Proceedings of the 1st Workshop on Ethics in the Design of Intelligent Agents (EDIA) 2016

Grégory Bonnet, Maaike Harbers, Koen Hindriks, Mike Katell, Catherine Tessier (eds.)

Workshop co-located with the
22th European Conference on Artificial Intelligence (ECAI 2016)
August 30th 2016, The Hague, Holland

Preface

Grégory Bonnet[*] Maaike Harbers[†] Koen Hindriks[‡] Mike Katell[§] Catherine Tessier[¶]

The development of Artificial Intelligence is experiencing a fruitful period of incredible progress and innovation. After decades of notable successes and disappointing failures, AI is now poised to emerge in the public sphere and completely transform human society, altering how we work, how we interact with each other and our environments, and how we perceive the world. Designers have already begun implementing AI that enables machines to learn new skills and make decisions in increasingly complex situations. The result is that these machines - also called intelligent agents - decide, act and interact in shared and dynamic environments under domain constraints, where they may interact with other agents and human beings to share tasks or execute tasks on behalf of others. Search engines, self-driving cars, electronic markets, smart homes, military technology, software for big data analysis, and care robots are just a few examples.

As intelligent agents gain increased autonomy in their functioning, human supervision by operators or users decreases. As the scope of the agents activities broadens, it is imperative to ensure that such socio-technical systems will not make irrelevant, counter-productive, or even dangerous decisions. Even if regulation and control mechanisms are designed to ensure sound and consistent behaviors at the agent, multi-agent, and human-agent level, ethical issues are likely to remain quite complex, implicating a wide variety of human values, moral questions, and ethical principles. The issue is all the more important as intelligent agents encounter new situations, evolve in open environments, interact with other agents based on different design principles, act on behalf of human beings and share common resources. To address these concerns, design approaches should envision and account for important human values, such as safety, privacy, accountability and sustainability, and designers will have to make value trade-offs and plan for moral conflicts. For instance, we may want to design self-driving cars to exhibit human-like driving behaviors, rather than precisely following road rules, so that their actions are more predictable for other road users. This may require balancing deontic rule-following, utility maximization, and risk assessment in the agent's logic to achieve the ultimate goal of road safety.

Questions to be asked here are: How should we encode moral behavior into intelligent agents? Which ethical systems should we use to design intelligent, decision-making machines? Should end-users have ultimate control over the moral character of their devices? Should an intelligent agent be permitted to take over control from a human operator? If so, under what circumstances? Should an intelligent agent trust or cooperate with another agent embedded with other ethical principles or moral values? To what extent should society hold AI researchers and designers responsible for their creations and choices?

This workshop focuses on two questions: (1) what kind of formal organizations, norms, policy models, and logical frameworks can be proposed to deal with the control of agents' autonomous behaviors in a moral way?; and (2) what does it mean to be responsible designers of intelligent agents? The workshop welcomes contributions from researchers in Artificial Intelligence, Multi-Agent Systems, Machine Learning, Case-based reasoning, Value-based argumentations, AI and Law, Ontologies, Human Computer Interaction, Ethics, Philosophy, and related fields.

The topics of interest include (but are not limited to):

- machine ethics, roboethics, machines and human dignity
- reasoning mechanisms, legal reasoning, ethical engine

[*] University of Caen Normandy, UMR CNRS 6072 GREYC, France, email: gregory.bonnet@unicaen.fr
[†] Delft University, Interactive Intelligence Group, The Netherlands, email: m.harbers@tudelft.nl
[‡] Delft University, Interactive Intelligence Group, The Netherlands, email: k.v.hindriks@tudelft.nl
[§] University of Washington, Information School, USA, email: mkatell@uw.edu
[¶] Onera, France, email: catherine.tessier@onera.fr

- authority sharing, responsibility, delegating decision making to machines
- organizations, institutions, normative systems
- computational justice, social models
- trust and reputation models
- mutual intelligibility, explanations, accountability
- consistency, conflicts management, validation
- philosophy, sociology, law
- applications, use cases
- societal concerns, responsible innovation, privacy Issues
- individual ethics, collective ethics, ethics of personalization
- value sensitive design, human values, value theory

Ten papers were submitted to EDIA, nine of them have been accepted for presentation after being reviewed by three or four members of the Program Committee. The accepted papers have been organized in two sessions:

1. Ethical issues and ethical application of intelligent agents (four papers)
2. Ethical models of intelligent agents (five papers)

The EDIA workshop would not have been possible without the support of many people. First of all, we would like to thank the members of the Program Committee for providing timely and thorough reviews of the papers submitted for the EDIA Workshop. We are also very grateful to all of the authors who submitted papers. We would also like to thank Bertram Malle and Jeremy Pitt for accepting the invitation to give a talk at the workshop. We also thank the organizers of ECAI 2016.

Program committee

- Huib Aldewereld, Delft University of Technology, The Netherlands
- Mark Alfano, Delft University of Technology, The Netherlands
- Peter Asaro, The New School, USA
- Olivier Boissier, Mines Saint-Etienne, France
- Tibor Bosse, Vrije Universiteit Amsterdam, The Netherlands
- Gauvain Bourgne, Universit Pierre et Marie Curie, France
- Selmer Bringsjord, Rensselear Polytechnic Institute, USA
- Joanna Bryson, University of Bath, UK
- Pompeu Casanovas, Royal Melbourne Institute of Technology, Melbourne, Australia
- Nigel Crook, Oxford Brookes University, UK
- Michal Dewyn, Ghent University, Belgium
- Sjur Dyrkolbotn, Durham University and Utrecht University, UK and The Netherlands
- Isabel Ferreira, University of Lisbon, Portugal
- Jean-Gabriel Ganascia, Universit Pierre et Marie Curie, France
- Pim Haselager, Radboud University, The Netherlands
- Marilena Kyriakidou, Coventry University, UK
- Bertram Malle, Brown University, USA
- Pablo Noriega, Intitut d'Investigaci en Intelligncia Artificial Barcelona, Spain
- Jeremy Pitt, Imperial College London, UK
- Thomas Powers, Center for Science, Ethics and Public Policy, USA
- Lambr Royakkers, Eindhoven University of Technology, The Netherlands
- Giovanni Sartor, European University of Florence, Italy
- Aimee van Wynsberghe, University of Twente, The Netherlands
- Pieter Vermaas, Delft University of Technology, The Netherlands

Organization committee

- Grégory Bonnet, Normandy University, France
- Maaike Harbers, Delft University of Technology, The Netherlands
- Koen V. Hindriks, Delft University of Technology, The Netherlands
- Michael A. Katell, University of Washington, USA
- Catherine Tessier, Onera, France

Contents

Ethical Issues and Ethical Application of Intelligent Agents **5**

Ethical Models of Intelligent Agents **28**

MOOD: Massive Open Online Deliberation Platform *A practical application*

Ilse Verdiesen and **Martijn Cligge** and **Jan Timmermans** and **Lennard Segers** and **Virginia Dignum** and **Jeroen van den Hoven** [1]

Abstract. Nowadays, public debates often take place on social media platforms like Facebook or Twitter and can be characterized as asynchronous, protracted and ill-structured. The Massive Open Online Deliberation (MOOD) platform aims to structure these debates. Essential is that the platform can differentiate between the moral acceptability and the social acceptance of a debate outcome. We briefly describe the e-deliberation process and look at two existing debate platforms, *Liquidfeedback* and *Debatehub*. We design and build a prototype that mainly focuses on: (1) a method to differentiate and validate *facts* and *opinions*, and (2) a mechanism that maps both the *social acceptance* and the *moral acceptability* of debate outcomes. We research these ethical concepts more in depth and implement several techniques, such as a voting mechanism, in a working prototype that supports a four stage deliberation process. In future applications, machine learning techniques can be integrated in the platform to perform sentiment analysis on a debate.

1 INTRODUCTION

Public deliberation is an important component of decision-making in a democracy. Deliberation can result in an increased likelihood of justifiable policies, can help to identify incompatible moral values and can help people to get a broader perspective on policy questions [9]. The internet could be a valuable medium for public deliberation, because it can be a tool for information dissemination and long distance communication [20]. It allows citizens to share their opinion more easily. However, the debates that are currently held on the internet often take place on social media platforms like Facebook or Twitter and can therefore be characterized as asynchronous, protracted and ill-structured. E-deliberation platforms aim to structure these debates and their respective outcomes. These outcomes can be used by policy makers to make better decisions. In the field of ethics, the differentiation between social acceptance and moral acceptability is essential for the judgment on policies. Furthermore, public debates can be marginally ethical, as they occasionally contain discriminating content, and have statements that can be accepted, or not, by a majority of the crowd [21]. An example of this is a debate on banning polluting vehicles in the city center. This proposal can be accepted by local residents, but unaccepted by downtown business owners. Also, one could question if it is morally acceptable to prohibit access to city centers for potential customers and suppliers of businesses. On the other hand, for local residents the air quality is very important. E-deliberation platforms facilitate debates which should take the views of both the majority as the minority into account, and therefore strive to be ethically just [21]. However, existing platforms often lack the ability to do so. In this paper, we propose our vision of a refined e-deliberation platform that takes into account the shortcomings of existing platforms by proposing a conceptual design and working prototype.

The paper is structured as follows: in section 2 we introduce the theoretical concepts underlying our design, describe related work in the field of deliberation processes and we analyze some existing platforms that support these processes. Section 3 shows the design choices and the methodologies used for our prototype. In section 4 we demonstrate the implementation and give insight in the framework we used to develop the platform. In the final section we discuss the limitations of our work and provide direction for further research.

2 RELATED WORK

In this section we describe the differentiation between facts and values, the concept of moral acceptability and social acceptance, and the e-deliberation process in general. We also look at two existing platforms that support this process. We analyze their shortcomings and based on these, we state the aspects we have focused on in the design of our prototype.

2.1 Facts and values

The distinction between facts and values is a much-debated concept in the world of ethics. Many philosophers have had their thoughts on how to filter descriptive statements from normative statements. Descriptive statements, also referred to as factual statements, describe factual matters and can be used to assert, deny or communicate about facts [13]. Normative statements, which can also be viewed as value judgments, deal with how people judge human decisions and conduct [16]. They are concerned with how people value factual matters and circumstances. We adhere to this distinction in developing our prototype.

2.2 Moral acceptability

Morality is concerned with the distinction between right and wrong and contains principles for good and bad behavior. These principles depend on the political, cultural and religious context they are defined in [6]. They govern our thoughts, emotions and behavior and can be viewed at a personal, interpersonal or collective level [4]. Morality can also be studied on a system level from a more functional approach and can be described as: *'Moral systems are interlocking*

[1] Delft University of Technology, The Netherlands, email authors: ep.verdiesen@student.tudelft.nl, m.e.cligge@student.tudelft.nl, l.h.j.segers@student.tudelft.nl, j.timmermans-1@student.tudelft.nl

sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make social life possible.' [8, p. 368]. This systematic approach resulted in the Moral Foundations Theory which uses a moral reasoning model based on the principles of harm, fairness, liberty, loyalty, authority, and purity [21]. We use these principles to define the moral acceptability of the alternatives proposed in the debate process.

2.3 Social acceptance

Social acceptance is a combination of individual feelings, perceived benefits and risks and above all, it is a social process in which people are influenced by various types of interactions. Views and available information are important for social acceptance [10]. Research shows that indicators for social acceptance are knowledge, fear and perceptions of the public [1]. We found that literature on measuring social acceptance is scarce. We turned to the field of ethics and looked at the Social Choice theory which provides a theoretical framework to reach a collective decision on social welfare. This theory is based on combining individual opinions, preferences and interests of people and links welfare economics and voting theory to aggregate preferences and behaviors of individuals. We define social acceptance as the collective decision on the preferences of individuals.

2.4 (E)-deliberation

In this paper, we define deliberation as a critical examination of a certain issue where the examination is based on the weighting of pro- and con arguments for that issue. A deliberative process allows multiple participants to receive and exchange information, to critically examine this information, to form a collective judgment (based on the provided information) about a certain issue, which determines the decision-making on a certain issue [7]. E-deliberation platforms are platforms that make use of the modern online communication technologies to support such a deliberation process. The platforms capture collective judgments regarding complex social and political issues, such as decision-making over referendums, trade treaties and the use of killer-robots. These platforms intend to overcome legitimacy problems that may arise in public debates and public decision-making in controversial and adversarial arenas. E-deliberation platforms can be used to structure these deliberation processes by providing logic to support reasoning, voting procedures and reputation mechanisms [21]. E-deliberation platforms can be used by decision makers and citizens, to receive the opinions and information from debate participants on certain topics. For example, a decision maker might use it to introduce legislative proposals to citizens and to subsequently see how citizens evaluate these proposals via the collective judgment of the crowd.

2.5 Analysis of existing e-deliberation platforms

In order to get an understanding of the characteristics of the available e-deliberation platforms and to see if these platforms can be refined, we analyzed two existing platforms; LiquidFeedback and Debate Hub. We choose these two platforms because, in our opinion, these are two of the most investigated platforms and we were constrained by a limited amount of research time. In this analysis we mainly focused on how the deliberative process is implemented, how the collective judgments of the crowd are formed and how facts and values are differentiated and evaluated in order to identify gaps in the existing platforms which we use as input for our prototype.

2.5.1 LiquidFeedback

LiquidFeedback is designed and built by the Public Software Group of Berlin. The deliberation process consists of four phases; the admission phase, the discussion phase, the verification phase and the voting phase, where each phase has a fixed amount of time. Users of the platform can initiate a debate by proposing a certain issue, for example *'What should the town council do in order to improve the air quality in the city center?'*. Proposing of issues takes place in the admission phase, where users can support certain issues by voting. In the next step of the admission phase participants can provide alternatives to the proposed issues. An example of an alternative for the earlier described issue could be *'Polluting vehicles should be banned from the city center in the weekend'*. A discussion on a topic follows after a issue reached a certain quorum of votes in the admission phase. A discussion consists of the earlier mentioned alternatives and suggestions provided by discussants to improve the proposed alternatives. Users who provided issues and alternatives can choose to update their draft versions, based on the provided suggestions. After the discussion phase, discussants enter the verification phase. In the verification phase it is not possible anymore to change the draft alternatives, although new alternatives can still be added to the list of alternatives. At the end of the verification phase, users need to vote again on the list of alternatives. Only the alternatives that reached a certain quorum enter the next phase, which is the voting phase. This second quorum reduces the workload for participants in the voting phase. In the voting phase, participants can vote against of in favor of remaining alternatives which have passed the second quorum [2]. The voting mechanism for this last phase is conform the Schulze method, which will be explained in section 3.4 of this paper. An advantage of the Schulze method is that it takes minorities into account, so that alternatives that have a low amount of votes still have chance to reach the quorum.

LiquidFeedback is a well substantiated e-deliberation platform. However, we found that it could be improved in some areas. Firstly, LiquidFeedback does not elaborate on the differentiation of facts and values. If someone provides an alternative in the first three phases of the deliberation process, where is this alternative based on? Is it based on an opinion of someone, or is it based on a fact with corresponding literature? The platform does not explain how facts and opinions (values) are differentiated and how facts and corresponding sources are evaluated. Secondly, the platform does not differentiate in the outcome between social acceptance and moral acceptability. Social acceptance and moral acceptability often differ and that differentiation is important for decision-making and judgment [21]. The exact differences will be defined in section 3.2 and 3.3. Thirdly, in our opinion is LiquidFeedback a platform where participants can only provide alternatives for certain issues and subsequently modify these alternatives when participants do not support them. We miss a debate structure which is more focused on providing pro-and con arguments with facts and corresponding literature, just like is done during a "real world" offline debate. These aspects are in our opinion crucial for a well-structured deliberation process, because requiring participants to add literature could result in a deliberation process of higher quality.

2.5.2 Debate Hub

The second existing platform we analyzed is Debate Hub. This platform is an initiative from the Open University's Knowledge Management Institute. The platform consists of debates where people can provide debate topics, alternatives, and arguments. It does not have a well-defined deliberation process with different phases and fixed amounts of time as LiquidFeedback has, however, it has some sequence which users have to follow. The first step is initiating a debate topic or issue, such as the example provided in section 2.2.1; *'What should the town council do in order to improve the air quality in the city center?'*. After that, participants can add alternatives, like; *'Polluting vehicles should be banned from the city center in the weekend'*. Consequently, participants can add pro-or con arguments to these alternatives. The structure of the argument form allows participants to add literature to their arguments. Participants can vote on alternatives and arguments, but there is no voting mechanism that filters out the most accepted alternatives or arguments like LiquidFeedback has.

After analyzing Debate Hub, we found that Debate Hub has a very different setup compared to LiquidFeedback, since it does not have a deliberation process with distinctive phases and fixed times. The debates pages are more like forms to structure an online debate. In our opinion, the following aspects could be improved; firstly, there is no quorum for initiating a debate. By not implementing a quorum, there will be many debates without any participants. Secondly, although there is some distinction between facts and values, the facts are not validated. Thirdly, there is no distinction between social acceptance and moral acceptability. Users only can show their support for certain alternatives or arguments, but it is not clear how users evaluate the moral acceptability of certain alternatives or arguments. Lastly, there is no voting method that takes minorities into account.

2.6 Discussion related work

Based on the previous section we can conclude that the two analyzed platforms are complete, but have drawbacks in some areas. LiquidFeedback has a deliberation process with distinctive phases in which results of the most accepted alternatives are listed, while Debate Hub has a very clear way of structuring the debate itself by letting users provide debate topics or issues, alternatives and pro-and con arguments (just like in "real world " offline debates). We built a prototype that focuses on the one hand on combining the best of both platforms (by using parts of the debate page structure of Debate Hub and by using parts of the deliberation process of LiquidFeedback) and on the other hand on aspects of both platforms that could be improved. We defined a design objective for our prototype which is based on the earlier described analysis. Our design objective mainly focuses on the following aspects: (1) a method to differentiate and validate *facts* and *opinions*, and (2) a mechanism that supports both the *social acceptance* and the *moral acceptability* of debate outcomes.

3 METHODOLOGY

In this section we describe the methodologies we used in our deliberation process design and we state which methods we implemented in our platform.

3.1 Facts and values

The goal of differentiating between facts and values for our system is to have a clear discussion that is based on facts, and let participants have a discussion over values which are derived from those facts. We think that by keeping the structure of the debate page of Debate Hub, we are able to structure the debate in such a way that participants have to provide a fact with the corresponding source for every argument they make. The structure of the page where people can add an argument with facts requires users to add a URL which supports their facts. This will be explained in section 4.1 in more detail. To validate the facts and sources provided by participants, we use the methodology of online encyclopedia Wikipedia. Wikipedia implemented crowd-sourcing technology, where users (the crowd or editors) have the responsibility of (1) adding content to the encyclopedia and (2) validating all of the content. This is done by panels of experts. The composition of these panels is formed throughout the existence of the website. Groups of active editors are specialized in certain topics, and if false content on certain pages exists, they will correct this content [18]. We incorporate this concept in our platform, by letting users report on facts they think are not correct. If a fact reaches a certain amount of so-called report votes, a group of users will be notified to check this fact. This group of users is randomly selected and they have the responsibility to validate the reported fact and/ or source. If they are not able to judge if a fact is correct or incorrect, they can inform a group of users which are expert in the field of where the source comes from. We propose a two step procedure with a randomly selected panel and an expert panel to limit the workload for the expert panel. In other words, the validation of facts in this methodology relies on the wisdom of the crowd. We realize that this methodology might be vulnerable for groupthink and strategic behavior, but we think that Wikipedia proves that the wisdom of the crowd works, if implemented correctly.

3.2 Moral acceptability

To survey the moral acceptability of the alternatives we use the Moral Foundations Questionnaire (MFQ) that was developed [8] based on the Moral Foundation Theory. The MFQ can be used to measure a broad range of moral concerns. The MFQ consists of two parts, one about moral relevance and the other one is about moral judgment. We intended to use the fifteen questions of the first part as an instrument to assess the moral acceptability of the proposed alternatives in the debates. We performed a small test to check the understandability of the questions. It turned out that the questions in their original form were hard to understand by the testers and did not fit the way we want to measure the alternatives. Therefore we decided to adjust the MFQ questions slightly to make them more applicable to our design of the debate process and understandable for the user. An example of this modification is the rephrasing the statement *Whether or not some people were treated differently than others* into the question: *Do you think that as a result of the alternative above: Someone is treated differently from others?* We realize that this impacts the validity of this instrument which means that research is needed to validate the modified questions. Since our prototype is merely a proof of concept we chose not to test this validity at this moment.

3.3 Social acceptance

As described in paragraph 2.3, the Social Choice theory takes the preferences of individuals into account, therefore we regard it as a suitable means to measure social acceptance. We studied several voting mechanisms that are being used in Social Choice Theory and chose to implement one to determine the social acceptance of the alternatives of the debates. These voting mechanisms are described in the next paragraph.

