Dialogues on Moral Theories

Guido Governatori, Francesco Olivieri, Régis Riveret
Data61, CSIRO, Australia

Antonino Rotolo
University of Bologna, Italy

Serena Villata
Université Côte d'Azur, I3S, Inria, CNRS, France

Abstract

Most ethical systems define how the individuals ought morally act, being part of a society. The process of elicitation of a moral theory governing the agents in a society requires them to express their own norms with the aim to find a moral theory on which all may agree upon. We address this issue by proposing a formal framework that can instantiate in agents' dialogues moral/rational criteria, such as the *maximin principle*, *Pareto efficiency*, and *impartiality*, which were used, e.g., by John Rawls' theory or rule utilitarianism.

1 Introduction

Many conceptions of autonomy have been developed in social science and philosophy [21, 36]. One successful approach in Artificial Intelligence (AI) and Multi-Agent Systems (MAS) sees an autonomous agent as "self-contained, reactive, proactive [...], typically with a central focus of control, that is able to communicate with other agents [...]. A more specific usage is to mean a computer system that is either conceptualised or implemented in term of concepts more usually applied to humans (such as beliefs, desires and intentions)". Autonomy, in particular, is "the assumption that, although we generally intend agents to act on our behalf, they nevertheless act without direct human or other intervention, and have some kind of control over their internal state" [41].

The authors are supported by the EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 for the project MIREL: *MIning and REasoning with Legal texts*.

A theoretical contribution to the idea of autonomy comes of course from moral philosophy, which takes this idea as crucial: the moral life of agents, their values, norms, and the related concept of moral responsibility are all meaningless without assuming that agents can deliberate and are decision-makers. Specific moral traditions, such as the Kantian one, claim in addition that precisely the idea of autonomy—conceived of as agent's capability of adopting right and universal norms of behaviour—pertains to the domain of *morality*, in contrast with other domains of practical reasoning, such as the *law*, where the grounds of agency is *heteronomy* and the binding force of norms is contextual and depends on external factors—typically, coercion [19].

The role of norms to characterise moral autonomy, and autonomous actions, has been differently, but successfully proposed by (competing) moral views like Kantianism and rule utilitarianism. Drawing from these traditions, we formally explore the following intuition:

Intuition. *Autonomous agents take decisions about the moral theory governing their society, and elicit new theories that would improve the welfare. Decision-making is performed through a collective deliberative procedure, called* moral dialogue.

On the basis of this intuition, a question arises: *how to define a generic formal framework to mimic the determination of moral theories in a society of autonomous individuals?* The research question can break down into subquestions, for example: *(i)* how to represent rational criteria which are the basis of several moral theories? *(ii)* how to represent building blocks for grasping at least basic aspects of both Rawls' approach to morality and rule utilitarianism?

To address these open challenges, we propose a generic framework capturing the determination of moral theories as a dialogical process. In our framework, moral views are represented through the form of rule-based theories expressing agents' norms, which are associated with an utility function. In other words, theories are nothing but normative systems on which agents argue in regard to their moral justification: agents have thus to deliberate about which theories should regulate the society through a dialogue. By putting on the table their arguments in the dialogue, agents determine what are the (possibly new) rules that should govern the society and its welfare, depending on a specific moral theory. We considered building blocks and rational criteria for characterising influential moral theories such as

- Rawls' contractualist model of deontological morality [32]; this approach sees moral principles for a society as resulting from a collective deliberative process among the agents belonging to such a society, a process which, if it runs under impartiality, supports the maximin principle; and
- rule utilitarianism [6, 16, 17]; this approach requires agents to follow the rules[1] that

[1] Hereafter, the terms rule and norm will be interchangeably used.

maximise utility, and which often connects morality to the theory of rational action, where expected-utility maximisation and Pareto efficiency are fundamental concepts [16].

Such moral theories express opposite views [16], and our approach aims at accommodating these different views (possibly amongst others). To sum up, our contribution is as follows:

- Agents propose in a dialogue the normative (moral) theory that they would prefer for their society;
- Each theory is associated with an utility that measures the impact of the proposed norms; the intended reading could be, for example, in terms of the consequence for the society if all agents would conform to such norms (as suggested by rule utilitarianism);
- Agents deliberate in a different way depending on which of the above theories are employed to compute the utility, leading then to the emergence of a society regulated by the selected moral theory.

Drawing ideas from Rawls' theory and rule utilitarianism, we provide a formal framework for moral dialogues which can accommodate the following rational/moral criteria:

- Welfare maximisation [16] and Pareto efficiency [13];
- Maximin principle [38, 32];
- Impartiality [18, 6].

Ethics and moral theories are becoming more and more relevant for AI in general [39, 26, 37, 28, 10, 12, 8], and for autonomous systems in particular [3], as evidenced by initiatives like The Moral Machine by the MIT[2] and the REINS Project [7]. Enhancements in robotics, knowledge representation and reasoning, and cognitive modelling cast a new light on these challenges making their discussion more urgent than in the past. The generic formal approach we propose is, up to our knowledge, the only formal framework devoted to represent and understand the determination of (well known) moral theories in a society of autonomous individuals. Our framework is in line with Artificial Moral Agents (AMA) proposed by Dignum [12]: our intelligent systems incorporate moral reasoning in their deliberation process, and they rely on argumentation theory to explain their behaviour in terms of moral concepts. From a broader perspective, our approach provides a first step towards

[2] http://moralmachine.mit.edu/

the so called “beneficial AI” [34], so that agents are designed to stick to a specific moral theory, e.g., Rawls’ maximin principle, emerging from a dialogue-based deliberation process and having as a goal the societal welfare.

