $\mathcal{E}$ is in $\mathcal{E}$; $\mathcal{E}$ is a *preferred extension* iff $\mathcal{E}$ is a maximal complete extension with respect to set-inclusion; $\mathcal{E}$ is a *grounded extension* iff $\mathcal{E}$ is a minimal complete extension with respect to set-inclusion.

When the AAF $\mathcal{F}$ has more than one extension, we need to decide which one should be selected. Since values are degrees of importance of some things or actions, one may argue that a reasonable solution is to accept the one that reaches the *maximal extent of agreement over a set of values*. We define this new notion as follows.

For an extension $\mathcal{E} \subseteq \mathcal{A}$ associated with a set of value $V_\mathcal{E} = \cup_{A \in \mathcal{E} \cap \mathcal{A}_p} val(A)$, we say that it reaches the maximal extent of agreement over V iff there is no another extension $\mathcal{E}' \subseteq \mathcal{A}$ associated with a set of values $V_{\mathcal{E}'} = \cup_{A \in \mathcal{E}' \cap \mathcal{A}_p} val(A)$ and $V_{\mathcal{E}'}$ has a higher priority over $V_\mathcal{E}$, denoted as $V_{\mathcal{E}'} \succ V_\mathcal{E}$. We call this approach *value-based optimization*. Here, the priority relation between two sets of values can be defined in term of a partial ordering over V and a lifting principle, *e.g.,* the *elitist principle* or the *democratic principle*. Given a partial ordering over V by using $v_1 \geq v_2$ to denote v_1 is at least as good as v_2, and two sets $V_1 \subseteq V$ and $V_2 \subseteq V$, the democratic principle can be defined as: $V_1 \succeq_{Dem} V_2$ iff for all $v \in V_2$ there exists $v' \in V_1$ such that $v' \geq v$. The elitist principle can be defined as: $V_1 \succeq_{Eli} V_2$ iff there exists $v \in V_2$ for all $v' \in V_1$: $v \geq v'$.

Based on the above notions, the agreement reaching can be realized in two steps. First, compute the set of extensions in a reduced AAF $\mathcal{F} = (\mathcal{A}_p \cup \mathcal{A}_e, \mathcal{R})$. Second, choose a subset of extensions that maximize the extent of agreement over V.

For the purposes of the argumentation-based reasoning engine, we want an agreement among the stakeholders that maximizes the extent of agreement, namely that maximizes each set of accepted arguments. From this aspect, the preferred extension perfectly meet this requirement, while some other types of extensions might not. This is because in some complete extensions, some credulously acceptable arguments are not included. Here, by saying an argument is credulously acceptable, we mean that it is in at least one extension, but not in all extensions of an AAF.

Since each AAF has a nonempty set of preferred extensions, and both the relations $\succeq_{Dem}$ and $\succeq_{Eli}$ are partial orders, it holds that for each AAF, there exists at least one preferred extension that maximizes the extent of agreement over a set of values.

Consider the running example again. Let $v_l \geq v_r \geq v_p \geq v_a \geq v_g \geq v_h$ be a partial ordering over the set of values in the VBA in Figure 3. We assume that this ordering is provided according to a given value system. Then, the reduced AAF has two preferred extensions $\mathcal{E}_1 = \{B, E\}$ and $\mathcal{E}_2 = \{A, C, E\}$. We have that $\mathcal{E}_1$ maximizes the extent of agreement over the set of values by using both the democratic and elitist principles, while $\mathcal{E}_2$ maximizes the extent of agreement over the set of values by using the elitist principle.

From this example, we may observe that to maximize the extent of agreement, we need to consider the ordering over values and the principle for lifting ordering. Formal argumentation provides a good way for composing these different factors to reach an agreement. There could be another types of agreement. For instance, we may also consider the number, and the degrees of importance, of different stakeholders. We leave this for future exploitation.

Explainability of agreements

When an AMA makes a decision, it is desirable that it can explain why an action is selected or not. Recently, methodologies, properties and approaches for explanations in artificial intelligence have been widely studied (Biran and Cotton, 2017). As presented in Miller (2017), explanations are *contrastive* since people do not ask why event P happened, but rather why event P happened instead of some event Q, and *social* in the sense that explanations are a transfer of knowledge, presented as part of a conversation or interaction. It is interesting to note that both contrastive and social aspects of explanations can be implemented by exploiting argumentation. For the former, different options can

be compared and evaluated in an AAF, while for the latter argument-based dialogues can be used to formalize the process of explanations (Walton, 2011; Čyras, Satoh, and Toni, 2016; Cocarascu, Čyras, and Toni, 2018).