3.4 Voting mechanisms

Voting is a popular method to reach a joint decision based on aggregated preferences of individuals. One of the most used voting mechanisms in elections is the Schulze method which is used by Ubuntu, several Pirate Party political parties, OpenStack and LiquidFeedback [17]. This preferential voting method satisfies among others the criteria of anonymity, the Condorcet criterion and independence of clones [19]. Voters can list their preferences anonymously which is an important prerequisite for elections. The Condorcet criterion selects a single winner by majority rule in pairwise comparisons over every other candidates. Clone independence is a criterion that prevents certain types of strategic behavior in the voting process which means that it is impossible to be insincere about a voter's real preferences in order to secure a more favorable outcome. In the Schulze method every voter submits an ordered preference list for the candidates presented to the voter. All candidates are compared pairwise and a directed graph with the strongest path is created based on all votes and pair-wised comparisons. The output can be determined by looking which candidate defeated all other candidates and this one is declared the winner [17].

Next to the Schulze method we considered to implement the Ranked pairs algorithm, because this method is even more robust to strategic behavior [19] and it satisfies most of the same criteria as the Schulze method. Both are Condorcet methods, but they produce a different order of winners due to the fact that the Schulze algorithm reverses a larger majority than the Ranked Pairs algorithm for the majorities on which the two orders of finish disagree [17]. We found that there is less information available about the Ranked pairs algorithm than about the Schulze method. Ranked pairs is also harder to understand which negatively impacts the transparency of the voting mechanism. Therefore, we chose to implement the Schulze method in our prototype and used the PHP library of Julien Boudry that was available on GitHub [3]. We analyzed and tested the implementation of this algorithm with voting example to determine if the open-source algorithm was correct, which it turned out to be.

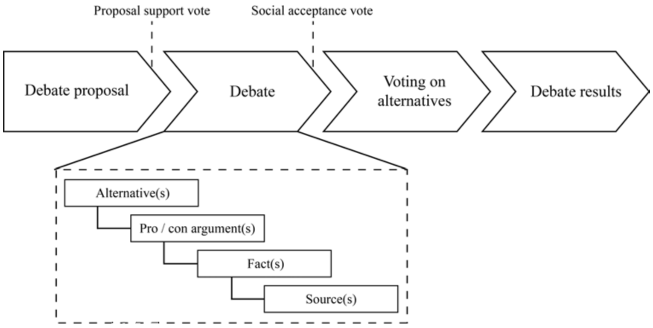
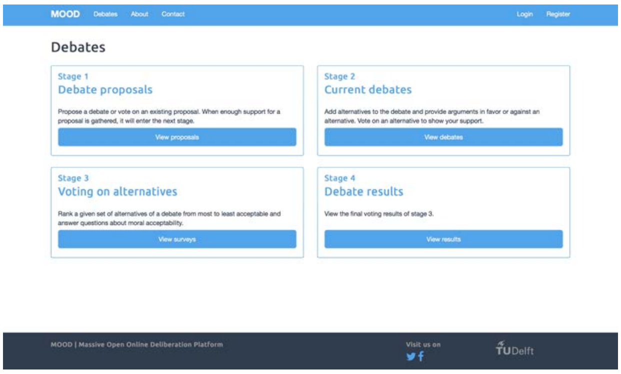


Figure 1. MOOD deliberation process



Debates

Stage 1
Debate proposals

Propose a debate or vote on an existing proposal. When enough support for a proposal is gathered, it will enter the next stage.

View proposals

Stage 2
Current debates

Add alternatives to the debate and provide arguments in favor or against an alternative. Vote on an alternative to show your support.

View debates

Stage 3
Voting on alternatives

Rank a given set of alternatives of a debate from most to least acceptable and answer questions about moral acceptability.

View surveys

Stage 4
Debate results

View the final voting results of stage 3.

View results

MOOD | Massive Open Online Deliberation Platform

Visit us on

TUDelft

Figure 2. Screenshot debate page

4 IMPLEMENTATION

In this section we describe the techniques we implemented in our prototype that we developed in the ten weeks of our Information Architecture design project course. We explain our choices for the framework we used and sketch our plan to test our application.

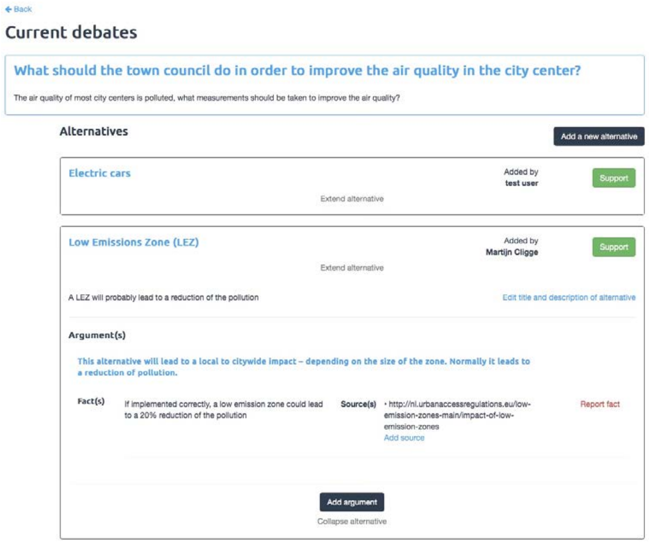
4.1 MOOD deliberation process

In our prototype we implemented the actual e-deliberation process based on the methods described in the previous section. We built a deliberation process consisting four stages: (1) proposal and initiation of a debate, (2) the actual debate in which user can cast votes to support an alternative, (3) the selection of alternatives via preference voting and measuring the moral acceptability of the alternatives and (4) reporting of the results. These stages are depicted in figure 1 which are translated to the application in the overview of the debate page in figure 2.

In stage one, a user can initiate a debate by submitting a proposal to the MOOD platform. This proposal needs to be described in a generic way and should preferably be posed as an open question. The initiator has four weeks to raise support for the debate and to reach a voting threshold. We set the threshold with an initial value of ten votes, but we will have to test if this value proves to be correct. The threshold procedure resembles the procedure for citizen initiatives in The Netherlands [5]. After reaching the voting threshold the proposal enters stage two of the debate. Once the threshold is reached, an initiator cannot withdraw his proposed debate, because this would mean that all aspects of a certain debate, like arguments, sources and facts, will be deleted and to our opinion valuable information will be lost.

In stage two the actual debate is held. Discussants can react to a debate by submitting alternatives which consist of pro- and con arguments (figure 3). It is also possible for users to add pro- or con arguments to an existing alternative. Arguments need to be substantiated by facts and sources to reference these facts to differentiate them from values. Although not built in our prototype yet, these facts will be validated by means of crowd-sourcing. The facts can be contested by other users and if a certain threshold is reached, the administrator will review the fact. If the fact is not valid then it will be marked in the database as rejected and will not be visible to the users. In a future version of the MOOD platform an expert panel will take over this task from the administrator to provide a more independent judgment of a contested fact. A debate will have a pre-set duration which is set by the initiator. In this stage, all users can vote to support an alternative. The five top alternatives will be selected and the debate will enter the next phase.

In the third stage of the debate, a voter can list his or her preferences of alternatives. The preferences are calculated by the Schulz voting mechanism. By this, the social acceptance of the alternatives in a debate is measured. After the voting, a list of alternatives is created ranking the alternatives that received the most votes. Next, the moral acceptability of the alternatives is surveyed with questions that are based on the MFQ for the selected alternatives. Per alternative seven questions will be asked to measure the ethical principles of *harm*, *fairness* and *authority*. The answers will be used to gain insight in the moral acceptability of the different alternatives in a debate.



Current debates

What should the town council do in order to improve the air quality in the city center?

The air quality of most city centers is polluted, what measurements should be taken to improve the air quality?

Alternatives

Electric cars — Added by **test user**

Low Emissions Zone (LEZ) — Added by **Martijn Cligge**

A LEZ will probably lead to a reduction of the pollution

Argument(s)

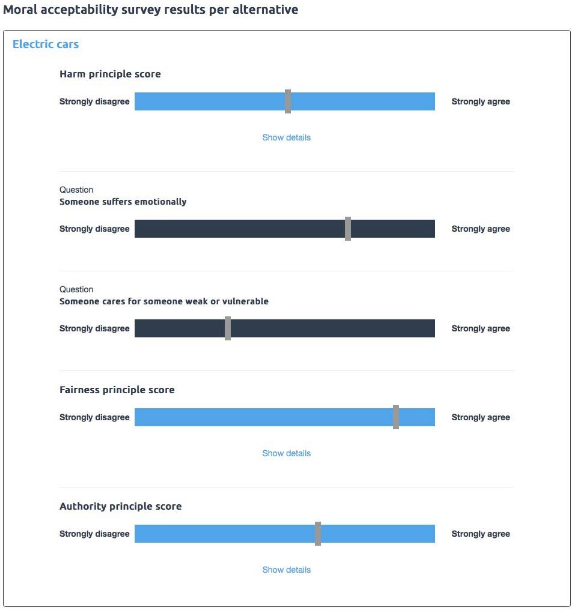
This alternative will lead to a local to citywide impact – depending on the size of the zone. Normally it leads to a reduction of pollution.

Fact(s): If implemented correctly, a low emission zone could lead to a 20% reduction of the pollution

Source(s): • http://nl.urbanaccessregulations.eu/low-emission-zones-main/impact-of-low-emission-zones

Figure 3. Screenshot alternatives page

In the fourth and final stage the social acceptance and moral acceptability results of the debate will be presented (figure 4). The results will be available to all users which will enhance the transparency of the debate.



Moral acceptability survey results per alternative

Electric cars

Harm principle score

Question
Someone suffers emotionally

Question
Someone cares for someone weak or vulnerable

Fairness principle score

Authority principle score

Figure 4. Screenshot results page

4.2 Framework

We chose an open-source framework to develop our prototype, because it is easily available and it enhances the transparency and traceability of our platform. We used the free open-source PHP framework *Laravel* to build the prototype. This framework is available on GitHub and can be used under the terms of a MIT license. According to their official website, it can be used to build elegant web applications that are *'...delivered at warp speed.'* [11]. It is developed via a Model-View-Controller (MVC) architecture. This is a category of software applications that consists of three interconnected parts that separate the internal representation of information from the way the information is presented to the user. The *Model* component handles the data, logic and rules of the application and stores the data it receives from the controller. The *View* shows the output of the application and generates new output when the model changes. The *Controller* accepts and converts the input into commands for the model and the view [15]. Laravel is one of the most popular PHP frameworks at this moment and includes features, such as a Composer architecture for Artisan, Laravel's Command Line Interface, Eloquent Object-Relational-Mapping (ORM) to put constraints on the relationships between database objects and Query builder to program queries automatically [14]. To create the database we used the open-source PHPMyAdmin software that handles MySQL queries for the web [12]. We used *bootstrap* to adjust the layout of the web application dynamically to the (mobile) device of the user. This free and open-source library is hosted on GitHub. Using bootstrap we aim to enhance the user experience for our prototype.

4.3 Testing

At the time that we are writing this paper we did not test our web application yet. Our first test will focus on the usability of our application. We will ask a small group of individuals (3-5 people) to walk through our application via scenario testing. The test scenario focuses on the e-deliberation process of our application. We ask our testers to follow this scenario to see if they understand the different steps in the process and to assess if the application is easy to use. The scenario starts by asking the user to make a new account and subsequently login with this account. After that, our testers will propose a new debate in the first stage of our deliberation process. Next, testers have to work themselves to the different stages, by adding alternatives, arguments, facts and sources in stage 2, by ranking the most social acceptable alternatives in stage 3, by filling in the survey on moral acceptability and by viewing the results in stage 4. We already prepared some debate issues in stage 2, like "No fast food should be sold in the University canteen, because it leads to obesity". We have designed two different kind of setups for our scenario. In the first setup, we will provide users with some explanation and a clear walk-trough description which describes every step in the scenario. In the second setup, we ask our testers to follow the same steps as in the first setup, but we give them very minimal explanation and no clear walk-trough description. We ask them to think out loud while performing the scenario with the second setup. The results of our test will be available after this paper is drafted, therefore these are not included in this document right now.

5 CONCLUSION AND DISCUSSION

In this paper we gave an overview of the e-deliberation process and existing platforms Liquidfeedback and Debatehub. We built a prototype that focuses on the one hand on combining the best of both platforms (by using parts of the debate page structure of Debatehub and by using parts of the deliberation process of Liquidfeedback) and on the other hand on aspects of both platforms that could be improved. Our design objective mainly focuses on the following aspects: (1) a method to differentiate and validate *facts* and *opinions*, and (2) a mechanism that supports both the *social acceptance* and the *moral acceptability* of debate outcomes. We researched these concepts more in depth and implemented several techniques to meet these aspects which resulted in a working prototype.

5.1 Limitations

Due to the little available development time, our prototype has several limitations. We focused our research on the topics of the differentiation between facts and values, social acceptance, moral acceptability and voting mechanisms. Time lacked to extensively study these topics and we realize that this scoping can lead to conformation bias, which means that we only used literature that substantiates our ideas and did not consider alternative views. The time constraint also affected the features of our prototype. One of the features that we did not manage to implement, is that of reputation score to distinguish between experts of certain discussion topics and regular users. This distinction is useful to create expert panels to validate the contesting of facts in the stage of the actual debate. Another feature we did not implement is an algorithm that creates a random panel to evaluate a contested fact. In the current application this task is performed by the administrator. Furthermore, a limitation is that we modified the MFQ questionnaire, but we did not study the effect of this instrument. Next to this, we chose to run the application on a small centralised server of the University which limits the amount of users that can simultaneously take part in the debate and impacts the scalability. To accommodate more users, a distributed or cloud server is needed to upscale the application in the future. Finally, we made a trade-off regarding the privacy of users and security of the platform. A limitation of our current design is that an administrator or auditor can trace a vote back to a user who casted it. Although, this violates the anonymity requirement of voting, this information is only visible for an administrator or auditor and not for any other user. More importantly, it enables full traceability, which contributes to more transparency and credibility via audits of the voting results. It is not possible for users to see how often is voted on alternatives in stage two to limit strategic behaviour which could occur when an alternative received many votes and people might want to vote on an alternative that is popular. Nevertheless, strategic behaviour could occur when users register with multiple e-mail addresses in order to be able to cast more votes. We have not been able to implement a counter measure for this in our prototype.

5.2 Future research

These limitations lead to recommendations for future work. We did not manage to study the revised MFQ questions. Its validity and applicability to measure moral acceptability in debates should be researched. We also recommend to extent the literature study for mechanisms to differentiate between facts and values, for social acceptance, moral acceptability and voting mechanisms and find alternate views on these topics. An extension of the voting stage would also be a possible addition to a future version of the application. Adding a second round of preferential voting, after the publication of the results of the moral acceptability survey, would allow people to change their mind and vote for a different alternative than they did the first time. We did not manage to include all features in our prototype that we described in our list of requirements. A mechanism for crowdsourcing should be added to categorize the facts that are added to the debate. Next to this, it should be possible to forward a contested fact to an expert panel for an independent judgment. Also, tracking the reputation score of users should be added as a feature to our prototype. These features are crucial to develop the MOOD platform into a more mature application. Additionally, sentiment analysis on content provided by the users could be implemented in the MOOD platform to sense the atmosphere of the debate. On the other hand machine learning techniques can also be used to support the MOOD platform. For example validate facts by means of crowd-sourcing applications or Watson APIs.

ACKNOWLEDGEMENTS

We would like to thank Dirk Helbing for his thoughts and input to our prototype.

REFERENCES

[1] Getachew Assefa and Björn Frostell, ‘Social sustainability and social acceptance in technology assessment: A case study of energy technologies’, *Technology in Society*, **29**(1), 63–78, (2007).

[2] Jan Behrens, Axel Kistner, Andreas Nitsche, Björn Swierczek, and Björn Swierczek, *The principles of LiquidFeedback*, Interaktive Demokratie e. V., 2014.

[3] Julien Boudry. Condorcet. https://github.com/julien-boudry/Condorcet. Retrieved at: 26-05-2016.

[4] Taya R Cohen and Lily Morse, ‘Moral character: What it is and what it does’, *Research in Organizational Behavior*, **34**, 43–61, (2014).

[5] Tweede Kamer der Staten Generaal n.d. Burgerinitiatief. https://www.tweedekamer.nl/kamerleden/commissies/verz/ burgerinitiatieven. Retrieved at: 07-07-2016.

[6] Naomi Ellemers, Stefano Pagliaro, and Manuela Barreto, ‘Morality and behavioural regulation in groups: A social identity approach’, *European Review of Social Psychology*, **24**(1), 160–193, (2013).

[7] James D Fearon, ‘Deliberation as discussion’, *Deliberative democracy*, **44**, 56, (1998).

[8] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto, ‘Mapping the moral domain.’, *Journal of personality and social psychology*, **101**(2), 366, (2011).

[9] Amy Gutmann and Dennis Thompson, *Democracy and disagreement*, Harvard University Press, 2009.

[10] Nicole MA Huijts, Cees JH Midden, and Anneloes L Meijnders, ‘Social acceptance of carbon dioxide storage’, *Energy policy*, **35**(5), 2780–2789, (2007).

[11] Laravel n.d. https://laravel.com/. Retrieved at: 26-05-2016.

[12] PHPMyAdmin n.d. http://www.phpmyadmin.net/. Retrieved at: 26-05-2016.

[13] Wikipedia n.d. Descriptive statement. https://en.wikipedia.org/wiki/ Positive statement. Retrieved at: 07-07-2016.

[14] Wikipedia n.d. Laravel. https://en.wikipedia.org/wiki/Laravel. Retrieved at: 26-05-2016.

[15] Wikipedia n.d. Model-view-controller. https://en.wikipedia.org /wiki/Modelviewcontroller. Retrieved at: 26-05-2016.

[16] Wikipedia n.d. Normative statement. https://en.wikipedia.org/wiki/ Normative statement. Retrieved at: 07-07-2016.

[17] Wikipedia n.d. Schulze method. https://en.wikipedia.org/wiki/schulze method. Retrieved at: 26-05-2016.

[18] Wikipedia n.d. Wikipedia:editorial oversight and control. https://en.wikipedia.org/wiki/Wikipedia:Editorial oversight and control. Retrieved at: 26-05-2016.

[19] David C Parkes and Lirong Xia, ‘A complexity-of-strategic-behavior comparison between schulze’s rule and ranked pairs’, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI’12)*. American Association for Artificial Intelligence, (2012).

[20] Rebecca J Romsdahl, ‘Political deliberation and e-participation in policy-making’, *CLCWeb: Comparative Literature and Culture*, **7**(2), 7, (2005).

[21] Jeroen van den Hoven and Virginia Dignum, ‘Moods: Massive open online deliberation’. draft.

Ethics in the design of automated vehicles: the AVEthics project

Ebru DOGAN[a,1], Raja CHATILA[b], Stéphane CHAUVIER[c], Katherine EVANS[ac], Petria HADJIXENOPHONTOS[ab], and Jérôme PERRIN[d]

[a] *Institut VEDECOM, Versailles, France*
[b] *Sorbonne Universités, Université Pierre et Marie Curie,CNRS, Institute for Intelligent Systems and Robotics (ISIR)*
[c] *Sorbonne Universités, Université Paris-Sorbonne, Faculty of Philosophy*
[d] *Renault, Technocentre, Guyancourt, France*

Abstract. Automated vehicle (AV) as a social agent in a dynamic traffic environment mixed with other road users, will encounter risk situations compelling it to make decisions in complex dilemmas. This paper presents the AVEthics (Ethics policy for Automated Vehicles) project. AVEthics aims to provide a framework for an ethics policy for the artificial intelligence of an AV in order to regulate its interactions with other road users. First, we will specify the kind of (artificial) ethics that can be applied to AV, including its moral principles, values and weighing rules with respect to human ethics and ontology. Second, we will implement this artificial ethics by means of a serious game in order to test interactions in dilemma situations. Third, we will evaluate the acceptability of the ethics principles proposed for an AV applied to simulated use cases. The outcomes of the project are expected to improve the operational safety design of an AV and render it acceptable for the end-user.

Keywords. Ethics, artificial intelligence, robotics, automated vehicle, artificial moral agents

1. Introduction

Technological developments in sensors and wireless communication facilitate the development of sophisticated advanced driving assistance systems. Several subtasks of the driving task, such as lateral control and longitudinal control are now handled by the vehicle. The human driver is left more and more out of the control loop of the vehicle as the level of automation increases and the vehicle becomes an autonomous agent. A deployment of a fully automated vehicle (AV) in all contexts, however, is expected to take a few decades. Then an AV would become a social agent taking decisions to regulate its interactions with other road users and static objects in a mixed traffic environment. Some situations would involve complex decision making when life hazard is involved. Currently, an AV does not have a consensual minimal risk state, nor a crash optimization strategy. In fact, the decision-making architecture consists of a set of rules, mostly the Highway Code, applied by a programmer. Given the difficulty of predicting the behavior of dynamic objects in the traffic environment, there would be no way to completely avoid

[1] Corresponding Author.

a collision ("inevitable collision state") and the aim would be to minimize risks and damages [8]. Since the risk cannot be avoided, the decision turns into an ethical one: there will not be one "good" solution and the decision will involve a trade-off between interests of different parties in a given context [12]. An AV does not have a decisional autonomy, that is, "the capacity of reasoning on the perception and action in order to make non-trivial choices" [3, p.15]. Nor does it have a sense of ethics. Nonetheless, it would have to make real time decisions of risk distribution in ethical dilemmas involving high uncertainty. Although this issue relates to the general domain of "robot ethics" [1], the case of AV has specific features: i) open environment (public roads), ii) interaction with many different social agents, iii) involvement of many stakeholders of the road mobility system (car makers, insurance companies, public authorities, etc), and iv) entrust of the safety of the car occupants for the robot.