The reader may argue that often people speak of *discovering* moral rules, rather than deliberating on moral rules. In this case, the appropriate dialogue would be like a mathematical dialogue, where people are engaged in the process of discovering independent truths, beyond agreeing on the rules to live by. The issue of discovering moral rules is out of the scope of this paper. We choose a more contractualist approach to deliberate on the *best* set of moral rules depending on the utility of the agents involved in the deliberation process.

The paper is organised as follows: first, we provide the definition of moral theory in our formal setting, and then we sketch how the utility of theories could be determined. After some problem definitions, we propose moral dialogues to address such problems, and we study some basic properties of these dialogues. We conclude with a comparison with the literature, and some future perspectives.

2 Moral Theory Setting

A moral theory defines norms stating what to do; it deals also with forming judgements about what one ought (e.g., morally) to do. We assume a logic language from which it is possible to build moral theories. A moral theory is made of a set rules and a superiority relation over the rules.

Definition 1 (Moral theory)**.** *A **moral theory** is a tuple $\mathcal{T} = \langle \mathcal{R}, \succ \rangle$ where $\mathcal{R}$ is a set of rules, and $\succ \subseteq \mathcal{R} \times \mathcal{R}$ is a superiority relation over the rules.*

In the remainder of the paper, a set of moral theories is denoted $\mathfrak{T}$, and we may just say theory instead of moral theory.

If a moral theory is meant to be applied in some context (represented by for example some facts), then such a moral theory can be embedded within a larger ‘contextual’ theory. In this paper, possible contexts are left implicit, assuming that everything can be parametrised with respect to them.

When agents argue about theories to govern their own society, they consider the utility springing from these theories.

Definition 2 (Agent theory utility distribution)**.** *Let $\mathfrak{T}$ be a set of theories, $\mathbb{V}$ an ordered set of values (on which the moral utility functions are computed), and Ag a set of agents. An **agent moral theory utility distribution** is a function*

$$U : \mathfrak{T} \to \prod_{0}^{|Ag|} \mathbb{V}.$$

Given a theory and n agents, the function returns a vector of $n + 1$ values, where the first value, conventionally, indicates the total welfare for the set of agents, and the remaining values define the value of the theory for each agent. The value of the theory for agent i corresponds to the projection on the i-th element of the vector, thus $U_i(\mathcal{T}) = \pi_i(U(\mathcal{T}))$. In the remainder, $U_{Ag}(\mathcal{T})$ denotes agents' utility $\pi_0(U(\mathcal{T}))$, and $U_i(\mathcal{T})$ the utility $\pi_i(U(\mathcal{T}))$ of agent i.

3 Theory Utility

In this section, we briefly illustrate two possible approaches for computing theory utility, i.e., from rule utility and from literal utility.

3.1 Theory Utility from Rule Utility

As argued in the context of rule utilitarianism, we can determine what is the value of each rule for each agent (based on the context in which the theory is used) by introducing the following function [16].

Definition 3 (Agent rule valuation)**.** *Let Ag and $\mathbb{V}$ be, respectively, a set of agents and an ordered set of values (on which a moral utility function is computed). An **agent rule valuation** is a function*

$$V\colon Ag \times \mathcal{R} \times \mathfrak{T} \to \mathbb{V}.$$

This function assigns to every rule in every theory a value to be used in a moral utility function. Based on this definition, we establish that the elements of the agent utility distribution are computed based on the utilities of the rules (and the context in which they appear), using the following equation

$$U_i(\mathcal{T}) = F^i_{r\in\mathcal{R}}(V(Ag_i, r, \mathcal{T})) \tag{1}$$

where F^i is a function/operator that agglomerates the individual values for a set of rules into a single value, and Ag_i denotes agent i.

If we move from rule utility, as discussed above, two options are possible for computing *theory utility*: in the first case, this is computed by other elements of the vector based on a predefined function (for instance, the total welfare is the sum of the welfare values for the individual agents); in the second case, there is an individual rule depending function designed for it working on values attributed to rules according to the context in which they appear, namely, we have a function

$$V_{Ag}\colon \mathcal{R} \times \mathfrak{T} \to \mathbb{V} \tag{2}$$

from which the total welfare can be computed as

$$U_{Ag}(\mathcal{T}) = F^0_{r \in \mathcal{R}}(V_{Ag}(r, \mathcal{T})). \tag{3}$$

Functions for rule valuation can be further specified, in particular with respect to some inference mechanisms as proposed later.

3.2 Theory Utility from Literal Utility

A more fine-grained approach to articulate the way in which utility springs from any theory $\mathcal{T}$ is based on the utility of conclusions that follow from arguing on $\mathcal{T}$.

Let us first give a basic language setting. A literal is a propositional atom or the negation of a propositional atom. Given a literal ϕ, its complementary literal is a literal, denoted as $\sim \phi$, such that if ϕ is an atom p then $\sim \phi$ is its negation $\neg p$, and if ϕ is $\neg q$ then $\sim \phi$ is q. If $Prop$ is a set of propositional atoms then $Lit = Prop \cup \{\neg p \mid p \in Prop\}$ is a set of literals.

For each literal l in a set Lit of literals and given a (possibly different) set of literals $\{l_1, \ldots, l_n\}$, we can define a function λ that assigns for each agent i in Ag an utility value, i.e., the utility that the state of affairs denoted by l brings to i in a context described by $l_1, \ldots, l_n$.

Definition 4 (Agent literal valuation)**.** *Let Ag and $\mathbb{V}$ be, respectively, a set of agents and an ordered set of values. An* ***agent literal valuation*** *is a function*

$$\lambda : Ag \times Lit \times \mathrm{pow}(Lit) \rightarrow \mathbb{V}.$$

If $E(\mathcal{T}) = \{c_1, \ldots, c_m\}$ is the set of conclusions of a theory $\mathcal{T}$, then an individual agent utility can be given by agglomerating the values of all conclusions.

$$U_i(\mathcal{T}) = \underset{\forall l \in E(\mathcal{T})}{F^i} \lambda(Ag_i, l, E(\mathcal{T})). \tag{4}$$

where F^i is a function/operator that agglomerates the individual values into a single value.