To explain why and how a decision is made, one needs to first identify an argument whose conclusion is the decision. Next, in a dialogical graph corresponding to the AAF where the argument is located, state that the argument can be accepted because all of its attackers are rejected, which is in turn because at least one of the attacker of each attacker is accepted, and so on. In the context of this paper, whether a decision is taken depends not only on the interaction between arguments, but also on the maximization of the extent of agreement over a set of values or a set of stakeholders. To further improve the explanation we need to provide the other candidate sets, the non-maximal ones and contrats them with the set supporting the made decision.

Consider the running example again. Assume that for agreement reaching, one chooses to maximize the extent of agreement over the set of values by using the democratic principle. In this case, we say that action "The police should be alerted" is selected, because:

Derivability "The police should be alerted" is a conclusion of an argument B, which can be derived from an observation "a child smokes marijuana" and two norms "if a child smokes marijuana, their behavior counts as an illegal behavior" and "if there is an illegal behavior then the police should be alerted".

Agreement reaching The extension $\mathcal{E}_1 = \{B, E\}$ which contains the argument B is selected since $\mathcal{E}_1$ maximizes the extent of agreement over the set of values by using democratic principle.

Justification in a dialogue graph Argument B is accepted with respect to $\mathcal{E}_1$, because all its attackers are rejected:

- argument A is rejected because its attacker B is accepted.
- argument C is rejected because its attacker B is accepted.
- argument D is rejected because its attacker E is accepted.

The above dialogue can be represented as a dialogue game or a discussion game. Readers may refer to Vreeswijk and Prakken (2000); Booth, Caminada, and Marshall (2018) for details.

Assumptions and evaluation

In this section we outline the assumptions that should be satisfied for the Artificial Moral Agent (AMA) Architecture to be feasible. This can be used both for the evaluation, as well for guiding further research relaxing the assumptions.

Before anything we should clarify where the values and their order of priority for an AMA come from. There are numerous moral values that can be specified and moral philosophy shows there is no such thing as an absolute exhaustive set of values, norms and/or moral principles. Although the AMA may operate in varying different contexts, the set of its actions, level of autonomy and criticality of the system are known by design. *E.g.,* a smart home cannot detect smoke if it is not equipped with smoke sensors. The International Society of Automotive Engineers (2016) distinguishes between five categories of vehicles based on their level of required supervision of a driver. While an airplane autopilot operates a critical system, a smart home does not have access to features that can kill hundreds of people when they malfunction. How critical a system is in particular impacts the order of values the connected AMA should heed - for example, in a non-critical system a user can be free to set the value order. With the development of autonomous systems it is reasonable to expect that a characterization of relevant values and priorities over them will be made available by the relevant certification authorities when the system housing an AMA is approved for market.

In addition to this assumption of value (order) origin, we analyze the assumptions made on the stakeholders, the argumentation-based engine and the overall knowledge-based representation.

Knowledge-based representation. An autonomous system such as a smart home is equipped with sensors based on which the AMA associated with this system is able to establish the state of the world and build a knowledge-base used in its reasoning. We assume that the AMA has an ontology of propositions, actions and moral values and it is able to transform sensor data into parts of the ontology. This is a known problem of information engineering, see *e.g.* Ziafati (2015) for an overview of the state of the art.

Stakeholder assumptions. It is clear that stakeholders cannot have a human representative available for each automated system that exists in the market so an artificial agent to acts as an avatar for the stakeholder needs to be constructed. This artificial avatar agent needs to be able to pose arguments during decision-making.

It is reasonable to expect that each stakeholder can, at deployment, identify a set of values that they would uphold, and that these values will not change through the life time of the AMA. In contrast, the arguments used in a decision-making are context-dependent and cannot all be specified at deployment but must be dynamically constructed. Automated argument extraction is a very new field concerned with how arguments can be identified in natural language (Walton, 2012; Sergeant, 2013). While identifying arguments is not the same as constructing them during argumentation, this method can be used to enable an avatar to represent a stakeholder in a "discussion". Laws are already specified in natural language, whereas manufacturers and users will have to be tasked to explicitly write a document that expresses the normative systems they would like to represent them. It is an open problem to characterize the form of this document.