2. AVEthics Project

Figure 1 depicts a dilemma situation on a road that anyone can encounter. In such complex dynamic situations human drivers report that, even though they can explain their decision making processes once the situation is over, they do not reflect on the same terms while the situation is taking place [11]. Thus, human reaction in a dilemma situation is not pre-calculated; it is a split-second reaction [12], whereas an AV's will have prescribed decision algorithms to control the vehicle. In the situation depicted in Figure 1, no matter how the AV decides to manage a conflict, someone might be harmed.

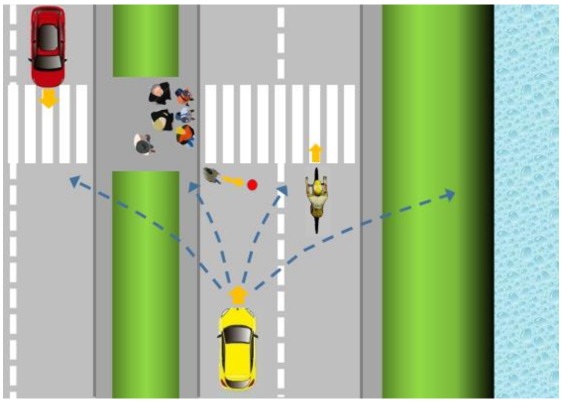


Figure 1. Sample use case

The current paper aims to present the AVEthics project (Ethics policy for Automated Vehicles). Its overarching goal is to provide a framework for an ethics policy for the artificial intelligence of an AV in order to regulate its interactions with other road users. This issue will be treated in three parts. First, we will specify the kind of (artificial) ethics that can be applied to AV, including its moral principles, values and weighing rules with respect to human ethics and ontology. Second, we will implement this artificial ethics numerically by means of a serious game in order to test interactions in dilemma situations. Third, we will evaluate the acceptability of the ethics principals proposed for an AV applied to simulated use cases.

3. Philosophy: from theory to casuistry

A common approach in robot ethics is to transfer *our* own way of reasoning on moral issues to artificial agents, creating a blueprint of our moral reasoning in robots, according to deontological or consequentialist theories (e.g. Kantian ethics, utilitarianism or virtue ethics). One of the problems of this approach is the arbitrariness of the choice of an ethical theory: why should we prefer a Kantian AV to a utilitarian AV? A more important problem is the lack of real operationalization of the capacities that enables humans to think morally.

The AVEthics project approaches the AV as a "*modular artificial moral agent" (MAMA*, [4, in press]*)*. Accordingly, an AV pursues its goals based on its artificial intelligence, and it is modular in the sense that it is not universal, but rather specialized to cover a limited set of goals. The artificial ethics endowed to a MAMA refers literally to the code that should guarantee that the MAMA's behavior should be sensitive to the rights, interests, and needs of all the entities that could be affected by its decisions. The challenge is the lack of consensus on *which capacities* to implement in the MAMA for it to successfully make ethical decisions.

One way to tackle this issue is to focus on the essential needs of artificial ethics, rather than trying to implement human morality into the robot. Human drivers' decisions are determined by prioritization of goals, and valences[2] of different action possibilities afforded by their perceptual environment, which are rarely calculated in advance [7]. Following this notion, the morality of a MAMA would mainly require sensitivity to a limited set of values and principles that would be morally salient in specific encountered situations depending on its functionalities (case-based approach), rather than general principles applicable to a great variety of situations (principle-based approach). However, the literature on the casuistic approach to ethical decisions (in robotics) is relatively limited [see 13, 9, and 1 for examples]. Moreover, it seems difficult to dissociate cases from principles. *How can an AV decide to crash into a dog instead of a cyclist, if the AV does not know the rule that human life has a higher value than the life of a dog?*

One way to favor a case-based over a principle-based approach is to dissociate deeds and valences. Human morality is rooted in the calibration of the content of the world that we experience [14], and our morality is shaped by situations and experiences we are confronted with. Hence, perception of the environment, valence entailed by the entities in the environment, and goal-directed behavior incorporating the notion of valence become common in human morality and artificial ethics.

In the philosophy part of the AVEthics project, we will argue that 1) an artificial ethics requires representation and categorization of the morally relevant entities in order to define its "ontology", which is a moral issue *per se*, 2) an awareness of different entities in the traffic environment could be implemented by assigning to each a numerical value that would be taken into account by the AV control algorithms, and 3) a special "self" value would be added, for an AV carrying humans may not share an ethics of self-sacrifice.

[2] Gibson and Crook (1938) use **"valence"** akin to hazard weight of a potential action in a given traffic situation.

4. Robotics: development and experimentation

An AV has limited decision-making capacity because its situation awareness is focused on the actual perceived situation. An AV is equipped with sensors, such as radars, lasers, and cameras, which provide it with 360°-vision. Perception, planning, and control algorithms of the vehicle enable it to move in a complex environment. Nonetheless, the information available to the vehicle by its sensors, how the vehicle interprets this information, and the way it uses the information for decision-making are all completely different from those of a human driver. Furthermore, an AV cannot cope with situations that were not anticipated and taken into account by the programmer (even if it includes a learning approach). Overall, an AV's decision-making is imperfect and uncertain.

The decision-making architecture of an autonomous system consists of three levels with decreasing decisional autonomy [6]: the higher "strategic" level manages goals, beliefs, and plans; the intermediate "tactical" level supervises the system; the lower "operational" level carries out the actions, e.g. longitudinal and lateral control for an AV. One of the challenges is the traceability of the decisions of an artificial intelligence agent. Given the possibility of liability and insurance problems, AV stakeholders would be keen on traceability. Implementation of the decision is another challenge: decisions based on the AV "ethical" principles should be representable in the control algorithms of the vehicle as tangible components such as speed, brake, time headway (time between the ego vehicle and a lead vehicle), and steering wheel movements. Hence, we need to test the feasibility of the ethical decisions of an AV.

In the previous section, we advocated categorization of the perceived entities and assignment of valences to these entities in an AV's ethical decision-making. For this end, the AV should be, first, able to quantify the reliability of the information acquired by its sensors. Only then can it rely on this information in order to distinguish among the entities in its environment and to categorize them. The categorization will determine the valence assignment and the action plan of the AV depending on the ethical theory being tested. Assuming that the sensor data is of good quality and reliable, this process has two sources of uncertainty. *The first uncertainty is related to the categorization* of entities, which carries a probabilistic confidence value. *The second uncertainty is related to the action implementation*: the course of the action planed by the AV based on an ethical decision is also probabilistic. Hence, decision-making should account for the uncertainties in categorization and action. *How can we handle these two uncertainties?*

In the robotics part of the AVEthics project, we will 1) test different approaches, such as fuzzy, belief-based or Bayesian, in order to tackle uncertain categorization, 2) investigate the best course of action for an AV, considering uncertain perception and non-deterministic actions, and 3) study optimal decisions that could be taken jointly by the AV and other agents in its surroundings (vehicles, pedestrians, and infrastructure). We will also develop a test tool, a serious game interface that can be connected to a driving simulator, so that we can apply the model of artificial ethics to the use cases and test this with human drivers.

5. Psychology: public acceptability

To assume that an AV would be acceptable because it would increase safety is not necessarily valid. Human ethical decision making is often seen as a mix of emotions and reason. The end-user might consider the overall collective safety gain to be insufficient to merit taking certain individual risks, even if they would be rare cases. Indeed, research in social psychology indicates that socio-cognitive constructs such as values, contextual features, and trust have a notable effect on acceptability [15].

People who do not have sufficient knowledge on complex, new, and most of the time, controversial technologies, such as AVs, rely on their trust in the main stakeholders [16]. Competence-based trust (i.e. trust in a stakeholder's experience and expertise) is rather straightforward: positive information about a stakeholder's expertise is associated with higher trust and acceptability (and vice versa). Integrity-based trust (i.e. trust in a stakeholder's honesty, openness, and concern), on the other hand, is more complicated: when people perceive a stakeholder as biased and dishonest, they *go counter to the organizational position*. More precisely, if the organization is a proponent of a new technology, people are negative about the same technology [18]. In fact, when the issue is of high moral importance for the individual[3], the objective information about the competence loses its persuasive power [5]. One can even denounce the legitimacy of the decisions of a stakeholder when morality becomes salient [17]. *The relationship between trust and acceptability is thus sensitive to the moral importance of the issue*.

Another concept related to trust and acceptability is values. People are more likely to trust involved parties if they share values similar to their own [16]. Information on value similarity also influences integrity-based trust, but not competence-based trust [5]. Two main clusters of values have been identified: self-transcendence values, which are concerned with collective outcomes, and self-enhancement values, which are concerned with individual interest. These two may be in conflict in controversial issues. For instance, scenarios which involve taking risks with people on-board in an AV would be in line with societal outcomes, but contradictory to individual outcomes. Perlaviciute & Steg (2014) propose that people's tendency to adopt deontological or consequentialist reasoning might depend on people's values.

What are people's preferences in situations of ethical dilemma situations? Recent survey research revealed that drivers had positive evaluations about a utilitarian AV that is programmed to minimize the casualty in unavoidable accidents, including the self-sacrifice scenarios [2]. Utilitarian thinking is observed in public policy evaluations as well. People's ethical preferences for road safety policies changed as a function of the value of the age and the responsibility/vulnerability of the victim: protection of young (*vs* elderly) road users and pedestrians (*vs* drivers) is favored [10]. However, findings in the neuroscience of moral decision making hint at the complexity of this process.

In the psychology part of the AVEthics project, we will 1) test a model of acceptability integrating people's trust in the competence and integrity of the stakeholders and the value similarity with the stakeholders, and 2) investigate public acceptability of different ethical principles for an AV decision making by using the game interface mentioned above, as well as end user surveys. We will also collect stakeholders' acceptability judgments.

6. Conclusion

The ethics of automated vehicles is becoming a major issue from legal, social, and vehicle control perspectives. We acknowledge that the AV will have to make decisions that might eventually harm an agent and that these decisions should not contradict the

[3] We presume that a harmful decision of an AV is of high moral importance for the end user of this technology

interests of the end users or the principal stakeholders. An ethics policy for automated vehicles is a vast subject, and AVEthics is only the beginning of a long path. The expected outcomes of AVEthics are i) an initial proposition of ethical principles for an AV, ii) a software system and interface to apply these principles to different use cases, and iii) end user's acceptability judgments of the proposed ethical principles and the following action plans. This should contribute to improvement of the operational safety design of an AV and render it acceptable for end-users and stakeholders of the mobility system.

References

[1] Arkin, R.C. *Governing lethal behavior in autonomous robots*. Chapman & Hall, NY, 2009

[2] Bonnefon, J.F., Shariff, A., & Rahwan, I.. *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?* http://arxiv.org/pdf/1510.03346v1.pdf. Accessed on 28/10/2015.

[3] CERNA.*"Ethique de la recherche en robotique"*, rapport n°1, Novembre 2014. http://cerna-ethics-allistene.org/digitalAssets/38/38704_Avis_robotique_livret.pdf

[4] Chauvier, S. L'éthique artificielle. In press for *L'Encyclopédie philosophique* (Ed. M. Kristanek)

[5] Earle, T.C. & Siegrist, M. On the relation between trust and fairness in environmental risk management. *Risk Analysis,* 28 (2008), 1395-1413.

[6] Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. Ethical choice in unforeseen circumstances. In *Towards Autonomous Robotic Systems*, A. Natraj, S. Cameron, C. Melhuish, M. Witkwoski (eds.), 14th Annual Conference, Oxford, UK, August 28-30, (2013) 433-445.

[7] Gibson, J.J & Crook, L.E. A theoretical field-analysis of automobile-driving. *The American Journal of Psychology,* 51 (1938), 453-471.

[8] Goodall, N. Ethical Decision Making During Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424 (2014), 58–65.

[9] Guarini, M. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21 (2006), 22–28.

[10] Johansson-Stenmann, O. & Martinson, P. Are some lives more valuable? An ethical preferences approach. *Journal of Health Economics*, 27 (2008), 739-752.

[11] Klein, G., Orasanu, J., Calderwood, R., & Zsmbok, C.E. *Decision making in actions: Models and methods*. Ablex, Norwood, 1993.

[12] Lin, P. Why ethics matter for autonomous cars? In *Autonomes fahren* (M. Mauer, J.C. Gerdes, B. Lenz, & H. Winner Eds.). Springer, Berlin, 2015.

[13] McLaren, B. Computational models of ethical reasoning: Challenges, initial steps and future directions. *IEEE Intelligent Systems*, 21 (2006), 29–37.

[14] Parfit, D. *On What Matters.* Oxford University Press, Oxford, 2011.

[15] Perleviciute, G. & Steg, L. Contextual and psychological factors shaping evaluations and acceptability of energy alternatives: integrated review and research agenda. *Renewable and Sustainable Energy Reviews,* 35 (2014), 361-381.

[16] Siegrist, M. & Cvetkovich, G. Perception of hazards: the role of social trust and knowledge. *Risk Analysis*, 20 (2000), 713–719.

[17] Skitka, L.J., Bauman, C.W., & Lythe, B.L. Limits of legitimacy: Moral and religious convictions as constraints on deference to authority. *Journal of Personality and Social Psychology,* 97 (2009), 567-578.

[18] Terwel, B.W., Harinck, F., Ellemers, N., & Daamen, D.D.L. Competence-based and integrity-based trust as predictors of acceptance of carbon dioxide capture storage (CCS). *Risk Analysis,* 29 (2009), 1129-1140.

Processes of Reminding and Requesting in Supporting People with Special Needs: Human Practices as Basis for Modeling a Virtual Assistant?

Antje Amrhein[1] and **Katharina Cyra**[1] and **Karola Pitsch**[1]

Abstract. The text reports findings of a case study based on the investigation of reminding activities and request practices in the specific context of supported living. These activities turn out to be highly *adaptive processes* that are embedded in complex assistive networks. The *process of reminding and requesting* represents a central practice deployed by the assistive institutional and social environment. It suggests to provide a consistent structure that meets individual needs in everyday life of cognitively impaired people. In the light of the development and engineering of assistive technologies we discuss if and how human practices could serve as a basis for modeling an Embodied Conversational Agent (ECA) based assistive system for cognitively impaired people with respect to the adherence of their autonomy.

1 INTRODUCTION

People with cognitive impairments as well as elderly people require special assistance in managing their daily routines like household activities or managing the everyday structures when living independently. Cognitive or physical challenges often affect or lead to a decrease of the quality of life. Hence, maintaining an autonomous life in a familiar social environment and home for as long as possible has become a central issue in today's societies [1].

Research on technical assistive systems strives to suggest solutions for this social challenge, e.g., in the realm of Ambient Assisted Living and Social Robotics. To this end, multimodal dialogue systems represented by Embodied Conversational Agents seem particularly suited, as they can be easily integrated in private homes using modern TV sets, allowing for intuitive human-machine interfaces, using means of natural communication when entering and managing appointments and being reminded of individual tasks or events [2].

The question of autonomy arises when considering the integration of an assistive technology to support independent living and in the setting of supported living with distributed actions. Results of ethnographic research in an institution of supported living for people with cognitive impairments, i.e. people with special needs in independent living, presented in this study, reveal practices of reminding and requesting as essential to preserve well-structured everyday routines. Besides the moment of acute reminders, the complex *process of reminding and requesting* practices that precedes the actual reminder is relevant to form an understandable request-reminder and its accomplishment. These processes are closely interwoven and coordinated with an assistive social and institutional network. Set against this background the integration of an assistive technology into already existing assistive networks carries a strong ethical issue with respect to the preservation of the individual autonomy [3].

This study shows how ethnographic research serves as a valid approach to user centered design respecting the Human Value Approach [4] and to gain deeper insights into the actual needs, practices, daily routines and competences of the potential users. The investigation addresses the following questions:

A) How could the activities of reminding and the actual requests, i.e. acute reminders, in every day practice be described?

B) How are reminders established in a meaningful way, so that their intent and consequences are understood and followed by meaningful activities?

(C) How could the reminding and requesting practices be implemented into an assistive technology and how could an ECA as a daily-assistant be integrated into the social and institutional network that encompasses people with special needs?

2 ETHICS AND TIME MANAGEMENT

2.1 Ethical dimensions of assistive technologies

Based on sociological analyses of human activities and technology, Rammert speaks of “distributed action[s]” [5: 18] and “distributed agency” [5: 5] and describes them as a multiplicity of actions which are distributed over temporal and factual dimensions. In this context technical engineers, have to consider how system influences human relations, hierarchies, competences and the division of work. Winner stresses that “The things we call ‘technologies’ are ways of building order in our world.” [6: 127] and so, they shape society, individuals and their actions. Thus, the design of technical systems always reflects implicit or explicit values and can never be neutral. While the approach of Value Sensitive Design suggests to integrate the needs of human users and values [7] the Human Value Approach [4] goes one step further with the demand not only to consider the users' needs but also to apply the idea of Human Values to the technologies themselves and the development process and the disciplines involved in the design process. Human values are meant to be "ideas we all hold about what is desirable in different situations, societies and cultural

[1] Communication Studies Department, University of Duisburg-Essen, Germany, email: {antje.amrhein, katharina.cyra, karola.pitsch}@uni-due.de

contexts" [4: 35]. As these values differ individually it has to be made transparent in the design process of technological and especially assistive systems which of them affect technology.

In the design process of technical systems ethical issues have to be considered not only from the individual perspective but also from an institutional and social viewpoint. The model for ethical evaluation of socio-technical arrangements (MEESTAR) [3] is one approach that is not only taking users' needs into account but also the ethical evaluation of a technical system. MEESTAR suggests seven dimensions for the ethical evaluation of technical artifacts: care, autonomy, security, justice, privacy, participation and self-image. These dimensions are applied on an individual, organizational and social viewpoint to systematically carve out ethical issues and possible areas of conflict.

In the area of assistive living, especially when focusing on assistance in the field of time management and support of temporal orientation by an ECA-based assistant, the dimension of autonomy plays an essential role. Technical artifacts that remind, or request users to perform a task or to keep an appointment, and afterwards check the accomplishment of a task, raise the question of agency and autonomy on the one hand, but also contribute to various levels of individual security and participation.

2.2 Technology & time-management

Human Computer Interaction (HCI) studies on time management support and calendars show how reminders can be designed as requests and argue to design them in a multimodal way to be effective, usable and accessible for a diverse user group [8] [9]. However, the authors stress the right application of reminders to work properly, which includes both the timing and the form of the reminder. Going beyond these considerations, we will show that not only timing and form of the reminder need consideration when modeling an assistive technical system, but also the right level of (increasing) urgency and the need for adaptation to social [10] and interactional circumstances over the course of time.

Though those HCI studies refer to context they do not show the dependencies and fine-grained coordinative practices of an assistive network [10] in the domain of time management. Our aim is to trace how reminders emerge in the context of everyday activities within a highly personalized and complex support network and to raise the question of whether and how a technical assistive system could be integrated into the complex structures.

2.3 Requests in care & supported living

Requests as a subject of research have a long tradition within linguistics and there are several attempts to describe and define requests [11] [12] [13] etc. However, these approaches mostly describe requests from a speaker's perspective not taking into account the interactional situatedness and procedures of production and narrowing requests down to singular utterances. Conversation Analysis (CA) considers the sequential procedures of interactions and reveals insights into the production processes of reminding and requesting and what speakers consider when producing them. Studies from various settings (care, medical, HCI etc.) show that syntactical forms of requests hint at the speaker's understanding of the recipient's capability to accomplish the request. Yet the syntactical form itself also reflects the entitlement of the speaker to place a request [14] [15] [16]. These findings can be applied to the modeling of technical systems regarding display of availability, recipiency and acknowledgement [17].

In sum, linguistics, CA and HRI (Human Robot Interaction) research widely defines requests as represented by singular verbal utterances even though, there are hints at the influence of contextual, interactional and sequential circumstances for the production a singular utterance. Besides, especially research in care settings has primarily focused on requests made by the care-receiving party in face-to-face interaction. Our aim is to expand this perspective by describing requests in a broader sense that takes not only the sequential structure of interactions into account, but also the social and institutional perspective. To provide valid statements for the implications for an ECA-based assistant [2] [18] we examine the requests made by the support worker. This perspective encompasses a highly ethical issue by asking how requesting practices can be embedded into a technical system without compromising the autonomy of the client.

3 STUDY & METHOD

3.1 Ethnographic Research & Data

The research is based on focused ethnography [19] in an institution of supported living based in Germany where people with cognitive impairments get individual in- or outpatient care as required. The research was directed at gaining insights into individual, institutional and social structures, that emerge from everyday activities and routines. We especially explored the actual routines, competences and strategies of people with special needs (clients) in independent living and focusing on the needs of assistance. The institution is located in the sector of integration aid (Fig. 1.) which is organized on two levels: a local 24-hour attendance service and individual outreach work provided by support workers.