As mentioned earlier, the utility of theories can be further specified by considering inference mechanisms to reason on the theories. Different mechanisms are possible, we propose next to use an argument-based reasoning setting.

4 Theory Utility in an Argumentation Setting

In this section, we sketch how the utility of a theory could be determined in an argumentation setting.

4.1 Argumentation Setting

To reason on theories, and in particular to determine justified and rejected conclusions of any theory, we adopt an argumentation setting. Different argumentation settings exist in the literature, see e.g. [14, 31, 20] amongst many others. We adopt an ASPIC$^+$-like setting, and the (internal) logical structure of arguments are specified in such a way that arguments are logical inference trees built out from rules (we adjust the definition in [31] to meet our definition of theory). In the following definition, for a given argument A, Conc returns its conclusion, Sub returns all its sub-arguments, Rules returns the set of rules in the argument and, finally, $\mathrm{TopRule}$ returns the last inference rule in the argument.

Definition 5 (Argument)**.** *Let $\mathcal{T} = (\mathcal{R}, \succ)$ be a theory where rules have the form $\psi_1, \ldots, \psi_n \Rightarrow \phi$ ($0 \leq n$), $\psi_1, \ldots, \psi_n, \phi \in Lit$. An* ***argument*** *$A$ constructed from $\mathcal{T}$ has the form $A_1, \ldots, A_n \Rightarrow_r \phi$, where*

- *A_k is an argument constructed from $\mathcal{T}$, and*
- *$r : \mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow \phi$ is a rule in $\mathcal{R}$.*

With regard to argument A, the following holds:

$$\begin{aligned}
&\mathrm{Conc}(A) = \phi\\
&\mathrm{Sub}(A) = \mathrm{Sub}(A_1) \cup \ldots \cup \mathrm{Sub}(A_n) \cup \{A\}\\
&\mathrm{TopRule}(A) = r : \mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow \phi\\
&\mathrm{Rules}(A) = \mathrm{Rules}(A_1) \cup \ldots \cup \mathrm{Rules}(A_n) \cup \{\mathrm{TopRule}(A)\}.
\end{aligned}$$

Arguments may support conflicting conclusions and thus attacks may appear between arguments. Effective attacks between arguments are defined here with respect to the superiority relation over rules which are used to build arguments.

Definition 6 (Attacks)**.** *An argument B attacks an argument A iff $\exists A' \in \mathrm{Sub}(A)$ such that $\mathrm{Conc}(B) = {\sim}\mathrm{Conc}(A')$, and $\mathrm{TopRule}(A') \not\succ \mathrm{TopRule}(B)$.*

Given a theory from which arguments are built and attacks determined, we can define an argumentation framework, along with standard argumentation semantics.

Definition 7 (Argumentation framework)**.** *Let $\mathcal{T} = (\mathcal{R}, \succ)$ be a theory. The* ***argumentation framework*** *$AF_{\mathcal{T}}$ determined by $\mathcal{T}$ is a tuple $\langle \mathcal{A}, \rightsquigarrow \rangle$ where $\mathcal{A}$ is the set of all arguments constructed from $\mathcal{T}$, and $\rightsquigarrow \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation.*

Definition 8 (Argumentation semantics)**.**

Conflict-free set**:** *A set $\mathcal{S}$ of arguments is conflict-free iff there exist no arguments A and B in $\mathcal{S}$ such that B attacks A.*

***Argument defence*:** *Let $\mathcal{S} \subseteq \mathcal{A}$ be a set of arguments. The set $\mathcal{S}$ defends an argument $A \in \mathcal{A}$ iff for each argument B attacking A there is an argument C in $\mathcal{S}$ that attacks B.*

***Complete extension*:** *Let $AF = (\mathcal{A}, \rightsquigarrow)$ and $\mathcal{S} \subseteq \mathcal{A}$. The set $\mathcal{S}$ is a complete extension of AF iff $\mathcal{S}$ is conflict-free and $\mathcal{S} = \{A \in \mathcal{A} \mid \mathcal{S}$ defends $A\}$.*

***Grounded extension*:** *A grounded extension $\mathrm{GE}(AF)$ of an argumentation framework AF is the minimal complete extension of AF.*

***Justified argument and conclusion*:** *An argument A and its conclusion are justified w.r.t. an argumentation framework AF iff $A \in \mathrm{GE}(AF)$.*

***Rejected argument and conclusion*:** *An argument A is rejected w.r.t. an argument framework AF iff $A \notin \mathrm{GE}(AF)$; its conclusion is rejected iff it is not justified.*

Other semantics could be employed, see e.g. [1, 2]. For our purposes and the sake of simplicity, we consider the grounded semantics only, and we leave how to deal with other semantics to future investigations.

4.2 Theory Utility from Rule Utility (cont'd)

On the basis of the argumentation setting to reason upon any theory, we can now provide a simple way for determining the utility of any theory from its rules. We recall that the assumption is that any rule r in any theory $\mathcal{T}$ can be associated with an utility value v (see Definition 3), which means that complying with this rule r produces utility v.

Suppose that the only rules that contribute to the utility function for an agent are the rules contributing to justified arguments/conclusions. So, assuming $\mathbb{V} = \mathbb{Z}$ one can create an agent rule valuation as follows:

$$V(Ag_i, r, \mathcal{T}) = \begin{cases} n \neq 0 & \text{if } r \in \mathrm{Rules}(A),\ A \in \mathrm{GE}(AF_{\mathcal{T}}) \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The individual utility U_i of agent i is then possibly the sum of the agent rule valuations:

$$U_i(\mathcal{T}) = \sum_{r \in \mathcal{R}} V(Ag_i, r, \mathcal{T}). \tag{6}$$

Hence, following an intuition from rule utilitarianism, this sketches how the utility of theories can be determined as the sum of the single agent rule valuations [17][3].