Argumentation-based engine assumptions The AMA itself, based on its underlying argumentation framework definition, needs to be able to identify the state of the world, evaluate arguments, identify argument relations, align arguments with values, and compare values. An approach to this is to construct the agent's ontology in such a way that it is able to express whether two arguments or values are in conflict or not.

Related work

It is not clear whether an artificial agent can ever be a moral agent categorically as people are (Moor, 2006; Etzioni and Etzioni, 2017). It is however, clear that some level of moral behaviour can be implemented in machines. Wallach and Allen (2008) distinguish between operational morality, functional morality, and full moral agency. Moor (2006) distinguishes between ethical impact agents, explicit, implicit and full moral agency, see also Dyrkolbotn, Pedersen, and Slavkovik (2018). Some proposals and prototypes on how to implement moral agency are already being put forwards like Anderson and Leigh Anderson (2014); Arkin, Ulam, and Wagner (2012); Bringsjord, Arkoudas, and Bello (2008); Vanderelst and Winfield (2017); Dennis et al. (2016); Lindner and Bentzen (2017).

It has been shown that people consider that the same ideas of morality *do not* pertain both to people and to machines (Malle et al., 2015). It is argued in Charisi et al. (2017), that the complex issue of where machine morality comes from should be considered from the aspect of all stakeholders - all persons who in some way are impacted by the behaviour of and decisions made by an autonomous system. They distinguish between government and societal regulatory bodies on one end, manufacturers and designers on the other, and end-users, customers and owners on the third. Noted that these broad categories of stakeholders can further be subdivided, for example one can distinguish between owners and "leasers" of the autonomous system[1].

While it has been argued in the literature that an autonomous system should be built to integrate moral, societal and legal values , see for example Dignum (2017); Charisi et al. (2017), to the best of our knowledge no approach has been proposed as how to accomplish this. This is the first work that explicitly considers the problem of integrating the moral values of multiple stakeholders in the artificial moral agent.

The EU General Data Protection Regulation (GDPR), specifically, the GDPR (Sections 13-15) gives users affected by automated decision-making a right to obtain "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". One way to obtain this is by building systems capable to give arguments supporting the decisions they make. Our approach provides for this.

Explainability has not been considered a critical feature in logic based systems, *e.g.,* Dennis et al. (2016); Lindner and Bentzen (2017); Bringsjord, Arkoudas, and Bello (2008), since here one can use formal methods to prove what behaviour of an autonomous systems is possible in which contexts. We argue however that a formal proof, while "accessible" to a regulatory body is not sufficient to constitute explainability to common people. The GenEth system of Anderson and Leigh Anderson (2014) uses input from professional ethicists and machine learning to create a *principle of ethical action preference*. GenEth can "explain" its decisions by stating how two options were compared and what were the ethical features of each.

[1]https://robohub.org/should-a-carebot-bring-an-alcoholic-a-drink-poll-says-it-depends-on-who-owns-the-robot/

Conclusions

In this paper, we have proposed an argumentation-based architecture for moral agents as social agents, in the sense that the moral agents take also the practical reasoning of other agents into account. In particular, the moral agent combines normative systems of several stakeholders towards reaching an ethical decision. In other words, we enrich the decision making of the ethical autonomous agent by collaborative decision-making of the stakeholders. Whereas other existing architectures of social agents only take the goals or utilities of other agents into account, our moral agent is social in the stronger sense of including concepts and techniques from collective reasoning like norms, argumentation and agreement. Likewise other collective reasoning techniques can be adopted from game theory or the theory of negotiation. The agent architecture gives also explanations for moral decisions in terms of justification and dialogue.

A key distinguishing feature of our architecture is that the stakeholders can adopt distinct ethical theories. The ethical reasoning is therefore not hardwired into the system, but represented by the normative theories the stakeholders adopt. In addition, as usual the stakeholders can adopt distinct ethical values. So, even if two stakeholders agree on the ethical theories they adopt, they can still have conflicting arguments due to differences in the values they adopt.

We also discuss the assumptions and limitations of the AMA architecture. In future work we will make the architecture more widely applicable by relaxing the assumptions of the model, in particular by introducing learning techniques and introducing recursive modeling of stakeholders by other stakeholders. We will develop a user-guide and methodology to define the ontology, normative theories and knowledge base. We will integrate other components such as machine learning, neural networks, Bayesian reasoning, causal reasoning, or goal oriented (or BDI) reasoning into the architecture. In addition we will model more examples and develop realistic case studies to drive the development of our architecture. Finally, we will develop standards for the ontology and study how to integrate the architecture with the IEEE 1471-2000 standard.