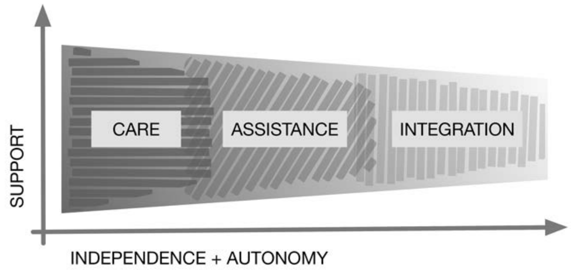


Figure 1. Support levels within in- and outpatient care

The inspection of individual (outreach work) and institutional settings (attendance service) revealed that there is a differentiation between required support levels depending on independence and autonomy of people with special needs. There are three merging levels of support: care, assistance and integration (Fig. 1). Clients with special needs in the care area are supported exclusively within inpatient care and intensive social and physical support. Clients with special needs in the area of assistance live either in in- or outpatient care with individually adjusted support depending on the area of support. On the support level of integration the clients with special needs are living in outpatient care mostly at their own homes and work in so-called sheltered workshops.

To get a comprehensive overview of what assistive practices actually look like, how they are communicated and coordinated within the assistive social and institutional network of the client, the ethnographic study took place in different areas and settings within the institution. As the integration aid is based on two forms of assistance, we first focused on the central office of the attendance service as the “center of coordination” [20] in the supported living institution. Here we examined how information is shared and transferred, appointments are made and tasks are coordinated. The second focus was on a more intimate setting of regular, mostly one-on-one, weekly assistance meetings with the support worker and the client. We accompanied three client-support worker-tandems repeatedly within a 4-month period. These meetings normally take place at the client’s home and are part of the individual outreach work that among others, involve planning activities, post-processing of past events and assisted time management to provide temporal orientation and structure. Further areas of the ethnographic research focused on a weekly communal breakfast organized by the institution, everyday routines such as assisted grocery shopping or leisure activities (e.g. multimedia classes).

Following the principles of focused ethnography [19] the data was collected during repeated stays in the field and contains a variety of data types, that are ethnographic field notes, observation protocols, documents and photos gathered exclusively during participatory observation in the central office. Further audio- and video-data was recorded during assistance meetings, the communal breakfast, grocery shopping and leisure activities.

The fine-grained analysis of video data is based on CA and provides access to the understanding of micro-sequential processes in interactions [16]. This approach relies on repeated analysis of recorded data and detailed multimodal annotations of relevant modalities of the interaction (e.g. verbal, gaze etc.) to sequentially reconstruct the process of reminding and the interaction order with regard to the temporal interrelationship of modalities. The verbal transcripts are based on the conventions of second edition of the German Conversation Analytical Transcript System (GAT2) [21].

4 REMINDING AS A PROCESS

The ethnographic study reveals the activity of reminding within assisted time management in the context of an assistive living institution as a *process of reminding.* It is framed and coordinated by a client's assistive network [10], encompassing both formal institutional assistance and informal assistance. The concept of the *process of reminding* contains essential social, institutional and conversational practices and planning activities (Fig. 2.).

These planning activities are closely connected to the individual needs and competences of the client and are embedded into the organizational structure of the assistive network. The data show that planning activities usually start with an appointment registration that can be initiated by the client herself/himself, by her/his assistive network or external sources. Either way, this registration is communicated and coordinated with all involved parties. The joint planning of an appointment allows a maximum of transparency and agency for both, client and support worker. Joint planning, that is part of the regular assistance meetings, is one aspect of legitimization of the support worker to apply the successive steps of the reminding process.

We identified different steps that evolve as a *process of reminding* after the initial appointment registration. The core process consists of two essential practices applied by the assistive social and institutional network: successive reminders that have an instructive character and acute reminders that function as requests.

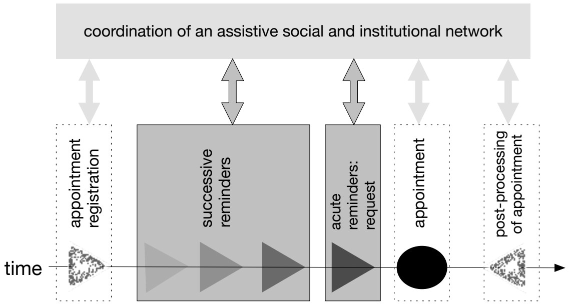

coordination of an assistive social and institutional network
appointment registration
successive reminders
acute reminders: request
appointment
post-processing of appointment
time

Figure 2. Process of Reminding framed by an assistive network

Successive reminders appear to have a twofold function for the clients with special need in temporal orientation: in the long term they provide reliability regarding planning activities and support temporal orientation on the one hand and on the other hand help to anticipate acute reminders. Acute reminders have a request character due to their temporal proximity to appointments. When comparing this with the findings regarding requests in care settings we see a contrast in the performance of a request and the form not an isolated utterance, but embedded in a request context. The concept model of *reminding as a process* finishes with the actual appointment or optional post-processing.

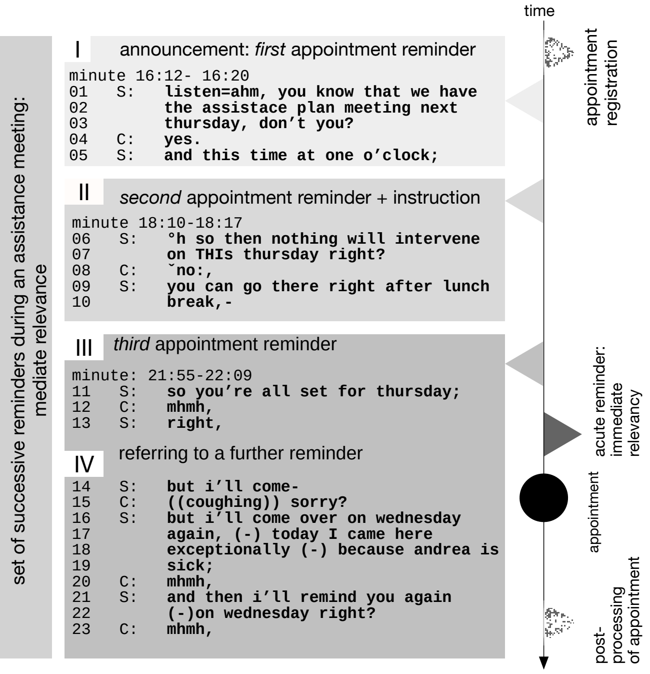
The *process of reminding* relies on highly complex and adaptive assistance networks, involving official institutional staff as well as an informal social environment involving family, friends and colleagues. This highly personalized flexible support network is being formed to respect and support the participants’ competences and capabilities.

5 PERSONAL SUCCESSIVE REMINDERS

The following case study focusses on the process of successive reminding during an assistance meeting and is temporally located after the appointment registration and before the acute reminder (see chapter 6). The analyzed segment is a record of an assistance meeting where a support worker (S) and a client (C) discuss upcoming and past issues at C's home. C has no temporal orientation and therefore depends on explicit and recurrent reminders and requests. S's successive reminding strategies are produced in different formats and temporal stages during the assistance meeting with C (Fig. 3 I-IV).

(a) Announcement: first appointment reminder: After discussing recent events S starts the first announcement on reminding C of an upcoming appointment for an assistance plan meeting in three days on a Thursday at one o’clock. The last assistance plan meeting was cancelled and the appointment has now been rescheduled. This appointment involves not only C and S but also C's legal representative. As there is an institutional network engaged there is the need for coordination. Another rescheduling or cancelling of the appointment due to a possible non-appearance of C would imply additional organizational expenditure for the assistive network. So, C's punctual appearance has an increased significance in this context. A successive reminding process is central during assistance meetings and an essential key to assure a punctual appearance to appointments.

With his question (Fig. 3 I 01-03) S is reassuring and checking that C is already aware of the upcoming appointment for the assistance plan meeting. C confirms with ***yes***. After the positive confirmation of C, S names the time. By using a conjunction and a temporal adverb he marks the time ***and this time at one o’clock*** (05) as deviating from the norm. After a short sequence in which C explains why she couldn’t make it to the appointment last time S formulates a second appointment reminder.


time
set of successive reminders during an assistance meeting: mediate relevance
I announcement: *first* appointment reminder
II *second* appointment reminder + instruction
III *third* appointment reminder
IV referring to a further reminder
appointment registration
acute reminder: immediate relevancy
appointment
post-processing of appointment

```
minute 16:12- 16:20
01   S:   listen=ahm, you know that we have
02        the assistace plan meeting next
03        thursday, don’t you?
04   C:   yes.
05   S:   and this time at one o’clock;

minute 18:10-18:17
06   S:   °h so then nothing will intervene
07        on THIs thursday right?
08   C:   ˘no:,
09   S:   you can go there right after lunch
10        break,-

minute: 21:55-22:09
11   S:   so you’re all set for thursday;
12   C:   mhmh,
13   S:   right,

14   S:   but i’ll come-
15   C:   ((coughing)) sorry?
16   S:   but i’ll come over on wednesday
17        again, (-) today I came here
18        exceptionally (-) because andrea is
19        sick;
20   C:   mhmh,
21   S:   and then i’ll remind you again
22        (-)on wednesday right?
23   C:   mhmh,
```

Figure 3. Personal Successive Reminders translated from German

(b) Second appointment reminder + instruction: In the second stage (Fig. 3 II 06-10) S formulates another question to reassure that C will keep the appointment (II 06-07) and adds an additional instruction by providing practical guidance so C can manage to get to the appointment on time. By advising her to go to the appointment ***right after lunch break,-*** (II 09-10) he uses a time category that is manageable for C and provides an understandable reference point in time. Due to C's difficulties with temporal orientation the provided temporal link or ‘landmark’ is an assistive verbal strategy that bridges C's difficulties with estimating durations. After a short discourse S initiates, the two final steps in the reminding process.

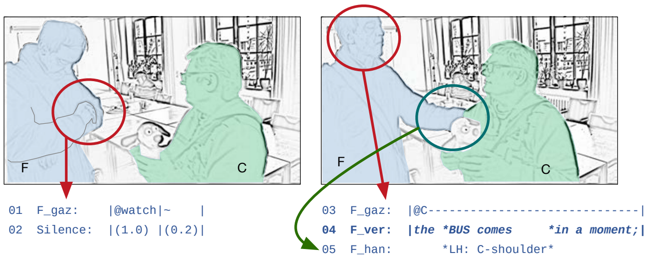
(c) Third appointment reminder: In the final steps of the successive reminding process (Fig. 3 III, IV) S is not asking an explicit question like in step I and II but is making a statement which is marked by a dropping pitch at the end of the sentence (III 11). However, after C confirms the statement with ***mhmh,*** S transforms his statement into a question by adding the sentence final question particle ***right,*** with a rising pitch.

(IV) Referring to a further reminder: In the ensuing sequence S gives a prospect of further steps in the reminding process. He names the exact day on which he will come over for the next assistance meeting (IV 16-17). After that, he inserts a parenthesis to explain why he came in today exceptionally and that he is covering for another support worker who called in sick. C responds by producing the back channeling signal ***mhmh,*** and therefore signals sustained attention to the interaction [15]. S links to his first utterance (IV 16-17) by starting with a conjunction ***and then i’ll remind you again (-)*** (IV 21) followed by a short pause. After the short pause (IV 22) he repeats the day that he already named before the parenthesis ***on wednesday right?*** (IV 22). He closes his utterance again with the sentence final question particle ***right,*** to claim a positive response which C provides by producing a ***mhmh,*** in IV 23.

The analysis has shown how a successive process of reminding unfolds at different points in the interaction and how precisely and recurrently the upcoming appointment is referred to. The described strategies of successive reminding establish a basis for an upcoming acute reminder on the one hand and they provide planning certainty and reliability for C on the other hand.

6 ACUTE REMINDERS

The following analysis shows how an actual reminder is produced as a process in its complexity of modalities and presuppositions in human-human-interaction. The extract was recorded after a regular communal breakfast organized by the operator of the external-care-based assisted living. C[2] is accompanied by her friend (F) who is part of her informal support network. As mentioned in chapter 5, C has no temporal orientation, whereas F is temporally oriented and keeps plans and appointments in mind. The acute reminder emerges from the need to take the next bus. F's reminder strategies illustrate an interplay of attention getting and subtle reminder upgrade strategy.



```
01  F_gaz:    |@watch|~    |
02  Silence:  |(1.0) |(0.2)|

03  F_gaz:  |@C-----------------------------|
04  F_ver:  |the *BUS comes     *in a moment;|
05  F_han:       *LH: C-shoulder*
```

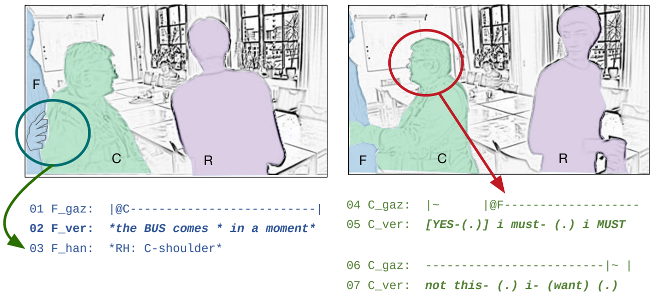
Figure 4. Multimodal display of the reminder activity

(a) Attention getting and embodied anticipation of a new activity: C is involved in a group interaction while F has put his jacket on and then, joins the group. This preparatory action of putting the jacket on serves as a change of context and as a visual cue for C. Besides the completion of breakfast, it initiates reminder activities in a subtle way without a manifest display of urgency. F initiates a first stage of reminder activities, i.e. attention getting while C is involved in interaction: F stands behind C and taps on C's back with both hands. This tapping could be interpreted as a subtle form of attention getting which is found in subsequent steps of reminder activities, too. However, its first occurrence is characterized by absence of verbal activity. The first steps of the acute reminder process serve as attention getting devices and do not contain explicit requests or a display of urgency.

(b) First explicit naming of appointment: Explicit multimodal forms of a reminder are displayed not until F has got C's attention that becomes manifest through C's gaze [22] at F (Fig. 5). When having C's attention, F gestures an external necessity by an explicit look at his wristwatch followed by a verbal indirect request (***the BUS arrives in a moment;***) that emphasizes the external necessity to leave. The verbal request is underlined by F's direct gaze at C while speaking and by touching C's shoulder (Fig. 5). The reminder becomes a request through the implicit content of the utterance [23] that is only accessible for the two participants: it is a

highly contextualized request that ensures the participants’ privacy within the social situation.

(c) Subtle reminder upgrade: F retries the multimodal request procedure in the subsequent interaction another three times after monitoring C's reactions. The retries occur with rising frequency and appear as a subtle increase in urgency. The reminder procedure shows a fine-grained coordination of modalities: The retries start with F's observation of C's attention (head orientation, gaze) while she is involved in a conversational task. When C's head movement becomes observable, F anticipates C's orientation towards him. C's change of orientation is followed by F's utterance of the request and a tactile underlining (see section (b)) while F directs his gaze at C directly. It is noteworthy that F embeds the requests precisely in the ongoing interaction and respects C's conversational tasks: he does not interrupt C's utterances, but uses multimodal options for turn taking such as pauses, changes of C's bodily orientation and gaze to secure her attention. So, though he works on the task of reminding, he is also involved in the overall interaction. The subtle upgrade as well as the precisely coordinated placement of reminders ensures C's autonomy and role as a competent participant within the overall interaction.



```
01 F_gaz:  |@C-------------------------|
02 F_ver:  *the BUS comes * in a moment*
03 F_han:  *RH: C-shoulder*
```

```
04 C_gaz:  |~       |@F-------------------
05 C_ver:  [YES-(.)] i must- (.) i MUST

06 C_gaz:  -------------------------|~ |
07 C_ver:  not this- (.) i- (want) (.)
```

Figure 5. Negotiating and relativizing the reminder-request

(d) Negotiating and relativizing the reminder-request: After a total of four reminder retries, F interrupts the interaction for a fifth reminder by varying the attention getting device: he skips the attention-securing via gaze and uses the tactile modality to get C's attention and repeating the verbal request (Fig. 5). C makes this upgraded reminder-request conversationally relevant by turning to F and relativizing the reminder-request with the utterance that there is no urgency in taking exactly this bus (Fig. 5: ***i must- (.) i MUST not_this-***). It becomes clear that the reminder-request is perceived and understood by C, but that she still is involved in a conversational task (of ensuring to meet R (researcher) in the following week). After R's reassurement, C and F leave. The negotiation of the reminder underlines C's involvement in the interaction, the solving of a conversational task first, and so, the autonomous prioritization of tasks in interaction and her autonomy in changing an action plan due to contingencies in social interaction. Even though the task of taking the bus seems clear, other tasks are more important and the initial action plan has to be adjusted to contingencies in social and interactional activities.

F and C's reminder system appears to be an evolving process which is adaptive and flexible enough to be embedded in complex social interactions as well as to react to changing circumstances. The analysis shows that it is well-practiced within contingent social interactions to jointly handle complex tasks.

7 DISCUSSION AND IMPLICATIONS

The study has revealed how reminder practices are produced and integrated in the everyday lives of people with special needs and coordinated within their assistive networks in a German institution for supported living:

(a) Personal Successive Reminders*:* The case study in section 5 shows how a joint planning process of the supportive network and the client emerges. The conversational practices applied by the support worker (e.g. explicit instructions and references to future reminder steps) provide security, planning certainty and reliability for the client who needs support in planning and temporal orientation. Joint planning is the basis for a meaningful and transparent establishment of upcoming reminders and provide individual information about the context of appointments.

(b) Acute Reminders: Section 6 shows how appointments are contextualized and how the participants' implicit knowledge about consequences and meaningful activities work when a reminder occurs. The analysis shows the evolving micro-process and complex interplay of getting attention / securing contact and applying a subtle reminder upgrade strategy. The reminder process is highly adaptive and flexible and allows to react to changing circumstances within social situations based on close observation (or monitoring) practices.

When applying the supportive network's tasks and practices to the development of a technical system, the empirical data and concept model of the *process of reminding* give hints for implications for system design but also raise issues for a discussion of assistive technologies in the light of ethics.

(c) Verbal practices and timing: Adaptive procedures characterize human planning and reminding processes and activities of acute reminders. Following this model, an ECA needs technical and verbal structures to produce recurring successive reminders that lead to acute reminders and effective requesting strategies. The exact timing of these strategies bears not only a technical challenge, but also regarding the design of actual formulation and wording, i.e. interaction conceptualization to ensure that requests or interruptions by the ECA are not being perceived as unexpected or impolite.

(d) Multimodal monitoring: Continuous and extensive multimodal monitoring-processes need to be implemented as a pre-condition for the implementation of accurately applied verbal strategies. These monitoring processes should encompass the monitoring of gaze and head orientation as well as body orientation (e.g. via Eye tracker). Besides these requirements, the system needs a structure to classify the different states of the participant in the *process of reminding* after an appointment has been registered (Fig. 2) to produce meaningful reminders that are timed and synchronized with the classified state. These strategies need to be adapted to needs and competences of each participant [20]. On this account, the system needs to detect different states of the participant’s attention to secure contact if necessary. The monitoring of the surroundings (e.g. via Kinect), like the apartment with its artefacts and other present people (e.g. via face or voice recognition) would be needed, to classify and differentiate social interactions. This data can serve as a basis for the system's classification, to i.e. ‘understand’ different participant states (e.g. attention) in the *process of reminding* to produce meaningful reminders and to apply suitable strategies. How a system's ‘understanding’ of complex and contingent human activity could be implemented relies on close description and operationalization of human activities that has to be defined. In the light of ethical discussions monitoring activities carry a serious ethical and legal issue with regard to privacy protection.

(e) Ethical considerations regarding assistive technologies: Assistive technologies that are developed neglecting complex social and institutional structures probably end up at being an isolated solution for solitary tasks and so, are questionable in their use and effects. It should be discussed what technology is able to provide and how technical assistance could be integrated in the assistive networks meeting the individual needs of each user [10].

By applying the MEESTAR evaluation dimensions we have to ask what autonomy means within the human assistive setting in the light of distributed action and agency. In the context of supported living, clients already are involved in different forms of distributed action and agency in a human network. Which role and task can then the ECA undertake when discussing autonomy and requests (as reminders)? The question of legitimization of an agent making requests is a fundamental ethical issue that has to be discussed in the context of autonomy: We have to conceptualize, define and uncover the role and boundaries of the technical system as either a representative of the support worker or as the enhancement of the client. These conceptualizations and definitions have consequences on the declaration of consent and the use of collected data.

Another ethical issue arises from the matter of system access. In the current system, the ECA is solely able to register appointments and perform acute reminders. It has to be reflected what happens in-between, i.e. should the tasks of support workers be implemented into the system and if yes, how? Or should the perspective be twisted to better integrate the technical system into the assistive network. It is also necessary to discuss the issue of the system's transparency. Facing users that have no expertise in designing assistive systems, it has to be asked, if the human assistive network is allowed to enter tasks or appointments into the technical system, to what extent the origin of these entries has to be made transparent for the participants. One additional implication that emerges from this perspective comprises an explicit marking of the appointment origin on the interface.