[3] We do not commit to any specific utility theory, which would require a more detailed machinery: see [17].

4.3 Theory Utility from Literal Utility (cont'd)

An alternative to determine theory utility consists in computing the utility from justified conclusions. For any theory $\mathcal{T}$ let us specify the set $E(\mathcal{T})$ of conclusions of $\mathcal{T}$ (Section 3.2) as follows:

$$E(\mathcal{T}) = \{\psi \mid \forall A \in \mathrm{GE}(AF_{\mathcal{T}}),\ \psi = \mathrm{Conc}(A)\}. \tag{7}$$

Hence, the function based on Equation (4) can be easily applied here.

We leave to future investigations the complete development of the two approaches sketched in this section to determine the utility of theories with respect to various contexts in an argumentation setting. Whatever the way a theory is associated with a utility value, we have then the problem of finding the 'right' theory.

5 Problem Definitions

As any theory can be associated with a utility, we may identify particular theories. For example, one may consider agents' utility optimal theories, i.e., theories maximising the agents' utility, or (strong) 'Pareto optimal theories', i.e., theories for which no agent can be made better off by making some agents worse off, or 'maximin optimal theories', i.e., theories maximising the utility of the worst off agents.

Definition 9 (Agents' utility optimal theory)**.** *Let Ag be a set of agents. A theory $\mathcal{T}^*$ is an **agents' utility optimal theory** amongst a set of theories $\mathfrak{T}$ iff there is no theory $\mathcal{T} \in \mathfrak{T}$ such that $U_{Ag}(\mathcal{T}) > U_{Ag}(\mathcal{T}^*)$.*[4]

Definition 10 (Pareto optimal theory)**.** *Let Ag be a set of agents. A theory $\mathcal{T}^*$ is a **Pareto optimal theory** amongst a set of theories $\mathfrak{T}$ iff there is no theory $\mathcal{T} \in \mathfrak{T}$ such that $U_i(\mathcal{T}^*) \leq U_i(\mathcal{T})$ for all $i \in Ag$ and $U_i(\mathcal{T}^*) < U_i(\mathcal{T})$ for some $i \in Ag$.*

Definition 11 (Maximin optimal theory)**.** *Let Ag be a set of agents. A theory $\mathcal{T}^*$ is a **maximin optimal theory** amongst a set of theories $\mathfrak{T}$ iff there is no theory $\mathcal{T} \in \mathfrak{T}$ such that $\min_{i \in Ag} U_i(\mathcal{T}) > \min_{i \in Ag} U_i(\mathcal{T}^*)$.*

We can now formulate the general problem of a moral theory elicitation.

Given: a set of agents Ag and a set of theories $\mathfrak{T}$;
Find: a specific moral theory $\mathcal{T}$ in $\mathfrak{T}$.

[4] Equivalently, we can say that a theory $\mathcal{T}^*$ is agents' utility optimal amongst a set of theories $\mathfrak{T}$ iff for all $\mathcal{T} \in \mathfrak{T}$ it holds that $U_{Ag}(\mathcal{T}) \leq U_{Ag}(\mathcal{T}^*)$.

The problem can be specified. For instance, one may seek an agents' utility optimal theory, or a Pareto optimal theory.

Definition 12 (Agents' utility theory problem)**.**

Given: *a set of agents Ag and a set of theories $\mathfrak{T}$;*

Find: *a theory $\mathcal{T}$ in $\mathfrak{T}$ which is agents' utility optimal amongst $\mathfrak{T}$.*

Definition 13 (Pareto theory problem)**.**

Given: *a set of agents Ag and a set of theories $\mathfrak{T}$;*

Find: *a theory $\mathcal{T}$ in $\mathfrak{T}$ which is Pareto optimal amongst $\mathfrak{T}$.*

Definition 14 (Maximin theory problem)**.**

Given: *a set of agents Ag and a set of theories $\mathfrak{T}$;*

Find: *a theory $\mathcal{T}$ in $\mathfrak{T}$ which is maximin optimal amongst $\mathfrak{T}$.*

Given a theory, a brute force solution is to compute every sub-theory along with a utility distribution and then retain the specific sub-theory we are after. However, this solution is of course not efficient. In the next section, we initiate possible alternative solutions by means of dialogues.

6 Moral Dialogues

A moral dialogue is the process through which agents propose their normative theories with the aim to improve on the current state-of-the-art theory. The normative system resulting from the dialogue is taken to be morally justified. Moral dialogues are based on dialogues.

Definition 15 (Dialogue)**.** *A **dialogue** is a sequence of theories $(\mathcal{T}_k)_{k=1,\dots K}$ such that*

- *theory $\mathcal{T}_1$ is an (arbitrary) initial theory;*
- *for every $\mathcal{T}_k$, there is a set of theories $\mathfrak{T}^k = \{\mathcal{T}_1^k, \dots, \mathcal{T}_n^k\}$ (proposed by some agents);*
- *theory $\mathcal{T}_{k+1} = Choice(\mathfrak{T}^k)$, where $Choice$ is a function that selects theory $\mathcal{T}_{k+1}$ out of a non-empty set $\mathfrak{T}^k$;*
- *theory $\mathcal{T}_K$ is terminal iff $\mathfrak{T}^K = \emptyset$.*

Definition 16 (Theories proposed in a dialogue)**.** *The set of theories $\mathfrak{T}^d$ proposed in a dialogue $d = (\mathcal{T}_k)_{k=1,\dots K}$ is $\bigcup_{k\in\{1,\dots K\}} \mathfrak{T}^k$.*

We can note that theory $\mathcal{T}_k$ may be included in $\mathfrak{T}^k$, possibly leading to some sort of equilibrium. However, in this paper, we are not interested in computing *equilibria* as, being interested in moral reasoning, we deal with principles and not with *moves* as in standard game theoretic approaches. For this reason, we rely on dialogues and not on games, though our dialogues may be seen as *mirroring* such games.