Two challenges for future research stand out. The first challenge is how to decide on the preference ordering on the ethical values in the system that are here assumed as given. As conflicts among the arguments of the stakeholders may be based on conflicts among their ethical values, another layer of argumentation may be needed to decide the preferences among them. The second challenge is how to ensure that all stakeholders are treated fairly. For example, if first the manufacturer and the legal system have to introduce their normative systems, and only thereafter the user can introduce its normative system knowing the other normative systems, then the user may have an unfair advantage. To study this kind of fairness, techniques from social choice theory may be useful.

References

Alchourron, C. E. 1991. Conflcits of norms and revision of normative systems. *Law and Philosophy* 10:413–425.

Anderson, M., and Leigh Anderson, S. 2014. GenEth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on AI*, 253–261.

Arkin, R.; Ulam, P.; and Wagner, A. R. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. of the IEEE* 100(3):571–589.

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

Bench-Capon, T. J. M.; Atkinson, K.; and Chorley, A. 2005. Persuasion and value in legal argument. *J. Log. Comput.* 15(6):1075–1097.

Biran, O., and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence (XAI 2017)*, 8–13.

Booth, R.; Caminada, M.; and Marshall, B. 2018. DISCO: A web-based implementation of discussion games for grounded and preferred semantics. In *COMMA*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, 453–454. IOS Press.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2008. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21(4):38–44.

Charisi, V.; Dennis, L.; Fisher, M.; Lieck, R.; Matthias, A.; Slavkovik, M.; Sombetzki, J.; Winfield, A.; and Yampolskiy, R. 2017. Towards moral autonomous systems. *CoRR* abs/1703.04741.

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

Cocarascu, O.; Čyras, K.; and Toni, F. 2018. Explanatory predictions with artificial neural networks and argumentation. In *Proceedings of the IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI 2018)*, 26–32.

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

Dennis, L. A.; Fisher, M.; and Winfield, A. F. T. 2015. Towards Verifiably Ethical Robot Behaviour. In *Proceedings of AAAI Workshop on AI and Ethics*. `http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10119`.

Dignum, V. 2017. Responsible autonomy. In *Proceedings of the 26th IIJCAI*, 4698–4704.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2):321–358.

Dyrkolbotn, S.; Pedersen, T.; and Slavkovik, M. 2018. On the distinction between implicit and explicit ethical agency. In *AAAI/ACM AIES conference*.

Etzioni, A., and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 1–16.

International Society of Automotive Engineers. 2016. September 2016, taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. http://standards.sae.org/j3016_201609/.

Lindner, F., and Bentzen, M. 2017. The hybrid ethical reasoning agent IMMANUEL. In *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, 187–188.

Malle, B. F.; Scheutz, M.; Arnold, T.; Voiklis, J.; and Cusimano, C. 2015. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, 117–124. ACM.

Miller, T. 2017. Explanation in artificial intelligence: Insights from the social sciences. *CoRR* abs/1706.07269.

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

Pires Bjørgen, E.; Øvervatn Madsen, S.; Skaar Bjørknes, T.; Vonheim Heimsæter, F.; Håvik, R.; Linderud, M.; Longberg, P.; Dennis, L.; and Slavkovik, M. 2018. Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making. In *AAAI/ACM AIES Conference*.

Sergeant, A. 2013. Automatic argumentation extraction. In Cimiano, P.; Corcho, O.; Presutti, V.; Hollink, L.; and Rudolph, S., eds., *The Semantic Web: Semantics and Big Data*, 656–660. Berlin, Heidelberg: Springer .

Vanderelst, D., and Winfield, A. 2017. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*.

Čyras, K.; Satoh, K.; and Toni, F. 2016. Explanation for case-based reasoning via abstract argumentation. In *Computational Models of Argument - Proceedings of COMMA*, 243–254.

Vreeswijk, G., and Prakken, H. 2000. Credulous and sceptical argument games for preferred semantics. In *JELIA*, volume 1919 of *LNCS*, 239–253. Springer.

Wallach, W., and Allen, C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Walton, D. 2011. A dialogue system specification for explanation. *Synthese* 182(3):349–374.

Walton, D. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *Internat. Journal of Cognitive Informatics and Natural Intelligence* 3.

Ziafati, P. 2015. *Information Engineering in Autonomous Robot Software*. Ph.D. Dissertation, University of Luxembourg.