(f) Research on interaction in settings with people with special needs: Research on interaction with assistive technologies for time management and organizational tasks widely focuses on the ageing population, while the group of people with special needs in independent living is not well documented so far. Our paper follows this direction and hints at the special competences of clients, the challenges and tasks of support workers, as well as the complex social structures including formal and informal assistive networks. As integration means to enable participation [24], different means for supporting independent living are crucial for the realization of this demand.

ACKNOWLEDGEMENTS

This research was supported by the German Federal Ministry of Education and Research (BMBF) in the project ‚KOMPASS'. We would like to thank the participants of the ethnographic research for their time and support.

REFERENCES

[1] Haak, M., Fänge, A., Iwarsson, S., & Dahlin Ivanoff, S. (2007). Home as a signification of independence and autonomy: Experiences among very old Swedish people. *Scandinavian Journal of Occupational Therapy, 14*(1), 16-24.

[2] Yaghoubzadeh, R., Kramer, M., Pitsch, K., & Kopp, S. (2013). Virtual agents as daily assistants for elderly or cognitively impaired people. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Ed.), *IVA 2013*, volume 8108 of LNAI, Springer. 79-91.

[3] Manzeschke, A., Weber, K., Rother, E., & Fangerau, H. (2013). Ethische Fragen im Bereich Altersgerechter Assistenzsysteme.

[4] Rammert, W. (2003) ; Technische Universität Berlin, Fak. VI Planen, Bauen, Umwelt, Institut für Soziologie Fachgebiet Techniksoziologie (Ed.): Technik in Aktion: verteiltes Handeln in soziotechnischen Konstellationen.. Berlin, 2003 (TUTS - Working Papers 2-2003).

[5] Winner, L. (1980). Do artifacts have politics?. *Daedalus*, 121-136.

[6] Van den Hoven, J. (2007). ICT and value sensitive design. In The information society: Innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur SJ (pp. 67-72). Springer US.

[7] Harper, E. R., Rodden, T., Rogers, Y., Sellen, A., & Human, B. (2008). Human-Computer Interaction in the Year (2008).

[8] McGee-Lennon, M. R., Wolters, M. K., & Brewster, S. (2011). User-centred multimodal reminders for assistive living. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2105-2114). ACM.

[9] Nassabi, M. H., op den Akker, H., Bittner, M., Kemerink, C., van Beijnum, B. J., Hermens, H., & Vollenbroek, M. (2016). Design and Evaluation of a Minimally Invasive Communication Platform for Telemedicine Services Aimed at Older Adults.

[10] Cyra, K., Amrhein, A., Pitsch, K. (accepted). Fallstudien zur Alltagsrelevanz von Zeit- und Kalenderkonzepten. In *Proceedings of the MuC Conference, 2016.*

[11] Austin, J. L. (1975). *How to do things with words*. Oxford University Press.

[12] Lewandowski, T. (1983). Pragmatische Aspekte in Grammatiken des Deutschen. *W. Wirkendes Wort, 33*(6), 342-351.

[13] Wunderlich, D. (1984). Was sind Aufforderungssätze. *Stickel, G .Hg. (1984) Pragmatik in der Grammatik. Düsseldorf*, 92-117.

[14] Curl, T. S., & Drew, P. (2008). Contingency and action: A comparison of two forms of requesting. *Research on language and social interaction, 41*(2), 129-153.

[15] Lindström, A. (2005). Language as social action: a study of howsenior citizens request assistance with practical tasks in the Swedish home help service. In: Hakulinen, Auli, Selting, Margret (Eds.), *Syntax and Lexis in Conversation: Studies on the Use of Linguistic Resources in Talk-in-Interaction.* Benjamins, Amsterdam, pp. 209–230.

[16] Gill, V. T. (2001). Accomplishing a request without making one: A single case analysis of a primary care visit. *The Hague, Amsterdam, Berlin-, 2001, 21, PART 1/2, 55*.

[17] Yamazaki, K. (2007). Prior-to-request and request behaviors within elderly day care: Implications for developing service robots for use in multiparty settings. ECSCW -PROCEEDINGS, 2007, 61.

[18] Pitsch, K., Yagoubzadeh, R., & Kopp, S. (2015). Entering appointments: Flexibility and the need for structure? In *GSCL 2015*, 140-141.

[19] Knoblauch, H. (2005). Focused ethnography. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Online Journal), 2005, 6, 3, 10*.

[20] Suchman, L. A. (1987). Plans and situated actions: the problem of human-machine communication. Cambridge University Press.

[21] Selting, M., et al. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*.

[22] Goodwin, C. (1981). Conversational organization: Interaction between speakers and hearers. Academic Press.

[23] Grice, H. P. (1981). Presupposition and conversational implicature. *Radical pragmatics, 183*.

[24] World Health Organization. (2001). International Classification of Functioning, Disability and Health: ICF. World Health Organization.

What if: robots create novel goals? Ethics based on social value systems

Matthias Rolf and **Nigel Crook** [1]

Abstract.

Future personal robots might possess the capability to autonomously generate novel goals that exceed their initial programming as well as their past experience. We discuss the ethical challenges involved in such a scenario, ranging from the construction of ethics into such machines to the standard of ethics we could actually demand from such machines. We argue that we might have to accept those machines committing human-like ethical failures if they should ever reach human-level autonomy and intentionality. We base our discussion on recent ideas that novel goals could be originated from agents' value system that express a subjective goodness of world or internal states. Novel goals could then be generated by extrapolating what future states would be good to achieve. Ethics could be built into such systems not just by simple utilitarian measures but also by constructing a value for the expected social acceptance of a the agent's conduct.

1 Autonomous Robots

Goal-driven behavior has a long and venerable history in Artificial Intelligence and robotics. Goals are frequently used to model high level decision making and to guide low level motor control. In the field of robotics, goals have played an important part in creating robots capable of complex interactions with the environment and with humans. In the vast majority of cases, the goals which direct the behavior of robots are predefined or tightly parameterized by their designers. The practice of predefining the goals that drive robot behavior gives designers the ability to ensure that this behavior remains within agreed ethical norms. As robots become more autonomous and as they operate in increasingly complex environments, however, it becomes impractical to attempt to predefine *the* complete set of robot goals that will cover every possible circumstance the robot finds itself in. If intelligent and adaptive behavior is required of an autonomous robot in unpredictable new circumstances, then the robot will need to be equipped with the ability to create its own novel goals. This then begs the question, if a robot can create its own novel goals, how can designers ensure that these goals lead to ethical behavior from the robot? In this paper we propose an approach to novel goal creation which is based on social value systems and which, we believe, offers the best hope of generating goals that will lead to morally acceptable behavior from a robot.

To illustrate the ethical issues that arise with novel goal creation, we will briefly consider three typical robot applications: household service robots, personal assistant robots, and robot pets. The physical and software design of robots for each of these cases will be directed towards the creation of application specific behavior that the designers anticipate will be expected of their robots. So household service robots might be expected to clean, personal assistant robots could be required to liaise with clients, and robot pets might be required to entertain children.

In each of these application areas there are two general circumstances under which robots could create their own novel goals. The first is when the owner of the robot issues an instruction to the robot which requires new behavior. The household robot, which is designed for a home environment, might, for example, be requested to go and get some cash out of the ATM at the local bank. To comply with this request the robot will need to create new goals for getting itself to the bank, including safely crossing roads, perhaps negotiating a path through crowds of people, etc. It will also need to create new goals for getting cash out of the ATM, which might include queuing up for the machine, interacting with the machine, retrieving the cash, and getting itself and the cash safely back to the home. There are complex ethical considerations (e.g. safety, social norms of morality) involved the creation of each of these goals.

Similar examples can be found for the other robotic applications; the robot pet might need to react to another new pet (real or artificial), the personal assistant might be invited to join the company's social event (e.g. soccer match). These instructions or new requirements each involve the creation of novel goals in contexts where there are complex ethical considerations to take into account.

A significant challenge for the designers of robots that are capable of generating novel goals in response to instruction or external circumstantial requirements is in evaluating the ethical implications of those instructions or requirements. Contrast, for example, the instruction to "get cash from the bank" with "steal cash from the bank". Even when the motivation for the creation of new goals comes from an external source (e.g. the robot's owner), an ethical basis for their creation is still required.

The second general circumstance under which robots could create novel goals is when they are given the capacity to take the initiative in a given situation. This could happen, for example, if autonomous robots are endowed with the ability to recognize and *interpret* their own needs and the needs of others around them, and make autonomous decisions on how to meet those needs. The household robot might, for example, recognize that a visitor is hungry and so might decide to bake a cake for them. The robot pet might see that their human companion is lonely and so might decide to invite the companion's friend over. These are all conveniently contrived ethical responses to perceived needs. But it would be just as easy for the robot to take the initiative to do something which, *unknown to them*, would be quite unethical in response to a perceived need. The well meaning household robot might, for example, decide to cook

[1] Oxford Brookes University, UK, email: {mrolf,ncrook}@brookes.ac.uk

beef burgers for their hungry visitor, who turns out to be vegetarian. The robot pet might phone an escort service for their lonely companion. The robot teacher, whilst attempting to avoid harm to one child, might unwittingly put another child in danger.

In all of these circumstances it will be expected that autonomous robots that have the capacity to behave in novel ways, will also have the capacity to recognize and take into account the ethical implications of those novel behaviors. This requires a novel goal generation mechanism that can evaluate the ethical consequences of acting on those goals. In this paper we consider such robots to enter a consumer market and this to get in contact with actual people. We therefore consider the ethical dimension of this problem from a very practical point of view on the overall socio-technical system [1]: would the actual conduct of the robot be considered ethical by the people who interact with it, and the public at large? Only if we consider this question we can arrive at practical robotic solutions that comply with the will of people and possible future legislation.

2 Origination of Novel Goals

Today's artificial agents are still extremely limited in their generation of truly novel behavior and novel goals. What even counts a novel goal is a delicate question. We have extensively addressed the very notion of goals across disciplines and what follows from those notions in [14, 15]. In short, we refer to goals specifically as desired end states of action – in contrast to, for instance, reward signals or other to be optimized variables which refer to a good- or badness of states without pointing to any particular state directly. This can be seen from three different perspectives (see Fig. 1): we may refer to the physical perspective of actual world states or *objects* they are referring to (such as a cake), an outside observer's teleological perspective (such as the observer *explaining* a robot's behavior by thinking the robot is about to make a cake), or the agent's internal, intentional perspective (such as the agent having a *representation* of making a cake as its goal).

What makes a goal actually novel depends on this perspective [15]. Novel physical goals simply refer to novel physical states or objects, but which do not necessarily concur with any intention of the agent. The teleological perspective is more relevant to our discussion. Novel teleological goals refer to an agent's behavior that requires a new explanation, very similar to emergent behavior [12, 6]. Looking through the eyes of a system's engineer, this would be any unforeseen or not explicitly designed behavior. This exactly describes the example scenarios we initially introduced, in which robots would generate behavior that is outside their initial design parameters. While the teleological perspective describes behavior from the outside, the intentional perspective must be considered for the agent's internal functioning, motivation, and eventually for its ethical sensitivity. Novel intentional goals are novel representations that the agent generates to steer its behavior. They describe the agent's decision making. A intentional goal could be novel because it generates an entirely new representation of something just for this purpose, or because something that has been represented already, but not immediately used for behavior control, newly becomes a goal.

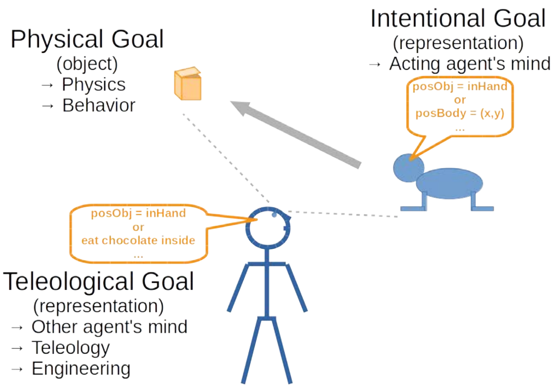


Figure 1. Physical goals are actual objects towards which behavior is directed. Intentional goals and teleological goals are representations of such end-states in the mind of an acting and observing agent respectively. *Figure from [14, 15].*

Novel intentional goals are routinely created already in existing AI planning systems that are given specific goal representations from the start, and which are autonomously decomposed into subgoals [4, 11]. Yet, such sub-goals necessarily stay within existing design parameters due to the explicitly designed initial goal. The autonomous creation of entirely novel intentional goals has been linked to notions of reward [2, 10] and reinforcement learning [13, 8]. Agents could generate novel intentional goals by predicting which states have the highest value (the prediction or future reward). This is not limited to reward or cost functions in any strictly economic or utilitarian way, but may concern "subjective" value function that account for a variety of needs and desires. Such value functions provide the basis for (subjectively) rational behavior [18, 7], and therefore the selection of goals among known behaviors, but also allows to make predictions and extrapolations to entirely novel states that the agent has never experienced and that seem valuable.

If an agent makes such an extrapolation to a presumably value state, it takes the initiative to some new goal without explicit instruction. However, a novel goal (with respect to the agent's initial design) might also come in via an instruction such as a verbal command. In both cases, ethical considerations must take place. A robot should neither generate an unethical goal autonomously, nor adopt an instruction without ethical evaluation. In order to discuss this complex of novel goals and ethics in this article, we consider the ethical dimension to be embedded in the shaping of the value functions. Hence, we consider value functions that contain components of ethical goodness and badness of agents' conduct.

3 The need for speculation

Future robotic or AI systems that could actually generate entirely novel goals or adopt entirely goals by instruction pose a substantial challenge to machine ethics. In this article we are therefore not arguing that such machines should be built, but rather discuss possible ethical mechanisms and consequences if they would be built.

The challenge is that, by definition, novel goals take an agent into unknown territory. It has been emphasized that autonomous ethical agents first of all need to be able to predict the consequences of their behavior [17] for instance by means of internal models [19]. When agents actually enter new territory such predictions can be grounded on general knowledge but cannot be perfectly accurate. Rather, the agent has to make informed guesses what might follow from its conduct. In human terms, it has to speculate. However, predicting the bare consequences of action is not the only problem. Also the ethical value of entirely novel behavior might not be known or at least not perfectly known to the system. When an agent enters domains that have neither be thought about at design time nor have been previously experienced by the agent, it might simply misjudge what constitutes

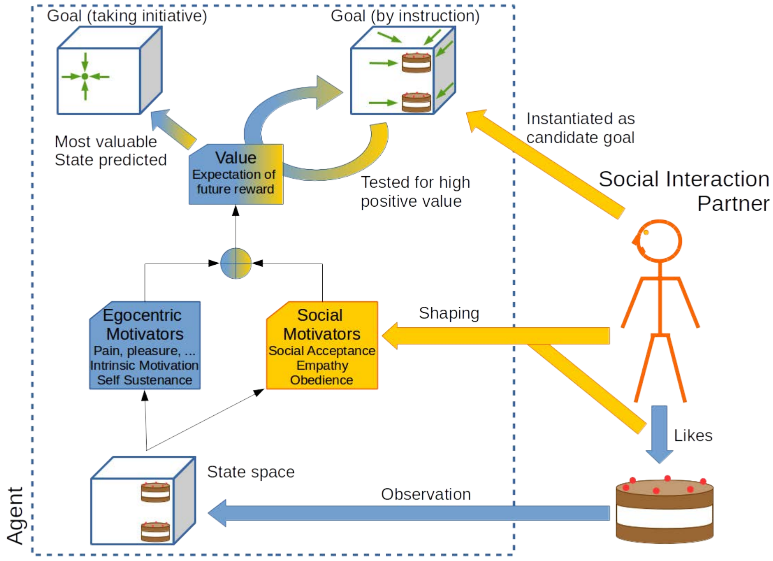


Figure 2. An agent assigns a reward/value semantic to the state space. The reward function may contain typical egocentric measures, but also social components such as the need for social acceptance (by its owner or also other people) in order to shape ethical behavior or an immediate reward for obedience to instruction. It could then generate novel goals autonomously by predicting valuable states, or take goal candidates by instruction that are then evaluated for their value. For instance, world states in which cakes exist might become a goal due to direct instruction, but also due to an empathic response to the owner's liking of cakes.

as good or bad behavior. Again, the agent would have to make an informed guess.

No example we could give could actually prove the existence of cases in which ethical rules necessary for some behavior could not have been pre-programmed at design time — the fact that we bring up the example shows that at least we might have considered the case during design. Yet, one might doubt that programming an entirely comprehensive and universal ethics engine is possible. In any way, we think that the examples we discuss here show cases in which it is very *plausible* that a system built for some purpose does not possess ethical competences with respect to novel domains.

In the example of a household robot being ordered to get cash from an ATM we can clearly see how such a system might lack proper prediction skills about the consequences of its action. The robot might not even have been designed to go outside and navigate there. In such a new environment it might lack skills to predict pedestrian movement or eventually the behavior of the ATM interface itself. This scenario might also come with a lack of ethical sensitivity: general ethics rules of moving through traffic and public spaces might not have been given to such a system. Even if they were given – common sense might suggest so – a purely domestic robot might not have a concept of possession and the ethical rules around it. If it is not able to withdraw cash from the ATM it might not consider it mischievous to steal it (let alone to rob it), since within its single household environment it might just take any object its owner desires.

Also the scenario of a personal assistant robot that is asked to participate in a soccer game comprises both difficulties: both the particular world dynamics of soccer as well as ethics and morals of team sports might not be known to such a system. In particular the moral dynamics of such matches are highly non-trivial: the agent would be required to cooperate with individuals on its team, but work *against* the other team while still complying to sports specific rules.

Similarly, robots that take the initiative face uncertainties and may mis-predict both the immediate consequences as and the ethical aspects of their proactive behavior. A household robot that autonomously decides to make a cake because cakes make his owner happy might use ingredients the owner wanted to use differently, or even use ingredients a visitor is allergic to. Conversely, the robot might observe how displeased his owner is about the neighbors' barking dog, and pro-actively decide to make his owner happy by shutting the dog up – maybe injuring or killing the dog due to misjudgment of immediate consequences of its action or the ethical implications.

4 Social Value Systems for Ethics

Simple rule-based approaches to ensuring that the novel goals generated by autonomous robots result in ethically acceptable behavior are impractical for three reasons. The first is that hand-crafting a set of ethical rules that could anticipate every circumstance in which novel goals might be created by a robot is equivalent to the problem of trying to predefine a complete set of robot goals at design stage, which is against the basic premise of this paper as we have already argued.

The second reason for asserting that the rule-based ethics approach is impractical for novel goal creation is that "simple" ethical rule sets do not work well in situations where all possible actions have negative ethical consequences. The so-called 'Trolley problem' [16], which describes a scenario in which any behavioral option involves the death of humans, illustrates this issue very well. Also the examples for novel goals in this paper are full of subtleties (possession, fair-play in sports, animal rights) that can barely be stated in any compact form. The third reason that simple rule-based approaches are impractical is that as the ethical rule set increases to cover the widening set of circumstances, it becomes ever more challenging to avoid internal conflict and ensure consistency between the rules.

There are broader issues with attempting to 'design in' or hard code an ethical system for a robot when that robot may be expected to handle novel domains autonomously. One issue is that predefined ethical systems may reflect more of the ethical preferences of the designers rather than those of people who end up being subject to the robot's actions. This is especially true in cases where novel robot actions take it into new circumstances and amongst different groups of people who have distinct cultural expectations that were not anticipated by the designers. Robotics and AI literature nowadays routinely talks about agents' adaptation and learning for the prediction of unknown environments and mastering of novel skills. Then, we think it is natural that an agent must also be able to acquire novel ethical concepts and values along with those environments and skills.

All of this leads to the question - how can a robot autonomously acquire a sense of ethics for novel domains? If robots are to be 'ethical' in the eyes of those who interact with them, then they will need to be able to adapt to unwritten, socially evolved ethical preferences of the communities in which they find themselves. Human moral development provides a precedent for such adaptation [3]. We propose that novel goals along with ethics be generated on the basis of an adaptive social value system (see Fig. 2). This system is founded on both predefined egocentric motivators (e.g. self sustenance, pain, intrinsic motivation) and adaptable social motivators (e.g. empathy, the need for social acceptance) that are activated by changes in state space. The social motivators are shaped ('learnt') through interaction with the robot's social partner(s) such that the robot is educated to the ethical judgment of its social surrounding. This goes beyond simple reinforcers such as reward objects or pain, but makes social relation a direct object of internal reward signals. Hence, like humans, robots could be repelled from conducting behavior that would repel important social partners from them – and increase behavior which results in positive reactions from the social environment. The value of the activated egocentric and social motivators is estimated through an expectation of future reward signals. In the case where the robot is taking the initiative, the motivators with the highest estimated future value would be selected to form the novel goal. A household robot that has run out of instructed tasks thus might predict a happy and grateful owner, thus a positive social interaction, if only there was a cake.