A dialogue is moral if, and only if, the choice function is moral. We concentrate on a few moral $Choice$ functions, each corresponding to a well established ethical/rational criterion: in an impartial choice, theories are drawn at random; following rule utilitarianism, other choices are maximising agents' utility choice, or a Pareto choice.

Definition 17 (Impartial choice)**.** *The choice function of a dialogue* $(\mathcal{T}_k)_{k=1,...K}$ *is an **impartial choice function** iff any theory* $\mathcal{T}_k$ *(*$2 \leq k$*) is drawn at random from* $\mathfrak{T} \subseteq \mathfrak{T}^{k-1}$ *with a uniform probability.*

Definition 18 (Agents' utility maximising choice)**.** *The choice function of a dialogue* $(\mathcal{T}_k)_{k=1,...K}$ *is an **agents' utility maximising choice function** iff any theory* $\mathcal{T}_k$ *(*$2 \leq k$*) is an agents' utility optimal theory amongst the set of theories* $\mathfrak{T}^{k-1}$.

Definition 19 (Pareto choice)**.** *The choice function of a dialogue* $(\mathcal{T}_k)_{k=1,...K}$ *is a **Pareto choice function** iff any theory* $\mathcal{T}_k$ *(*$2 \leq k$*) is a Pareto optimal theory amongst the set of theories* $\mathfrak{T}^{k-1}$.

Randomising theories is also a basis to mimic an intuition developed by moral philosophers such as John Rawls [32], who argued the importance of the so-called second-order impartiality, according to which norms and principles are impartially evaluated and selected by agents by demonstrating that they would be selected by a group of impartial persons who were choosing the moral rules for their society. In particular, Rawls said that a normative theory of a just society should be chosen by self-interested rational agents in the original position, i.e., a position in which agents are rational and possess broad knowledge about the world, but are denied specific information regarding their own particular identities and personal convenience. On this basis, Rawls argued that rational agents with risk aversion should yield the 'Difference Principle', according to which inequalities are to the greatest benefit of the least advantaged members of society, i.e., a maximin choice [15].

Definition 20 (Maximin choice)**.** *The choice function of a dialogue* $(\mathcal{T}_k)_{k=1,...K}$ *is a **maximin choice function** iff any theory* $\mathcal{T}_k$ *(*$2 \leq k$*) is a maximin optimal theory amongst the set of theories* $\mathfrak{T}^{k-1}$.

Example. *Let us consider a group of three autonomous individuals: agent* 1 *has high incomes because of its high salary, agent* 2 *has high incomes because of tax evasion, and agent* 3 *has low incomes. Suppose we have an initial theory* $\mathcal{T}_0$ *with utility distribution* $[5, 3, 1, 1]$

(where 5 *is the global utility), and suppose theories* $\mathcal{T}_1$*,* $\mathcal{T}_2$ *and* $\mathcal{T}_3$ *are proposed such that* $U(\mathcal{T}_1) = [8, 2, 0, 6]$ *(i.e., taxes slightly raised for upper classes, tax evasion is severely punished, and public subsidies are introduced for lower classes),* $U(\mathcal{T}_2) = [6, 2, 2, 2]$ *(i.e., taxes slightly raised for upper classes together with public subsidies for lower classes and imprisonment for tax fraud is lowered from 5 years to 3 years), and* $U(\mathcal{T}_3) = [7, 3, 3, 1]$ *(i.e., a tax evasion amnesty is proposed). If the agents' utility maximising choice is adopted then* $\mathcal{T}_1$ *is elicited, while the maximin choice yields* $\mathcal{T}_2$*, and the Pareto choice results into* $\mathcal{T}_3$*.*

One may complement agents' utility maximising choices, or Pareto choices, or maximin choices with an impartial choice when there exist multiple agents' utility, or Pareto, or maximin theories, respectively, in a set of theories $\mathfrak{T}^k$. For example, a moral dialogue $(\mathcal{T}_k)_{k=1,\dots,K}$ can have an impartial and agents' utility maximising choice function if any theory $\mathcal{T}_k$ is drawn at random from the set theories maximising agents' utility.

Whatever the moral choice function, a moral theory may not emerge from the theories proposed in a dialogue, especially if theory $\mathcal{T}_{k-1}$ is not included in $\mathfrak{T}^{k-1}$. We further investigate moral dialogues in that regard in the next section.

7 Moral Optimising Dialogues

The use of dialogues and their iterative nature suggests a few different (search) strategies to find an optimal theory in a set of theories. We do not propose any particular strategies in this paper, leaving them to future work. In this section, we simply give some basic properties of moral dialogues that may characterise such strategies.

7.1 Agents' Utility Optimising Dialogue

For the terminal theory of a dialogue to be agents' utility optimal amongst the theories proposed in the dialogue, it is sufficient that the dialogue has an agents' utility maximising choice function whose output theory $\mathcal{T}_k$ is always included in the proposed theories $\mathfrak{T}^k$.

Proposition 1. *The terminal theory of a dialogue* $d = (\mathcal{T}_k)_{k=1,\dots K}$ *with an agents' utility maximising choice function is agents' utility optimal amongst the set of theories* $\mathfrak{T}^d$ *proposed in the dialogue if for any* $\mathcal{T}_k$*, it holds that* $\mathcal{T}_k \in \mathfrak{T}^k$*.*

Therefore, such an agents' utility maximising dialogue solves the problem given in Definition 12, where $\mathfrak{T} = \mathfrak{T}^d$.

However, the terminal theory may not be a strict 'improvement' of the initial theory. For this reason, one may consider dialogues to elicit agents' utility optimal theories based on the idea of improving theories.

Definition 21 (Agents' utility improving theory)**.** *Let Ag a set of agents. A theory $\mathcal{T}^*$ is an **agents' utility improvement** of a theory $\mathcal{T}$ iff $U_{Ag}(\mathcal{T}^*) > U_{Ag}(\mathcal{T})$.*

Proposition 2. *A theory is an agents' utility optimal theory amongst a set of theories $\mathfrak{T}$ iff there exist no agents' utility improvements in $\mathfrak{T}$ of the theory.*

Proof. By Definition 9, a theory $\mathcal{T}^*$ is agents' utility optimal amongst a set of theories $\mathfrak{T}$ iff there exists no agents' utility improving theories in $\mathfrak{T}$ of $\mathcal{T}^*$. □

Consequently, the initial theory is not optimal if there exists an improvement.

Proposition 3. *The terminal theory of a dialogue $d = (\mathcal{T}_k)_{k=1,...K}$ with a agents' utility maximising choice function is agents' utility optimal amongst the set of theories $\mathfrak{T}^d$ proposed in the dialogue and it is an agents' utility improvement of the initial theory, if for any $\mathcal{T}_k$, it holds that $\mathcal{T}_k \in \mathfrak{T}^k$, and there exists a theory $\mathcal{T}_k$ which is an agents' utility improvement of $\mathcal{T}_{k-1}$.*

In other words, if there exists no improvement in a dialogue then the initial theory remains the optimal theory, and a moral dialogue is not necessary to find the optimal theories.

7.2 Pareto Optimising Dialogue

Moral dialogues can be similarly tuned to elicit Pareto optimal theories.

Proposition 4. *The terminal theory of a dialogue $d = (\mathcal{T}_k)_{k=1,...K}$ with a Pareto choice function is Pareto optimal amongst the set of theories $\mathfrak{T}^d$ proposed in the dialogue if for any $\mathcal{T}_k$, it holds that $\mathcal{T}_k \in \mathfrak{T}^k$.*

Therefore, a Pareto improving dialogue solves the problem given in Definition 13, where $\mathfrak{T} = \mathfrak{T}^d$.

As the terminal theory may not be an improvement of the initial theory, we can consider Pareto improving theories, i.e., theories leading to a utility gain, without any agents being made worse off.

Definition 22 (Pareto improving theory)**.** *Let Ag a set of agents. A theory $\mathcal{T}^*$ is a **Pareto improvement** of a theory $\mathcal{T}$ iff $U_i(\mathcal{T}^*) \leq U_i(\mathcal{T})$ for all $i \in Ag$ and $U_i(\mathcal{T}^*) < U_i(\mathcal{T})$ for some $i \in Ag$.*

Proposition 5. *A theory is a Pareto optimal theory amongst a set of theories $\mathfrak{T}$ iff there exist no Pareto improvements in $\mathfrak{T}$ of the theory.*

Proof. By Definition 10, a theory $\mathcal{T}^*$ is Pareto optimal amongst a set of theories $\mathfrak{T}$ iff there is no Pareto improving theory of $\mathcal{T}^*$. □

Proposition 6. *The terminal theory of a dialogue* $d = (\mathcal{T}_k)_{k=1,...K}$ *with a Pareto choice function is Pareto optimal amongst the set of theories* $\mathfrak{T}^d$ *proposed in the dialogue and it is an agents' utility improvement of the initial theory, if for any* $\mathcal{T}_k$*, it holds that* $\mathcal{T}_k \in \mathfrak{T}^k$*, and there exists a theory* $\mathcal{T}_k$ *which is a Pareto improvement of* $\mathcal{T}_{k-1}$*.*

7.3 Maxmin Optimising Dialogue

Similarly to agents' utility and Pareto improving choice functions, maximin can be accommodated in dialogues.

Proposition 7. *The terminal theory of a dialogue* $d = (\mathcal{T}_k)_{k=1,...K}$ *with a maximin choice function is maximin optimal amongst the set of theories* $\mathfrak{T}^d$ *proposed in the dialogue if for any* $\mathcal{T}_k$*, it holds that* $\mathcal{T}_k \in \mathfrak{T}^k$*.*

Therefore, a maximin improving dialogue solves the problem given in Definition 14, where $\mathfrak{T} = \mathfrak{T}^d$.

Definition 23 (Maximin improving theory)**.** *Let* Ag *a set of agents. A theory* $\mathcal{T}^*$ *is a* ***maximin improvement*** *of a theory* $\mathcal{T}$ *iff* $\min_{i \in Ag} U_i(\mathcal{T}) > \min_{i \in Ag} U_i(\mathcal{T}^*)$.

Proposition 8. *A theory is a maximin optimal theory amongst a set of theories* $\mathfrak{T}$ *iff there exist no maximin improvements in* $\mathfrak{T}$ *of the theory.*

Proof. By Definition 10, a theory $\mathcal{T}^*$ is maximin optimal amongst a set of theories $\mathfrak{T}$ iff there is no maximin improving theory of $\mathcal{T}^*$. □

Proposition 9. *The terminal theory of a dialogue* $d = (\mathcal{T}_k)_{k=1,...K}$ *with a maximin choice function is maximin optimal amongst the set of theories* $\mathfrak{T}^d$ *proposed in the dialogue and it is an agents' utility improvement of the initial theory, if for any* $\mathcal{T}_k$*, it holds that* $\mathcal{T}_k \in \mathfrak{T}^k$*, and there exists a theory* $\mathcal{T}_k$ *which is a maximin improvement of* $\mathcal{T}_{k-1}$*.*

Notice that, while Rawls justifies the maximin choice on the assumptions of agents with risk aversion and an impartial choice (from the 'original position'), we do not aim here at directly covering this justification.