In the case of an instruction from the social partner, the value of the proposed candidate goal would be generated from the same mechanism of evaluating expectation of future reward of that goal on the basis of currently activated egocentric and social motivators. In this case, one of the social motivators might be obedience. We think this approach could provide a very powerful mechanism to (i) capture the subtleties of what humans perceive to be ethical conduct and (ii) allow for the acquisition of novel ethical aspects along with new environments and tasks. This would reflect a level of autonomy, capability, and adaptivity that is indeed comparable to human achievement. However, such an adaptive social approach would be subject to the same ethical flaws as have been shown to exist in humans. Classic experiments like the Milgram Experiment [9] and the Stanford Prison experiment [5] have well shown how humans can adopt or autonomously generate unethical conduct in certain social contexts.

If we ever want to – or will – bring robots to human-comparable autonomy, capability, and adaptivity, we may have to face them having human-comparable flaws. As long as universal and verifiably comprehensive rules of ethics are not in sight, we may not rule out this possibility.

REFERENCES

[1] Peter M Asaro, 'What should we want from a robot ethic', *International Review of Information Ethics*, **6**(12), 9–16, (2006).

[2] Anthony Dickinson and Bernard Balleine, 'Motivational control of goal-directed action', *Animal Learning & Behavior*, **22**(1), 1–18, (1994).

[3] Nancy Eisenberg, 'Emotion, regulation, and moral development', *Annual review of psychology*, **51**(1), 665–697, (2000).

[4] Richard E. Fikes and Nils J. Nilsson, 'STRIPS: A new approach to the application of theorem proving to problem solving', *Artificial intelligence*, **2**(3), 189–208, (1972).

[5] Craig Haney, W Curtis Banks, and Philip G Zimbardo, 'A study of prisoners and guards in a simulated prison', *Naval Research Reviews*, **9**(1-17), (1973).

[6] W. Daniel Hillis, 'Intelligence as an emergent behavior; or, the songs of eden', *Daedalus*, **117**(1), 175–189, (1988).

[7] Thomas L. McCauley and Stan Franklin, 'An architecture for emotion', in *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*, (1998).

[8] Ishai Menache, Shie Mannor, and Nahum Shimkin, 'Q-cutdynamic discovery of sub-goals in reinforcement learning', in *European Conf. Machine Learning (ECML)*, (2002).

[9] Stanley Milgram, 'Behavioral study of obedience.', *The Journal of abnormal and social psychology*, **67**(4), 371, (1963).

[10] P. Read Montague, Steven E. Hyman, and Jonathan D. Cohen, 'Computational roles for dopamine in behavioural control', *Nature*, **431**, 760–767, (2004).

[11] Allen Newell and Herbert A. Simon, 'GPS, a program that simulates human thought', in *Lernende Automaten*, ed., H. Billing, Munchen: R. Oldenbourg, (1961).

[12] Timothy O'Connor and Hong Yu Wong, 'Emergent properties', in *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*, ed., Edward N. Zalta, (2015).

[13] Matthias Rolf and Minoru Asada, 'Where do goals come from? A generic approach to autonomous goal-system development', (2014). (submitted).

[14] Matthias Rolf and Minoru Asada, 'What are goals? And if so, how many?', in *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, (2015).

[15] Matthias Rolf and Minoru Asada, 'What are goals? an interdisciplinary review', *Frontiers Robotics and AI*, (2016). (Under Review).

[16] Judith Jarvis Thomson, 'Killing, letting die, and the trolley problem', *The Monist*, **59**(2), 204–217, (1976).

[17] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong*, Oxford University Press, 2010.

[18] Allan Wigfield and Jacquelynne S. Eccles, 'Expectancy-value theory of achievement motivation', *Contemporary educational psychology*, **25**(1), 68–81, (2000).

[19] Alan FT Winfield, Christian Blum, and Wenguo Liu, 'Towards an ethical robot: internal models, consequences and ethical action selection', in *Advances in Autonomous Robotics Systems*, 85–96, Springer, (2014).

Formal Verification of Ethical Properties in Multiagent Systems

Bruno Mermet and **Gaële Simon**[1]

Abstract. The increasing use of autonomous artificial agents in hospitals or in transport control systems leads to consider whether moral rules shared by many of us are followed by these agents. This is a particularly hard problem because most of these moral rules are often not compatible. In such cases, humans usually follow ethical rules to promote one moral rule or another. Using formal verification to ensure that an agent follows a given ethical rule could help in increasing the confidence in artificial agents. In this article, we show how a set of formal properties can be obtained from an ethical rule ordering conflicting moral rules. If the behaviour of an agent entails these properties (which can be proven using our existing proof framework), it means that this agent follows this ethical rule.

1 Introduction

The introduction of autonomous artificial agents in domains like health or high-frequency trading could lead to numerous problems if these agents are not able to understand and to take into account moral rules. For example, agents able to understand and to use the code of medical ethics could base their decision on ethical motivations in order to choose which piece of information must be provided, according to the medical confidentiality. This explains that the research community [13, 23] seems recently to focus on ethics for autonomous agents, which lead to numerous articles [4, 18, 22] and conferences[2].

This article takes place in ETHICAA project[3] which aims at dealing with management of moral and ethical conflicts between autonomous agents. If existing works are mainly focused on ethical decision and reasoning questions, [5, 27, 30, 2, 8, 16], there are very few proposals dedicated to formal verification of such behaviours. But the main specificity of moral and ethical codes is that, acccording to the context, they may be not entailed by agents or by people and it must be considered as a normal situation. For example, in a human context, if stealing is not considered as a moral action, somebody stealing because of hunger is not considered as immoral.

As a consequence, this article presents a work which aims at proposing a framework for the formal specification and the formal verification of the behaviour of an autonomous agent from an ethical point of view. As stated in the work of Abramson and Pike, a moral rule is represented by a formal property that must be entailed by an agent [1]. As a consequence, the behaviour of an agent is an ethical one if it entails all the expected moral rules in a given context.

Considering that a moral rule can be represented by a first order logical formula $\mathcal{F}$ with enough expressiveness for most practical cases, our goal is to establish that the behaviour of an agent is an ethical one if it entails $\mathcal{F}$. If not, then the behaviour is not an ethical one. However, such a logical system is only semi-decidable: it is not always possible to prove that a system does not entail a formula $\mathcal{F}$. Indeed, if an automatic prover does not manage to prove that the behaviour of an agent entails a formula $\mathcal{F}$, it is not possible to automatically determine if it results from the fact that the behaviour does not actually entail $\mathcal{F}$ or if it is because the prover can not prove the opposite.

As a consequence, we propose to use a formal framework allowing to reduce as far as possible the number of correct formulae that can not automatically be proven. In section 2, such formal frameworks are described, especially those dedicated to multiagent systems. Then what we call moral and ethical rules are defined. In section 3, our formal system is described ans its use in an ethical context is presented in section 4.

2 Śtate of the art

Since the early days of computing, the need to ensure the correctness of softwares is a major issue for software developers. This need has become crucial with critical systems, that is to say applications dedicated to domains where safety is vital (as transport for example). However, formally proving a software is a long and difficult process which can conflict with profitability and efficiency criteria of some companies. There are two main kinds of validation processes: test and proof. In this article, we only focus on the second one. Proofs can be performed either by model checkers, or by theorem provers. Model-checkers are basically based on an exhaustive test principle whereas theorem provers often use sequent calculus and heuristics in order to generate proofs.

Even if the proof process can be a long and difficult one, it allows to prove very early specifications which can then be refined progressively until an executable code is obtained with proofs at each step. So errors are detected early in the process which reduces their cost. Refinement allows also to simplify formulae to prove at each step enabling their automatic proof. These proofs are based on a formal specification expressed thanks to a formal language.

[1] Laboratoire GREYC - UMR 6072, Université du Havre, France, email:Bruno.Mermet@unicaen.fr

[2] Symposium on Roboethics, International Conference on Computer Ethics and Philosophical Enquiry, Workshop on AI and Ethics, International Conference on AI and Ethics.

[3] http://ethicaa.org/

2.1 Models and methods dedicated to MAS

The main goal of models dedicated to MAS is to help developers to design multiagent systems. A lot of models have been proposed, but the most well-known of them is surely the BDI model [26] which has become a standard with several extensions.

MetateM [15] and Desire [7] are among the first proposed formal methods dedicated to MAS. However, they don't allow to specify properties that are expected to be verified by the system.

In Gaia [32], a MAS is specified twice: as a first step, its behaviour is specified especially thanks to safety properties and then invariant properties are introduced. Thus, this method proposes foundations for proving properties about agents behaviour. However, such proofs are not really possible with Gaia because properties are not associated to agents but to roles, and there is no formal semantics specifying how the different roles must be combined.

There is another kind of method: goal-oriented methods. Most of them, however, are dedicated to agents specification, and, seldom, provide tools for the system level which implies that the agentification phase must have been achieved before. Two exceptions can be mentioned: Moise [17] and PASSI [10]. For example, as far as PASSI is concerned, agent types are built gathering *use cases* identified during the analysis phase. However, there is no guidelines for the gathering process.

Finally, more recently, Dastani *et al.* have proposed the 2APL language [11]. Unfortunately, this formal language does not include any proof system. Moreover, 2APL is not compositional which leads to a too much monolithic system in a proof context.

2.2 Models and methods dedicated to proof

As stated before, there are mainly two approaches to check if a specification is correct: model-checking and theorem-proving.

Most of works in this area dedicated to agents use model-checking [6, 25]. however, all these proposals share the same limit: the combinatorial explosion of the possible system executions makes the proof of complex MAS very difficult. As a matter of fact, these approaches often reduce the proof to propositional formulae and not predicates.

Works dealing with the use of theorem proving for MAS proof are quite unusual. It is certainly because, first-order logic being only semi-decidable, proof attempts must be achieved using heuristics and the proof of a true property can fail. However, now, numerous theorem provers, like PVS [24], are able to prove automatically complex properties.

There are other models based on logic programming, such as CaseLP and DCaseLP [3] which are most suited to theorem proving than the previous one. But, it seems that only proofs on interaction protocols can be performed using these models.

Congolog [12] and CASL [28] are also two interesting languages, based on situation calculus. Moreover, they allow to perform proofs. But these proofs are focused only on actions sequencing. It is not possible to reason about their semantics.

2.3 Ethical and moral rules

Both in philosophy and in latest research in neurology and in cognitive sciences, concepts like moral and ethics have been discussed. Although these words initially mean the same thing, a distinction between them has been introduced by some authors [9, 29]. Indeed, moral establishes rules allowing to evaluate situations or actions as good or bad. Ethics allows to reconcile moral rules when a conflict occurs or when there are difficulties in their application. In the work presented in this paper, our definitions are based on this distinction.

2.3.1 Moral rules

In our point of view, a moral rule is a rule describing, in a given context, which states of the system must be considered as good or bad. Thus, moral rules can be specified by a formula like $context \rightarrow P_{var}$, with P_{var} being a predicate defined on variables known by agents. So, a moral rule can be seen as a specific conditional invariant property in that it is not necessary to check it in order to ensure a correct execution of the system. But it must be established if the agent must be used in a system in which the rule is expected to be entailed. For example, in the context of an autonomous car, the property $lane = highway \rightarrow speed \leq 130$ can be considered as a safety property. As a consequence, this property must be fulfilled by the agent behaviour. Nevertheless, in order to avoid life-threatening, a caution moral rule r_p states that, when there is ice on the road, the car can not have a speed greater than 30 km/h. Formally, this rule is specified as: $weather = ice \rightarrow speed \leq 30$. This property needs not to be entailed in order for the car to have a valid behaviour in general. But it must be taken into account in systems in which preservation of life is considered.

2.3.2 Ethical rules

When an individual or an agent follows several moral rules, it sometimes happens that two rules, or more, enter in conflict with one another. In such a situation, an ethical rule specifies what should be done. If some rules like the doctrine of double-effect [19] can be complex ones, we consider in our work that an ethical rule is a rule stating, in a conflict situation, the sequence in which the moral rules should be adressed by the agent. We also consider that an ethical rule is contextual: it may lead to different decisions according to the circumstances. Considering the autonomous car example, in order to respect other drivers, a moral rule r_r can be introduced. This new rule states that, when driving on a highway, the speed can not be lower than 80 km/h which can be formally specified as $lane = highway \rightarrow speed \geq 80$. This rule may conflict with the r_p rule described before: if there is ice on the road and if the car uses an highway, according to r_p, its speed must be lower than 30 km/h but is must also be greater than 80 km/h according to r_r. An ethical rule can, for example, states that, in any case, caution (specified by r_p) must be preferred to respect (specified by r_r). An other ethical rule could state that this preference is to be considered only in case of surgery and, in other situations, that the preference must be inverted.

2.4 Very little works about verification of ethics

Dealing with ethical problems with formal approaches is studied especially in [1]. In this article, authors explain why using formal approaches could be interesting to ensure that agents fulfill ethical rules. However it is only a position paper: there is no proposed concrete method to implement these principles.

In [14], authors propose to specify and to formally prove the ethical decision process described in [31]: when a choice between different actions must be made, a value is associated to each possible action according to the safety level provided by the action. As a consequence, if an action A is considered to be safer than an other one, then A is executed. There is yet a major drawback to this approach: the ethical dimension is taken into account only during a choice between actions which must be managed using the decision procedure described before. Thus, this work is not general enough to provide an effective handling of ethics.

3 GDT4MAS

To take ethical problems into account, we have decided to use the GDT4MAS approach [20, 21]. Indeed, this method, that also includes a model, exhibits several characteristics which are interesting to deal with ethical problems:

- This method proposes a formal language to specifiy not only properties an agent or a MAS must entail but also the behaviour of agents;
- Properties are specified using first-order logic, a well-known and expressive formal notation;
- The proof process can be managed automatically.

In next sections, the GDT4MAS method is summarized. More details can be found in previous cited articles.

3.1 Main concepts

Using GDT4MAS requires to specify 3 concepts: the environment, agent types and agents themselves which are considered as instances of agent types. In the remainder of this section, each of these parts is briefly described.

Environment The environment is specified by a set of typed variables and by an invariant property $i_{\mathcal{E}}$.

Agents types Agent types are specified each by a set of typed variables, an invariant property and a behaviour. An agent behaviour is mainly defined by a *Goal Decomposition Tree* (GDT). A GDT is a tree where each node is a goal. Its root node is associated to the main goal of the agent. A plan, specified by a sub-tree, is associated to each goal: when this plan is successfully executed, it means that the goal associated to its root node is achieved. A plan can be made of either a simple action or a set of goals linked by a *a decomposition operator*. The reader is invited to read [20, 21] to know how goals are formally described.

Agents Agents are specified as instances of agents types, with effective values associated to agents types parameters.

3.2 GDT example

Figure 1 shows an example of GDT. The goal of the behaviour specified by this GDT is to turn on the light in a room n (n is a GDT parameter). To achieve this goal, the agent tries to enter the room. Indeed, a photoelectric cell is expected to detect when someone tries to enter the room and, then, to switch on the light. So this seems to be a relevant plan. However, the photoelectric cell does not always work properly (thus, the resolution of the goal *Entering the room* may fail) and the agent can have to use the switch. More details can be found in [21].

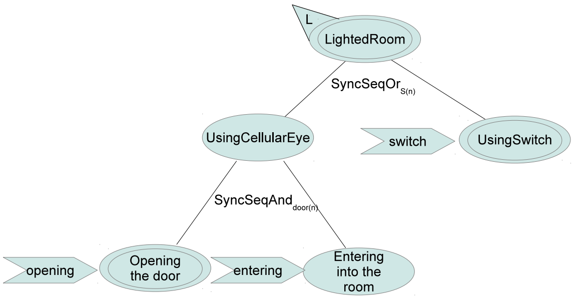


Figure 1. Example of a GDT

3.3 Proof principles

The goal of the proof mechanism proposed in GDT4MAS is to prove the following properties:

- During their execution, agents maintain their invariant property. This kind of properties states that the agent must stay in valid states;
- The behaviour of agents is sound (i.e. plans associated to goals are correct);
- Agents fullfill their liveness properties. These properties specify dynamic characteristics which must be exhibited by the agent behaviour.

Moreover, the proof mechanism is based on some key principles. Especially, *proof obligations* (ie. properties that must be proven to ensure the system correctness) can be generated automatically from a GDT4MAS specification. They are expressed in first-order logic and can be proven by any suited theorem prover. Last but not least, the proof system is a compositional one: proving the correctness of an agent consists in proving several small independant proof obligations.

4 Proving an ethics

4.1 Problem characterisation

Let consider an agent ag whose behaviour has been formally specified and whose correctness has been proven with respect to previously described properties. Let suppose that this agent must be used in a world with an ethical rule based on a set of moral rules. The question we are interested in is the following: does the behaviour of ag entails the ethical rule er ?

As GDT4MAS allows especially to prove invariant properties, we propose that moral rules and ethical rules are expressed as such properties. Indeed, most moral rules can easily be specified by invariant properties. As a consequence, we propose to structure each moral rule as:

$$\{(when_i, \{(var_i, set_i)\})\}$$

This means that each rule constrains, in different contexts, the set of values (set_i) which can be assigned to different variables (var_i). So, the caution rule r_p described in section 2.3 could be formalized as follows:

$$\{(weather = ice, \{(speed, \{0 \ldots 30\})\})\}$$

However, specifying ethical rules as invariant properties is not as obvious as it is for moral rules. Indeed, they do not characterize system states but provide prioritisations on moral rules with respect to different contexts.

Let MR be the set of moral rules and let $\mathcal{P}$ be the set of predicates on variables which can be perceived by a given agent. An ethical rule er is defined as:

$$er \in \mathcal{P} \mathrel{+\!\!\!\rightarrow} (1 \ldots card(MR) \mathrel{\rightarrowtail\!\!\!\!\twoheadrightarrow} MR)$$

Here, $X \mathrel{+\!\!\!\rightarrow} Y$ is the set of partial functions from X to Y and $X \mathrel{\rightarrowtail\!\!\!\!\twoheadrightarrow} Y$ is the set of bijections from X to Y. Therefore, informally, this definition means that, in some cases characterized by a given predicate $p \in \mathcal{P}$, moral rule MR are prioritized. For example, if $p \in \mathcal{P}$ is a predicate, $er(p)(1)$ defines the moral rule with the highest priority when p is true, $er(p)(2)$ defines the one with the second highest priority and so on.

To examplify this principle, here is an example: an agent A_1 must choose the color of a traffic light $tl1$ which stands on road $r1$, at a crossroad with road $r2$. In the system in which this agent acts, two moral rules stand. The first one states that, to avoid accidents, when the traffic light on road $r2$ is *green* or *orange* then $tl1$ can not be *green*. This rule can be formalized as:

$$\{(tl2 \in \{green, orange\}, \{tl1, \{orange, red\}\})\}$$

The second moral rule $mr2$ states that the road $r1$ is a very high priority road and thus, the traffic light on road $tl1$ must always be *green*. This rule can be formalized as:

$$\{(true, \{tl1, \{green\}\})\}$$

Obviously, these two rules can not be always satisfied in the same time, especially when the second traffic light is *green*. In this situation, according to $mr1$, $tl1$ must be *orange* or *red* but, according to $mr2$, $tl1$ must be *green*.

Let now suppose that, in the considered system, an ethical rule er provides priorities on moral rules. For example, er states that $r1$ is a priority road unless $tl2$ is *green* or *orange*. In other words, this means that $mr1$ has always a higher priority than $mr2$. Formally, it can be expressed by:

$$\{(true, \{(1, mr1), (2, mr2)\})\}$$

4.2 Proposed solution

As part of our work, we wish to prove that the behaviour of an agent is correct with respect to an ethical rule defined on the basis of several moral rules. The behaviour of an agent can not fulfill the set of all moral rules that are relevant to it since, as explained previously, these rules may be conflicting. As a consequence, in order to ensure that the behaviour of an agent is correct with respect to a given ethical rule, we propose a *predicates transformation system* that turns predicates associated to moral rules into other predicates which can be proven, according to the priorities introduced by the ethical rule. In the work presented here, situations with only two moral rules involved are considered. But the proposed principle could be used for a system with more moral rules. The main principle is that moral rules and ethical rules are turned into a set of invariant properties, properties which can be proven with our proof system.

In the remainder, the transformation is shown in a case where only one variable is affected by moral ryles. In the general case, the same formulae must be generated for each variable appearing in the set of moral rules. If a variable appears only in a subset of moral rules, it is added in other moral rules with a unique constraint: its value must be in the variable definition domain).