8 Related Work

Autonomy and agency are central properties in robotic systems, assisted living applications, responsible vehicles and many other application domains. As the complexity of the situations faced by such autonomous agents is increasing, agents' decision-making has to take into account new elements like ethical and moral considerations. For these reasons, the last years have seen a raising number of projects tackling this issue, e.g., the REINS project about responsibility [7], the ETHICAA project about defining regulation modes to manage ethical conflicts within socio-technical systems [3], and the MIT Moral Machine about autonomous self-driving cars.

New approaches and position papers are also appearing in the literature: Charisi *et al.* [8] report and discuss the immediate challenges faced by the problem of the engineering of machine ethics, and Dignum [12] describes the leading ethics theories, and proposes alternative ways to ensure ethical behaviour by autonomous agents. Our approach is closer to the goals of the ETHICAA project regarding the *reasoning perspective* where a representation of ethical principles is provided together with decision-making models [3]. Following [25], our agents converge on a moral theory by assessing the arguments and values at stake under specific ethical principles.

Our approach can also be related to the work on computational justice in self-organising electronic institutions proposed by Pitt *et al.* [30], where agents can agree on a set of rules to self-organise and self-regulate a distribution of resources [29]. We take a more abstract stance, and provide a dialogical framework for agents' deliberation about moral theories.

An argumentation-based perspective to ethical systems design is proposed by Verheij [37]. Based on the assumptions that ethical system's decisions depend on the *values* and the rules embedded in the system, Verheij [37] studies the issue of the comparison of values in value-guided argumentation, through techniques connecting qualitative and quantitative primitives from evidential argumentation applied to value-guided argumentation. The problem tackled in this paper is different from our goal, as well as the methodology proposed to address it. The only shared point is the fact of relying on rules to represent ethical decisions and theories, and on (two different kinds of) formal argumentation frameworks.

Another approach to ethics in a multi-agent scenario has been proposed by Cointe *et al.* [9], where the authors studied the problem of ethical judgement, i.e., the assessment of the appropriateness of agents' behaviours with respect to moral convictions and ethical principles. No specific ethical theories are considered. Again, the goal and methodology to address it differ from our paper, even if the two approaches may be seen as complementary, as after the elicitation of a new moral theory to ensure the welfare the society then the agents need to judge whether the behaviour of the other agents complies or not with it.

Dennis *et al.* [11] propose a theoretical framework for ethical plan selection that can be formally verified. The authors formally verify that the agent chooses to execute the most ethical available plan, given its belief set. This approach focuses on the formal verification of ethical decision-making within autonomous agents. This goal differs from ours, as we are interested in a dialogical assessment of moral theories. Moreover, further differences arise, i.e., ethical principles are considered as abstract while we take into account three well known moral theories. The common point is the representation of moral theories using (ethical) rules.

Other contributions have been presented about formal models of ethics but with different goals than the one we addressed in this paper. A logic-based approach to model moral reasoning with deontic constraints is presented by Wiegel [40]. It is an approach to represent the theory of good, and ethical reasoning is addressed in the meta-level with the aim to support the adoption of a less restrictive model of behaviour. Other approaches like [4, 35] provide a direct translation of some well-known ethical principles as Kant's Categorical Imperative or Thomas Aquinas' Doctrine of Double Effect into logic programming.

Other approaches aim at formalising, with the help of logic, classical game theory and evolutionary game theory, the influence of social preferences and moral values on the decision-making process of autonomous agents, with special emphasis on fairness values based on Rawls' maxmin criterion. More precisely, Lorini [22] proposes a logical formalisation of social preferences based on Rawls' fairness principle. Lorini [23] again proposes a logical formalisation of the relationship between moral values and preferences. Lorini and Muhlenbernd [24] provide a game-theoretic and evolutionary analysis of Rawls' fairness principle and its connection with the concepts of responsibility and guilt. The latter work focuses on the integration of moral and ethical aspects into the decision-making processes of rational autonomous agents and it is related with some existing economic theories of morality [5]. Another recent game-theoretic account of morality and its connection with guilt has been proposed in the area of multi-agent systems by Pereira *et al.* [27].

9 Conclusion

Autonomous agents raise the problem of eliciting moral norms governing the behaviour of the agents in a society. We addressed the elicitation problem of moral norms as normative theories by proposing a generic dialogical framework. We showed how the framework can accommodate different and well-established moral choices (such as utility maximising, Pareto and maximin choices), possibly leading to different moral theories. By doing so, different ethical criteria can be compared. To the best of our knowledge, there exists no other formal comparative framework designed to deal with the elicitation of moral theories in MAS showing the above mentioned features.

Future work directions are multiple. They include a deeper comparative investigation of moral principles. From the philosophical point of view, we also aim to extend this framework to cope with the theory of the *ideal observer* (as a condition for ensuring impartiality in ethics) [18, 6]. The idea is that the moral judgement is well founded if it is accepted by the set of fully-informed agents in the society based on well founded rational arguments. A goal is to prove that these conditions are equivalent to a situation where moral rules are selected by an infinite set of agents who know everything. Eventually, given the attention paid to computational self-organising institutions (see e.g. [30, 33]), it would be interesting to investigate how dialogues on moral theories can fit with these implementations.

References

[1] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowledge Eng. Review*, 26(4):365–410, 2011.

[2] Pietro Baroni, Guido Governatori, and Régis Riveret. On labelling statements in multi-labelling argumentation. In *Proc. of the 22nd Euro. Conf. on Artificial Intelligence*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 489–497. IOS Press, 2016.

[3] Aline Belloni, Alain Berger, Olivier Boissier, Grégory Bonnet, Gauvain Bourgne, Pierre-Antoine Chardel, Jean-Pierre Cotton, Nicolas Evreux, Jean-Gabriel Ganascia, Philippe Jaillon, Bruno Mermet, Gauthier Picard, Bernard Rever, Gaële Simon, Thibault de Swarte, Catherine Tessier, François Vexler, Robert Voyer, and Antoine Zimmermann. Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *Papers from the 2015 AAAI Workshop on Artificial Intelligence and Ethics*. AAAI Press, 2015.