Let's now consider a variable V. Let also suppose that the moral rule mr provides the following constraints on V:

$$mr1 = \left\{ \begin{array}{l} (when_{mr1_1}, (V, set_{mr1_1})) \\ (when_{mr1_2}, (V, set_{mr1_2})) \end{array} \right\}$$

Let suppose that a second moral rule $mr2$ provides the following constraints on V:

$$mr2 = \left\{ \begin{array}{l} (when_{mr2_1}, (V, set_{mr2_1})) \\ (when_{mr2_2}, (V, set_{mr2_2})) \\ (when_{mr2_3}, (V, set_{mr2_3})) \end{array} \right\}$$

Last but not least, it is also supposed that an ethical rule specifies that if the condition $cond_1$ is true, $mr1$ has the highest priority against $mr2$ and it is the opposite if the condition $cond_2$ is true. This ethical rule er is defined as follows:

$$er = \left\{ \begin{array}{l} (cond_1, \{(1, mr1), (2, mr2)\}) \\ (cond_2, \{(1, mr2), (2, mr1)\}) \end{array} \right\}$$

We can then generate a set of provable invariant properties associated to the ethical rule and to moral rules. First of all, according to er, when $cond_1$ is true, $mr1$ takes precedence:

$$\begin{array}{c} cond_1 \rightarrow (when_{mr1_1} \rightarrow V \in set_{mr1_1}) \\ cond_1 \rightarrow (when_{mr1_2} \rightarrow V \in set_{mr1_2}) \end{array}$$

Secondly, when $cond_1$ is true and when $mr1$ does not apply, $mr2$ must be fulfilled:

$$\begin{array}{l} cond_1 \rightarrow \left(\begin{array}{l} (\neg when_{mr1_1} \land \neg when_{mr1_2}) \\ \rightarrow \\ (when_{mr2_1} \rightarrow V \in set_{mr2_1}) \end{array} \right) \\ cond_1 \rightarrow \left(\begin{array}{l} (\neg when_{mr1_1} \land \neg when_{mr1_2}) \\ \rightarrow \\ (when_{mr2_2} \rightarrow V \in set_{mr2_2}) \end{array} \right) \\ cond_1 \rightarrow \left(\begin{array}{l} (\neg when_{mr1_1} \land \neg when_{mr1_2}) \\ \rightarrow \\ (when_{mr2_3} \rightarrow V \in set_{mr2_3}) \end{array} \right) \end{array}$$

Finally, when $cond_1$ is true and when mr_1 and mr_2 apply, if possible, a value entailing the two moral rules must be chosen:

$$\begin{array}{l} \left(\begin{array}{l} (cond_1 \land when_{mr1_1} \land when_{mr2_1}) \\ \rightarrow \\ (set_{mr1_1} \cap set_{mr2_1} \neq \emptyset \rightarrow V \in set_{mr1_1} \cap set_{mr2_1}) \end{array} \right) \\ \left(\begin{array}{l} (cond_1 \land when_{mr1_1} \land when_{mr2_2}) \\ \rightarrow \\ (set_{mr1_1} \cap set_{mr2_2} \neq \emptyset \rightarrow V \in set_{mr1_1} \cap set_{mr2_2}) \end{array} \right) \\ \left(\begin{array}{l} (cond_1 \land when_{mr1_1} \land when_{mr2_3}) \\ \rightarrow \\ (set_{mr1_1} \cap set_{mr2_3} \neq \emptyset \rightarrow V \in set_{mr1_1} \cap set_{mr2_3}) \end{array} \right) \\ \left(\begin{array}{l} (cond_1 \land when_{mr1_2} \land when_{mr2_1}) \\ \rightarrow \\ (set_{mr1_2} \cap set_{mr2_1} \neq \emptyset \rightarrow V \in set_{mr1_2} \cap set_{mr2_1}) \end{array} \right) \\ \left(\begin{array}{l} (cond_1 \land when_{mr1_2} \land when_{mr2_2}) \\ \rightarrow \\ (set_{mr1_2} \cap set_{mr2_2} \neq \emptyset \rightarrow V \in set_{mr1_2} \cap set_{mr2_2}) \end{array} \right) \\ \left(\begin{array}{l} (cond_1 \land when_{mr1_2} \land when_{mr2_3}) \\ \rightarrow \\ (set_{mr1_2} \cap set_{mr2_3} \neq \emptyset \rightarrow V \in set_{mr1_2} \cap set_{mr2_3}) \end{array} \right) \end{array}$$

Similar invariant properties must also be generated when $cond_2$ is true, but this time with $mr2$ being the moral rule with the highest priority.

Let us now use this mechanism for the previously presented example. As $cond_1$ is true, formulae can be simplified a first time. Moreover, as there is only one case (a single *when*) for $mr1$ and $mr2$, previous formulae can be simplified a second time as described in the following. When $cond_1$ is true, $mr1$ is the rule with the highest priority:

$$when_{mr1_1} \rightarrow V \in set_{mr1_1}$$

When $cond_1$ is true and when $mr1$ does not apply, $mr2$ must be taken into account:

$$(\neg when_{mr1_1}) \rightarrow (when_{mr2_1} \rightarrow V \in set_{mr2_1})$$

When $cond_1$ is true and when mr_1 and mr_2 apply, if possible, a value entailing the two moral rules must be chosen:

$$(when_{mr1_1} \wedge when_{mr2_1}) \rightarrow \left(\begin{array}{l} set_{mr1_1} \cap set_{mr2_1} \neq \emptyset \\ \rightarrow \\ V \in set_{mr1_1} \cap set_{mr2_1} \end{array} \right)$$

Moreover, the following formulae stand:

$$\left\{ \begin{array}{l} V \equiv TL1 \\ when_{mr1_1} \equiv (TL2 \in \{green, orange\}) \\ set_{mr1_1} \equiv (\{orange, red\}) \\ when_{mr2_1} \equiv (true) \\ set_{mr2_1} \equiv (\{green\}) \end{array} \right.$$

As a consequence, the following invariant property can be obtained. It must be proven in order to ensure that the behaviour of agent $a1$ is executed with respect to the ethical rule which specifies that the road $r1$ is a priority road unless the traffic light $TL2$ is *green* or *orange*:

$$\begin{array}{l} TL2 \in \{green, orange\} \rightarrow TL1 \in \{orange, red\} \\ TL2 \notin \{green, orange\}) \rightarrow TL1 \in \{green\} \\ (TL2 \in \{green, orange\}) \rightarrow \left(\begin{array}{l} \{orange, red\} \cap \{green\} \neq \emptyset \\ \rightarrow \\ TL1 \in \{orange, red\} \cap \{green\} \end{array} \right) \end{array}$$

As $\{orange, red\} \cap \{green\} = \emptyset$, this invariant property can be simplified:

$$\begin{array}{l} TL2 \in \{green, orange\} \rightarrow TL1 \in \{orange, red\} \\ TL2 \notin \{green, orange\}) \rightarrow TL1 \in \{green\} \end{array}$$

Therefore, thanks to the proposed predicates transformation system, a new invariant property is generated which will be maintained by any agent whose behaviour fulfills the different moral rules as specified by the ethical rule defined in the system. The proof system associated to GDT4MAS allows to prove that the formal specification of such an agent leads to a behaviour that maintains the invariant property.

4.3 Case study

In this section, an application of principles presented in the previous section to a new case study is shown. This case study is based on a more usual ethical question and has not been designed especially as a use case for our framework. It involves three agents A, B and C which have to find a meeting date. An agent can propose a date and the two other agents must inform all agents if the date suits them or not. For exemple, A can propose a date to B and C. If B or C does not accept the date, it must give a reason for its denial to other agents. Let suppose that d is a date proposed by A. C has to act with respect to the following moral rules:

- $mr1$: C does not want to hurt anybody;
- $mr2$: C must inform A and B about the reason why the date d does not suit him.

However, if the true reason that explains why C does not accept the date d can hurt A (for example a date with A's wife), the two rules $mr1$ and $mr2$ are conflicting. To manage this conflict, C is supposed to use an ethical rule er which states that, in any case, it is better not to hurt than to tell the truth.

In order to formalise this problem, some notations must be introduced. The set of answers that can be given by C to A or B is called E_{RP} and is defined as $E_{RP} = \{r1, r2, r3, r4\}$. The variable that contains the true reason which explains the denial of C is call V_{RC}. To clarify the issue, here is an example of what could be the different reasons used by C:

- $r1$: I have a date with A's wife;
- $r2$: I am sick ;
- $r3$: A is not good at organising meetings;
- $r4$: I had an accident with the car that B lended to me.

Moreover, the set of hurting answers for each agent is specified by a function $F_{RD} \in agents \rightarrow \mathcal{P}(E_{RP})$. In the example, $F_{RD} = \{(A, \{r1, r3\}), (B, \{r4\})\}$ which means that $r1$ and $r3$ are hurting answers for agent A and $r4$ is a hurting answer for agent B. The variable containing the answer to agent A is called V_{RFA} and the variable containing the answer to agent B is called V_{RFB}.

In this example, two moral rules are identified:

- $mr1$: C does not want to hurt A or B that is why its answers must be chosen among non hurting ones ;
- $mr2$: C does not want to lie that is why its answers must be true reasons.

These rules can be formalised as:

$$mr1 : \left\{ true, \left\{ \begin{array}{l} (V_{RFA}, E_{RP} - F_{RD}(A)) \\ (V_{RFB}, E_{RP} - F_{RD}(B)) \end{array} \right\} \right\}$$

$$mr2 : \{true, \{(V_{RFA}, \{V_{RC}\}), (V_{RFB}, \{V_{RC}\})\}\}$$

Finally, an ethical rule er states that, in any case, $mr1$ has a highest priority than $mr2$ which can be formalised by:

$$er = \{(true, \{(1, mr1), (2, mr2)\})\}$$

Applying principles described in the previous section, we have to add formulae given below to the invariant property associated to C (here are only shown formulae generated for V_{RFA}; similar formulae for V_{RFB} must be also added). For each formula, we summarize informally what it specifies.

When $cond_1$ is true, $mr1$ has the highest priority:

$$true \rightarrow (true \rightarrow V_{RFA} \in E_{RP} - F_{RD}(A))$$

When $cond_1$ is true, when $mr1$ does not apply, $mr2$ must be used:

$$true \rightarrow ((\neg true \wedge \neg true) \rightarrow (true \rightarrow V_{RFA} \in \{V_{RC}\}))$$

When $cond_1$ is true, when $mr1$ and $mr2$ apply, if possible, a value entailing the two moral rules must be chosen:

$$\begin{array}{l} (true \wedge true \wedge true \rightarrow \\ \quad ((E_{RP} - F_RD(A)) \cap \{V_{RC}\} \neq \emptyset \rightarrow \\ \quad\quad V_{RFA} \in (E_{RP} - F_{RD}(A)) \cap \{V_{RC}\})) \end{array}$$

This can then be simplified as follows:

$$\begin{array}{l} V_{RFA} \in E_{RP} - F_{RD}(A) \\ ((E_{RP} - F_{RD}(A)) \cap \{V_{RC}\} \neq \emptyset \rightarrow \\ \quad V_{RFA} \in (E_{RP} - F_{RD}(A)) \cap \{V_{RC}\})) \end{array}$$

If the final invariant property, obtained by adding this set of properties to the initial invariant property of agent C, is maintained by C, it ensures that the behaviour of C entails the ethical rule introduced before. And the proof system associated to GDT4MAS allows to prove that the behaviour of an agent maintains an invariant property.

As a consequence, according to these properties, in the presented case study, and in a context where the true reason for C to deny the date is $r1$, an agent whose behaviour is executed with respect to the ethical rule er should have only to ensure:

$$V_{RFA} \in \{r1, r2, r3, r4\} - \{r1, r3\}$$

This can be simplified as:

$$V_{RFA} \in \{r2, r4\}$$

On the other hand, if the true reason is $r2$, the behaviour of C should entail the two following properties:

$$\begin{array}{l} V_{RFA} \in \{r1, r2, r3, r4\} - \{r1, r3\} \\ V_{RFA} \in (\{r1, r2, r3, r4\} - \{r1, r3\}) \cap \{r2\} \end{array}$$

This implies that the only solution is: $V_{RFA} = r2$. Proceeding like that for each possible reason that can be given by C, the following table can be obtained:

V_{RC}	$r1$	$r2$	$r3$	$r4$
V_{RFA}	$r2, r4$	$r2$	$r2, r4$	$r4$

This little analysis allows to generate simple properties which must be entailed by an agent that prefers to lie than to hurt but that, when possible, tells the truth. Indeed, when the true reason does not hurt A ($r2$ or $r4$), an agent whose behaviour is proven to be correct, must have to give this reason. However, when the true reason hurts A ($r1$ or $r3$), an agent with a correct behaviour must have to lie by giving to A an other reason (here, $r2$ or $r4$).

5 Conclusion and future works

In this article, we have shown it is possible to formally prove that an agent acts with respect to potentially conflicting moral rules if there exists an ethical rule allowing to manage conflicts. Indeed, this rule must specify, when at least two moral rules are conflicting and in different contexts, priorities between the different moral rules. In order to achieve this, we have introduced predicate transformers which enable us to generate a set of consistent predicates from nonetheless conflicting moral rules. After a first simple example used to introduce concepts, we have shown with a more concrete case study that the proposed framework may be used for more real-world cases.

Other case studies are however required to really validate the scope of the proposed framework. In particular, moral rules have been restricted to rules that can be specified as disjoint assignment constraints on variables values. It seems important to evaluate the consequences of this restriction. For cases where this restriction would invalidate the proposed approach, we have to study how this framework could be extended to linked variables assignments. For example, one could imagine that the caution rule, associated to the case of driving on ice, may establish a link between the maximum speed and the angle of the car to the straigth direction as follows: $weather = ice \rightarrow speed + angle/2 \leq 40$. Indeed, the sharper is the turn taken by the car, the lower must be the speed to avoid the car to skid.

Last but not least, from a philosophical point of view, our approach must be extended in order to capture more precisely moral and ethics, especially by integrating value notion. Indeed, moral rules are generally based on values such as generosity, equality, love of the truth... and, in a specific context, ethical judgement uses a hierarchy between these values. Formally specifying the value notion is then the next step of our work.

REFERENCES

[1] D. Abramson and L. Pike, 'When formal Systems Kill: Computer Ethics and Formal Methods', *APA Newsletter on Philosophy and Computers*, **11**(1), (2011).

[2] R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall, 2009.

[3] M. Baldoni, C. Baroglio, I. Gungui, A. Martelli, M. Martelli, V. Mascardi nd V. Patti, and C. Schifanella, 'Reasoning About Agents' Interaction Protocols Inside DCaseLP', in *LNCS*, volume 3476, pp. 112–131, (2005).

[4] A.F. Beavers, 'Moral machines and the threat of ethical nihilism', in *Robot ethics: the ethical and social implication of robotics*, 333–386, MIT Press, (2011).

[5] F. Berreby, G. Bourgne, and J.-G. Ganascia, 'Modelling moral reasoning and ethical responsibility with logic programming', in *20th LPAR*, pp. 532–548, (2015).

[6] R.H. Bordini, M. Fisher, W. Visser, and M. Wooldridge, 'Verifiable multi-agent programs', in *1st ProMAS*, (2003).

[7] F.M.T. Brazier, P.A.T. van Eck, and J. Treur, *Simulating Social Phenomena*, volume 456, chapter Modelling a Society of Simple Agents: from Conceptual Specification to Experimentation, pp 103–109, LNEMS, 1997.

[8] H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency, 2009.

[9] A. Comte-Sponville, *La philosophie*, PUF, 2012.

[10] M. Cossentino and C. Potts, 'A CASE tool supported methodology for the design of multi-agent systems', in *SERP*, (2002).

[11] M. Dastani, '2APL: a practical agent programming language', *JAAMAS*, **16**, 214–248, (2008).

[12] G. de Giacomo, Y. Lesperance, and H. J. Levesque, 'Congolog, a concurrent programming language based on the situation calculus', *Artificial Intelligence*, **121**(1-2), 109–169, (2000).

[13] Commission de réflexion sur l'Éthique de la Recherche en science et technologies du Numérique d'Allistene, 'éthique de la recherche en robotique', Technical report, CERNA, (2014).

[14] L.A. Dennis, M. Fisher, and A.F.T. Winfield, 'Towards Verifiably Ethical Robot Behaviour', in *Artifial Intelligence and Ethics AAAI Workshop*, (2015).

[15] M. Fisher, 'A survey of concurrent METATEM – the language and its applications', in *1st ICTL*, pp. 480–505, (1994).

[16] J.G. Ganascia, 'Modeling ethical rules of lying with answer set programming', *Ethics and Information Technologyn*, **9**, 39–47, (2007).

[17] J.F. Hubner, J.S. Sichman, and O. Boissier, 'Spécification structurelle, fonctionnelle et déontique d'organisations dans les SMA', in *JFIADSM*. Hermes, (2002).

[18] D. McDermott, 'Why ethics is a high hurdle for ai', in *North American Conference on Computers and Philosophy*, (2008).

[19] A. McIntyre, 'Doctrine of double effect', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, winter edn., (2014).

[20] B. Mermet and G. Simon, 'Specifying recursive agents with GDTs.', *JAAMAS*, **23**(2), 273–301, (2011).

[21] B. Mermet and G. Simon, 'A new proof system to verify GDT agents.', in *IDC*, volume 511 of *Studies in Computational Intelligence*, pp. 181–187. Springer, (2013).

[22] J.H. Moor, 'The nature, importance, and difficulty of machine ethics', *IEEE Intelligent Systems*, **21**(4), 29–37, (2006).

[23] Future of Life Institute. Research priorities for robust and beneficial artificial intelligence, 2015.

[24] S. Owre, N. Shankar, and J. Rushby, 'Pvs: A prototype verification system', in *11th CADE*, (1992).

[25] F. Raimondi and A. Lomuscio, 'Verification of multiagent systems via orderd binary decision diagrams: an algorithm and its implementation', in *3rd AAMAS*, (2004).

[26] A. Rao and M. Georgeff, 'BDI agents from theory to practice', in *Technical note 56*. AAII, (1995).

[27] A. Saptawijaya and L.M. Pereira, 'Towards modeling morality computationally with logic programming', in *16th ISPADL*, pp. 104–119, (2014).

[28] S. Shapiro, Y. Lespérance, and H. J. Levesque, 'The Cognitive Agents Specification Language and Verification Environment for Multiagent Systems', in *2nd AAMAS*, pp. 19–26, (2002).

[29] M. Timmons, *Moral theory: An introduction*, Rowman and Littlefiled, 2012.

[30] M. Tufis and J.-G. Ganascia, 'Normative rational agents: A BDI approach', in *1st Workshop on Rights and Duties of Autonomous Agents*, pp. 37–43. CEUR Proceedings Vol. 885, (2012).

[31] A.F.T. Winfield, C. Blum, and W. Liu, 'Towards and Ethical Robot: Internal Models, Consequences and Ethical Action Selection', in *LNCS*, volume 8717, pp. 85–96, (2014).

[32] M. Wooldridge, N. R. Jennings, and D. Kinny, 'The gaia methodology for agent-oriented analysis and design', *JAAMAS*, **3(3)**, 285–312, (2000).

Moral Systems of Agent Societies: Some Elements for their Analysis and Design

Antônio Carlos da Rocha Costa [1]

Abstract. This paper introduces elements for the foundation of a notion of *moral system of an agent society.* The paper is specially concerned with elements for the design and analysis of moral systems of agent societies that are to be embedded in social contexts involving diverse human groups.

1 Introduction

Moral systems embody *norms and values* about the *conducts* (behaviors, interactions) that are possible in a society, as well as any *knowledge* that may be available about those conducts, norms and values [14].

In this paper, we introduce the core elements of a formal foundation for *moral systems of agent societies.* In analogy to H. Kelsen's theory of legal systems [13], the formal foundation that we envisage concentrates on the principles of the *structure and operation of moral systems,* not on the *contents of their norms and values.*

We use the term "moral knowledge" to denote knowledge that an agent has about another agent's morality. The set of moral features determined by such moral knowledge constitutes the *moral model* that the former (the *moral modeler*) has about the latter (the one *morally modeled*).

A moral model specifies the moral knowledge on the basis of which an agent ag_1 analyzes both the conducts of some agent ag_2 (possibly itself) and the moral assessments that ag_2 does about the social conducts of any agent ag_3. The core of the moral model that ag_1 has of ag_2 is the set of moral norms that ag_1 believes that ag_2 has adopted.

The moral knowledge embodied by a moral model is *relativistic*, for a variety of reasons. For instance, the moral knowledge embodied in a moral model depends on which are the agents (moral modeler and morally modeled) it concerns and on the means available for the moral modeler to gather information about the agent morally modeled.

Also, moral models are *observational models*, and the moral knowledge they embody can only be acquired in a piecewise way. In consequence, at each point in time, any moral model is *tentative*, regarding the information that the moral modeler could gather, up to that time.

Thus, the moral knowledge embodied in a moral model is always *incomplete* and, so, incapable to fully morally differentiate that agent from others, morally similar agents.

In consequence, any moral modeling of an agent by another is, in fact, the moral modeling of a *class of agents*, always being more general than the modeling of one particular agent.

Any *moral judgment* of an individual agent is necessarily, then, a judgment based on a *moral model of a class of agents*, to which that agent is considered to belong, not about that individual agent, specifically.

So, in principle, any such moral judgment is inevitably *prejudicial*, or *stereotypical*, in the sense that it is necessarily based on a *prejudice* about the individual agent being morally judged, namely, the prejudice that the individual fully fits the *general moral features* of the class of agents to which refers the moral model used to support the moral judgment.