[4] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *Proc. of 20th Inter. Conf. on Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548. Springer, 2015.

[5] K. Binmore. *Natural Justice*. Oxford University Press, 2005.

[6] Richard Brandt. *Ethical Theory*. Prentice Hall, 1959.

[7] Jan M. Broersen. Responsible intelligent systems - the REINS project. *KI*, 28(3):209–214, 2014.

[8] Vicky Charisi, Louise A. Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan F. T. Winfield, and Roman Yampolskiy. Towards moral autonomous systems. *CoRR*, abs/1703.04741, 2017.

[9] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *Proc. of the 15th Inter. Conf. on Autonomous Agents & Multiagent Systems*, pages 1106–1114. ACM, 2016.

[10] Boudewijn de Bruin and Luciano Floridi. The ethics of cloud computing. *Science and Engineering Ethics*, 23(1):21–39, 2017.

[11] Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[12] Virginia Dignum. Responsible autonomy. In *Proc. of the 26th Inter. Joint Conf. on Artificial Intelligence*, pages 4698–4704. ijcai.org, 2017.

[13] Allan M. Feldman. welfare economics. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

[14] Guido Governatori, Michael J. Maher, Grigoris Antoniou, and David Billington. Argumentation semantics for defeasible logic. *J. Log. Comput.*, 14(5):675–702, 2004.

[15] John C. Harsanyi. Can the maximin principle serve as a basis for morality? a critique of john rawls's theory. *American Political Science Review*, 69(2):594–606, 1975.

[16] John C. Harsanyi. Morality and the theory of rational behavior. *Social Research*, 44:623–656, 1977.

[17] John C. Harsanyi. Rule utilitarianism and decision theory. *Erkenntnis*, 11(1):25–53, 1977.

[18] Troy Jollimore. Impartiality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[19] I. Kant and M.J. Gregor. *Practical Philosophy*. Kant, Immanuel, 1724-1804. Works. Engl. 1992. Cambridge University Press, 1999.

[20] Ho-Pun Lam, Guido Governatori, and Régis Riveret. On ASPIC$^+$ and Defeasible Logic. In *Proc. of 6th Conf. on Computational Models of Argument*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 359–370. IOS Press, 2016.

[21] Richard Lindley. *Autonomy*. Atlantic Highlands, NJ: Humanities Press International, 1986.

[22] Emiliano Lorini. From self-regarding to other-regarding agents in strategic games: a logical analysis. *Journal of Applied Non-Classical Logics*, 21(3-4):443–475, 2011.

[23] Emiliano Lorini. A logic for reasoning about moral agents. *Logique & Analyse*, 58:177–218, 2016.

[24] Emiliano Lorini and Roland Mühlenbernd. The long-term benefits of following fairness norms under dynamics of learning and evolution. *Fundam. Inform.*, 158(1-3):121–148, 2018.

[25] Michel Meyer. *Principia Moralia*. Fayard, 2013.

[26] Brent Daniel Mittelstadt and Luciano Floridi. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, 22(2):303–341, 2016.

[27] Luís Moniz Pereira, Tom Lenaerts, Luis A. Martinez-Vaquero, and The Anh Han. Social manifestation of guilt leads to stable cooperation in multi-agent systems. In Kate Larson, Michael Winikoff, Sanmay Das, and Edmund H. Durfee, editors, *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*, pages 1422–1430. ACM, 2017.

[28] Luís Moniz Pereira and Ari Saptawijaya. *Programming Machine Ethics*, volume 26 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer, 2016.

[29] Jeremy Pitt, Dídac Busquets, and Régis Riveret. Procedural justice and fitness for purpose of self-organising electronic institutions. In *Proc. of the 16th Inter. Conf. on Principles and Practice of Multi-Agent Systems*, pages 260–275. Springer, 2013.

[30] Jeremy Pitt, Dídac Busquets, and Régis Riveret. The pursuit of computational justice in open systems. *AI Soc.*, 30(3):359–378, 2015.

[31] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[32] John Rawls. *A Theory of Justice*. London: Oxford University Press, 1971.

[33] Régis Riveret, Alexander Artikis, Jeremy V. Pitt, and Erivelton G. Nepomuceno. Self-governance by transfiguration: From learning to prescription changes. In *Proc. of the 8th IEEE Inter. Conf. on Self-Adaptive and Self-Organizing Systems*, pages 70–79. IEEE Computer Society, 2014.

[34] Stuart J. Russell. Provably beneficial artificial intelligence. *Exponential Life, The Next Step*, 2017.

[35] Ari Saptawijaya and Luís Moniz Pereira. Logic programming for modeling morality. *Logic Journal of the IGPL*, 24(4):510–525, 2016.

[36] J. B. Schneewind. *The Invention of Autonomy*. Cambridge University Press, 1998.

[37] Bart Verheij. Formalizing value-guided argumentation for ethical systems design. *Artif. Intell. Law*, 24(4):387–407, 2016.

[38] Abraham Wald. *Statistical Decision Functions*. Wiley, 1950.

[39] Toby Walsh, editor. *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop*, volume WS-15-02 of *AAAI Workshops*, 2015.

[40] Vincent Wiegel and Jan van den Berg. Combining moral theory, modal logic and mas to create well-behaving artificial agents. *I. J. Social Robotics*, 1(3):233–242, 2009.

[41] Michael Wooldridge and Nicholas R. Jennings. Agent theories, architectures, and languages: A survey. In *Proc. of the Workshop on Agent Theories, Architectures, and Languages on Intelligent Agents*, ECAI-94, pages 1–39. Springer, 1995.

Received \jreceived