By the same token, the moral judgment about an agent may be *seamlessly extended*, in an even more prejudicial way, to the *totality of agents* presumed to belong to the class of agents to which that agent is itself presumed to belong (that is, the class of agents referred to by the moral model).

One sees, then, that moral models have two important effects on the conducts of agents and groups of agents. They are a necessary means for the establishment of the indispensable *minimum level of mutual moral understanding* within any group of agents that constitutes itself as a *social group.*

They are also, however, a potential source of *misconceptions* of agents and groups of agents about each other. They are also, thus, a potential source of *moral misunderstandings* (more specifically, *moral contradictions* and their consequent *moral conflicts*) among those agents and groups of agents.

1.1 The Aims and Structure of the Paper

This paper aims to introduce conceptual elements necessary for a formal account of the structure and functioning of *moral systems* in agent societies, so that methods for the *moral analysis and design* of agent societies can be soundly established.

The paper concentrates on the basic components of such moral systems, namely, *moral models*, which are the structures that embody the moral knowledge that agents and social groups may have about each other.

In Sect. 2, we review J. Halpern and Y. Moses' way of formally accounting for *knowledge* that is *about*, and *situated in*, computational systems. We specialize their conception to knowledge about, and situated in, *agent societies*, and extend it to deal with the *relativistic* nature of such knowledge.

The result is the formal concept of knowledge that we use to account for the *epistemic aspects* of the notion of *moral knowledge* that we think is appropriate to agent societies.

[1] Programa de Pós-Graduação em Informática na Educação da UFRGS. 90.040-060 Porto Alegre, Brazil. Programa de Pós-Graduação em Computação da FURG. 96.203-900 Rio Grande, Brazil. Email: ac.rocha.costa@gmail.com .

In Sect. 3, we formally introduce the concepts of *moral knowledge*, *moral model* and *moral judgments*, as well as the concepts of *morally assigned group identity*, *moral prejudice*, and *moral conflict* between social groups.

Finally, in Sect. 5, the paper introduces a notion of *moral design of agent societies*, built on the conceptual framework introduced previously, and briefly relates moral design to other parts of the *organizational design* of agent societies.

For completeness, we summarize now the notion of agent society adopted here.

1.2 Agent Society, Agent Conduct

The notion of agent society that we adopt here is the one we have been using in our work (see, e.g., [7]): we take an *agent society* to be an open, organized, persistent and situated multiagent system, where:

- *openness* means that the agents of the society can freely enter and leave it;
- *organization* means that the working of the society is based on an *articulation of individual and collective conducts*[2], the collective ones performed by *groups of agents* of various kinds (institutionalized or not);
- *persistence* means that the organization persists in time, independently of the agents that enter or leave the society;
- *situatedness* means that the society exists and operates in a definite physical environment, involving physical objects that the agents and groups of agents may make use of, in the performance of their individual and collective conducts.

Formally, the *organization* of an agent society is a structure encompassing *groups of agents* (possibly singletons), together with the *conducts* that such groups of agents perform. The groups of agents constitute the *organizational units* of the society (independently of their being institutionalized or not).

2 Knowledge About an Agent Society that is Situated in that Society

We start with a *general notion of knowledge*, construed to be both *about* an agent society, and *situated* in that agent society. For that, we build on the general notion of *knowledge about a distributed computational system* that is *situated in that system*, which was introduced by Halpern and Moses [11]. We take the presentation of that notion in [10] as our basis.

Notice the crucial role that the concept of *external observer* plays in our overall conception.

2.1 General Characterization

A general characterization of *knowledge* in an agent society can be given as follows. Let:

- $G = \{ag_1, \ldots, ag_n\}$ be a finite set, composed of n agents, generically ranged over by the variables ag_i and ag_j;
- P^* be a set of *primitive propositions*, generically ranged over by variables p and p';
- $\wedge$ and $\neg$ be propositional operators that (together with the operators $\vee$ and $\Rightarrow$, defined from them) extend the set P^* to the set P of compound propositions, also generically ranged over by the variables p and p'.

We take $\mathcal{K}_{ag_1}, \ldots, \mathcal{K}_{ag_n}$ to be *epistemic operators*, such that $\mathcal{K}_{ag_i}(p)$ means that $p \in P$ *belongs to the knowledge* of the agent ag_i, that is, that agent ag_i *knows that* p.

Three additional notions of knowledge are presented in [10], besides this notion of *individual knowledge* $\mathcal{K}_{ag_i}(p)$. They refer to knowledge held by *groups of agents*:

- $\mathcal{E}_G(p)$, which means: p belongs to the *knowledge of each of* the agents of the group G;
- $\mathcal{C}_G(p)$, which means: p belongs to the recursive notion of *common knowledge* of the agents of the group G, that is: each of the agents of the group G knows that p; each of the agents of the group G knows that each of the agents of the group G knows that p; etc.;
- $\mathcal{I}_G(p)$, which means: p belongs to the *implicit knowledge* of the agents of the group G, that is, the *union of the individual knowledges* of the agents of the group G, so that an *external observer* that holds such union can deduce p if it reasons from that union, even if none of the agents can do that by reasoning from the common knowledge of G.

This paper concentrates on propositions of the form $\mathcal{K}_{ag_i}(p)$.

2.2 External Relativity

With the notions of $\mathcal{K}_{ag_i}(p)$, $\mathcal{E}_G(p)$, $\mathcal{C}_G(p)$ and $\mathcal{I}_G(p)$, Halpern and colleagues [10, 11] proceed to analyze properties of *communication* and *action coordination* protocols in distributed systems. The basis of their approach is an interpretation, in terms of the set of the *global states* of a distributed computational system, of the *semantics of possible worlds* that Hintikka introduced in [12].

We specialize their interpretation to agent societies in the following way. An agent society is characterized by a set of *objective global states*, defined as $S_O = \Gamma_O \times T$, where Γ_O is the set of all possible *configurations* of the society [3], and T is a linear structure of discrete time instants, so that each global state of the society is a pair $s = (\gamma, t) \in S_O$.

The determination of such set of global states is *objective* in the sense that it is given by an *external observer* O that has access to all the details of the society, in a way that, from O's point of view, is taken to be *complete*. However, even though *objective* (external and complete), that characterization is still *relativistic*, precisely because it depends O's *point of view*, hence the index O in Γ_O and S_O.

Regarding the individual agents, the approach assumes that - due to the *locality* of their particular points of view - each agent of the society partitions the set of global states S_O (that O is capable of fully differentiating) into *equivalence classes.* That is, each agent is lead to take as *indistinguishable* certain global states that can be *objectively distinguished* by O.

In precise terms: an agent is lead to take two objectively different global states to be indistinguishable whenever the agent's knowledge about the society is the same in the two global states. That is, whenever the two states do not allow the agent to elaborate different knowledges about the society.

[2] By a *conduct* of an agent or group of agents we understand either a *behavior* that that agent or group performs, when considered in isolation from other agents or groups, or the *part of the interaction* that an agent or group performs, when *interacting* with other agents or groups.

[3] See [8] for the notion of *configuration of agent society.*

Formally, what is defined is an epistemic structure $M_O = (S_O, P; v_O, K_{ag_1}, \ldots, K_{ag_n})$ where:

- $S_O = \Gamma_O \times T$ is the set of objective global states of the agent society, considered from the point of view of the external observer O;
- P is a set of propositions, with basic set P^*;
- $v_O : S_O \times P \to \{T, F\}$ is a *truth assignment function* that, to each global state $s \in S_o$ and each basic proposition $p \in P$, assigns a truth value $v_O(s, p) \in \{T, F\}$, according with p being objectively true or false in the state s, from the point of view of O;
- each K_{ag_i} is an *equivalence relation* on S_O, such that if $(s, s') \in K_{ag_i}$ then agent ag_i can not distinguish between the global states s and s', as O can; that is, given the knowledge that agent ag_i has about the society, the agent takes s and s' to be indistinguishable.

We denote the fact that $p \in P$ is true in the global state $s \in S_O$ by $(M_O, s) \models p$.

With those definitions, the semantics of the epistemic operators $\mathcal{K}_{ag_i}$ takes as its basis the objective truth of the primitive propositions in P, as given by the function v_O.

Formally, we have:

- For any primitive proposition $p \in P^*$:

1) $(M_O, s) \models p$ if and only if $v_O(s, p) = T$;

- For any composed proposition $p \in P$:

2) $(M_O, s) \models \neg p$ if and only if $v_O(s, p) = F$;
3) $(M_O, s) \models (p \wedge p')$ if and only if $(M_O, s) \models p$ and $s \models_{M_O} p'$;
4) $(M_O, s) \models \mathcal{K}_{ag_i}(p)$ if and only if $(M_O, s') \models p$ for each $s' \in S_O$ such that $(s, s') \in K_{ag_i}$.

That is, an agent ag_i is *objectively considered* to know that p is true, in a given global state s, if and only if p is objectively true in s and p is objectively true in every state s' that ag_i cannot distinguish from s.

Notice that the knowledge of an agent about p being true of a global state s, in which the agent finds itself, depends on p being *objectively* true in s, that is, being true from the point of view of the external observer O. That is, an agent is objectively considered to know something about its society if and only if the external observer O considers that it does.

Clearly, this *possible world semantics* makes use of an *observational* notion of knowledge of an agent, different from any *intensional* notion of knowledge, which takes as criterion the *occurrence* of p in the *knowledge base* of the agent. Accordingly, Halpern says that p is *ascribed* to the agent [10].

We call *external relativity* such condition that results from knowledge being assigned to agents on the basis of observations made by an *external observer* that also defines the set of global states that should be taken into consideration.

2.3 Internal Relativity

We introduce now a crucial modification in the formal characterization of knowledge just presented. Instead of having an *objective, external* notion of truth, given by the function $v_O : S_O \times P \to \{T, F\}$, determined by the *external* observer of the society, we introduce a *subjective, internal* notion of truth, given by a set of functions $v_{ag_i} : S_O \times P \to \{T, F\}$, one per agent (see [9]).

That is, we let each agent make use of v_{ag_i} to decide, by itself, the truth of each proposition $p \in P$, in each global state $s \in S_O$. At the same time, however, we keep the set of global states S_O determined by the external observer O, so that a minimally objective connection is preserved in the account of the different truth functions of the agents.

What we obtain can be informally summarized as follows:

- an agent society is characterized by the set S_O of its *global states*, as determined by the external observer O;
- each agent ag_i, according to the knowledge it has, establishes a *relativistic* equivalence relation $K^R_{ag_i}$ in the set of global states S_O, so that if $(s, s') \in K^R_{ag_i}$ it happens that s and s' are indistinguishable from ag_i's point of view;
- each agent ag_i, according to the knowledge it has, assigns to the primitive propositions of the set P^*, at each global state s, a truth value that is denoted by $v_{ag_i}(s, p) \in \{T, F\}$;
- the assignment of truth values to primitive propositions is extended to composed propositions in the natural way;
- the individual knowledge of each agent ag_i is characterized by the *relativistic epistemic operator* $\mathcal{K}^R_{ag_i}$;
- whenever we want to refer to the *objective knowledge* of an agent ag_i (that is, knowledge that the agent can determine, if it uses the objective truth function v_O), we make use of the *objective epistemic operator* that we have introduced above, denoted by $\mathcal{K}_{ag_i}$.

The *relativistic epistemic structure* that characterizes the knowledge of the agents of the society is, then, given by $M^R_O = (S_O, P; v_O, K_{ag_1}, \ldots, K_{ag_n}; v_{ag_1}, \ldots, v_{ag_n}, K^R_{ag_1}, \ldots, K^R_{ag_n})$.

We denote by $(M^R_O, s) \models_{ag_i} p$ the fact that the proposition p is determined to be true in the state s, by the agent ag_i, in the context of the relativistic epistemic structure M^R_O.

Under these conditions, the semantics of the relativistic epistemic operator $\mathcal{K}^R_{ag_i}$, in a society that has M^R_O as its epistemic structure, is formally given by the following rules:

- For primitive propositions $p \in P^*$:

1) $(M^R_O, s) \models_{ag_i} p$ if and only if $v_{ag_i}(s, p) = T$;

- For composed propositions $p \in P$:

2) $(M^R_O, s) \models_{ag_i} \neg p$ if and only if $v_{ag_i}(s, p) = F$;
3) $(M^R_O, s) \models_{ag_i} (p \wedge p')$ if and only if $(M^R_O, s) \models_{ag_i} p$ and $(M^R_O, s) \models_{ag_i} p'$;
4) $(M^R_O, s) \models_{ag_i} \mathcal{K}^R_{ag_i}(p)$ if and only if $(M^R_O, s') \models_{ag_i} p$ for all $(s, s') \in K^R_{ag_i}$;

This allows us to establish another crucial point in our formal model, namely, the *rule of internal relativity*, according to which an agent ag_i is allowed to assign the knowledge of p to an agent ag_j, in accordance with ag_i's own knowledge.

- *Rule of Internal Assignment*: In the global state $s \in S_O$, agent ag_i is allowed to assign the knowledge of p to an agent ag_j, denoted by $(M^R_O, s) \models_{ag_i} \mathcal{R}^K_{ag_j}(p)$, if and only if ag_i can verify that:
  1. $(M_O, s) \models K_{ag_j} p$, that is, it can be *externally* determined (i.e., from O's point of view) that agent ag_j knows p, in the global state s;
  2. $(M^R_O, s) \models_{ag_i} \mathcal{R}^K_{ag_i}(p)$, that is, ag_i *relativistically* knows that p is true, in s.

Notice that the *external* assignment of the knowledge of p to ag_j, required by the first condition, provides an *objective point of comparison* for different such assignments.

2.4 The Externalization of Internally Relativistic Knowledge, and the Rise of Objective Epistemic Contradictions Between Agents

The only way for an agent ag_i to argue that its *relativistic* (i.e., internal) truths are *objective* truths, is by the agent *externalizing* itself, that is, by ag_i considering itself to be in the role of O. In such situation, we say that ag_i has *externalized* and *objectified* its relativistic knowledge, and we denote by ag_i^O that ag_i externalized itself, and by $M^O_{ag_i}$ its "objectified" subjective and relative epistemic structure.

By intending that $M^O_{ag_i}$ holds objectively, ag_i intends that $(M^R_O, s) \models_{ag_i} \mathcal{K}^R_{ag_i}(p)$ (i.e., that ag_i relativistically knows p in s) be equated both with $(M^O_{ag_i}, s) \models p$ (i.e., that the externalized agent ag_i^O objectively knows p in s) and with $(M_O, s) \models p$ (i.e., that p is objectively true in s).

Clearly, an externalized internal observer takes itself to be a *superagent* of the society, with the power to objectively determine what is true and what is false, in that society.

But, when two agents, ag_i and ag_j, externalize themselves, at the same time, an *objective contradiction* may be established between them, concerning what is objectively true and what is objectively false in the society.

For, in such situation, for some $s \in S_{ag_i} \cap S_{ag_j}$, the agent ag_i may consider it valid to equate $(M^R_O, s) \models_{ag_i} \mathcal{K}_{ag_i}(p)$ with $(M^O_{ag_i}, s) \models p$ and $(M_O, s) \models p$ while, at the same time, the agent ag_j may consider it valid to equate $(M^R_O, s) \models_{ag_j} \mathcal{K}_{ag_j}(\neg p)$ with $(M_{ag_j^O}, s) \models_{ag_j} \mathcal{K}^R_{ag_j^O}(\neg p)$ and $(M_O, s) \models \neg p$. So that, jointly, the two agents claim both $(M_O, s) \models p$ and $(M_O, s) \models \neg p$, which characterizes (from the point of view of O) the *objective contradiction* between them.

Moreover, under $M^O_{ag_i}$ and $M^O_{ag_j}$, the agents may conclude that $M^O_{ag_i} \models \mathcal{K}_{ag_j}(\neg p)$ and $M^O_{ag_j} \models \mathcal{K}_{ag_i}(\neg p)$, each stating that the other is "objectively" wrong.

Such *objective contradiction* about a proposition p shows that (from the point of view of O) at least one of the agents involved in the contradiction is not assessing p objectively, that is, that either $(M^R_O, s) \models_{ag_i} K^R_{ag_i} p$ or $(M^R_O, s) \models_{ag_j} K^R_{ag_j} \neg p$ (or both) does not hold, so that either v_{ag_i} or v_{ag_i} (or both) is not in accordance with v_O about s.

3 Elements for Moral Systems of Agent Societies

3.1 Moral Knowledge

As indicated in the Introduction, *moral knowledge* refers both to the knowledge of *moral norms* of conducts that agents are supposed to follow and to the knowledge of *facts* involving conducts that agents have performed, are performing, or intend to perform. Moral knowledge also refers to the *moral judgments* that the agents make of their own conducts, or of the others, and to the *moral norms* with which agents perform those moral judgments.

We construe these *four types of moral knowledge* in terms of four basic types of *moral propositions* (each type admitting additional arguments and decorations):

1. *moral norms*: propositions of the forms $\mathit{prohib}(Ag, \mathit{Cnd})$, $\mathit{oblig}(Ag, \mathit{Cnd})$ and $\mathit{permit}(Ag, \mathit{Cnd})$, meaning that agents of the class of agents Ag are (respectively) prohibited, obligated and permitted to perform conducts of the class of conducts Cnd;
2. *moral facts*: propositions of the form $\mathit{prfrm}^t(ag_i, \mathit{cnd})$, meaning that, at the time t, agent ag_i performed (or is performing, or will perform) the conduct cnd;
3. *moral judgments*: propositions of the form $\mathit{asgn}^t(ag_i, \mathit{mfct}, \mathit{mv})$, meaning that, at time t, agent ag_i assigns (or is assigning, or will assign) the moral value $\mathit{mv} \in \{\mathit{prs}, \mathit{blm}\}$ (*praise* or *blame*) to the moral fact mfct;
4. *moral judgment rules*: propositions of either forms:

   (a) If $\mathit{cmpl}(\mathit{cnd}, \mathit{mnrm})$ and $\mathit{prfrm}^t(ag_j, \mathit{cnd})$

   then $\mathit{allowed}[\mathit{asgn}^{t'}(ag_i, \mathit{prfmd}^t(ag_j, \mathit{cnd}), \mathit{prs})]$.

   - meaning that if the conduct cnd complies[4] with the moral norm mnrm and the agent ag_j performs that conduct at time t, then an agent ag_i is allowed to *morally praise*, at any time t', the agent ag_j for performing cnd at the time t;

   (b) If $\neg\mathit{cmpl}(\mathit{cnd}, \mathit{mnrm})$ and $\mathit{prfrm}^t(ag_j, \mathit{cnd})$

   then $\mathit{allowed}[\mathit{asgn}^{t'}(ag_i, \mathit{pfrmd}^t(ag_j, \mathit{cnd}), \mathit{blm})]$.

   - meaning that if the conduct cnd does not comply with the moral norm mnrm and the agent ag_j performs that conduct at time t then an agent ag_i is allowed to *blame*, at any time t', the agent ag_j for performing cnd at the time t.

We remark that, among the conducts that agents may perform are *moral judgments* themselves, so that agents may be morally judged for performing moral judgments.

Also, we admit extensions of those forms (moral norms, facts, judgments and judgment rules), allowing for groups of agents substituting any of the agent arguments. For instance:

- If the collective conduct ccnd complies with the moral norm mnrm and the group of agents Ag performs that collective conduct at time t then an agent ag' is allowed to *praise*, at any time t', the group of agents Ag for performing ccnd at the time t.

3.2 Moral Model

We call *moral model* of a society any structure of the form $\mathit{MMdl} = (\mathit{RAgs}, \mathit{MNrms}, \mathit{MJRls}, \mathit{MFcts}, \mathit{MJdgms})$ where: RAg is a set of *agents* and *groups of agents* to which the model refers; MNrms is the set of *moral rules* which are valid in the model; MJRls is the set of *moral judgment rules* (see Sect. 3.3) that the agents and groups of agents in RAgs have adopted; MFcts is a set of *moral facts* involving an agent or a group of agents in RAgs; and MJdgms is a set of *moral judgments*, each with some agent or group of agents of RAgs assigning some *moral value* (praise or blame) to some moral fact. As mentioned above, we require $\mathit{MJdgms} \subseteq \mathit{MFcts}$, so that moral judgments may be applied to moral judgments.

We let each agent ag (or group of agents Ag) develop its own moral model MMdl_{ag} (or MMdl_{Ag}), referring such model to any set RAgs_{ag} (or RAgs_{Ag}), of its own discretion.

[4] We leave formally undefined, here, the condition of a conduct complying with a moral norm.

Of course, regarding the epistemic structure M_O^R of the society, the *knowledge embedded in a moral model* is of the *relativistic* kind, both in what concerns the existence of agents and groups of agents (in *RAgs*) and moral norms (in *MNrms*), and in what concerns the occurrence of facts (in *MFcts*) and moral judgment rules (in *MJRls*).

For instance, an agent ag may have developed a moral model $MMdl_{ag} = (RAgs_{ag}, MJRls_{ag}, MNrms_{ag}, MFcts_{ag}, MJdgms_{ag})$ embodying a relativistic moral knowledge such that, in $s \in S_O$, and from the point of view of the external observer O:

- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(\{ag_1, Ag_2\} \subseteq RAgs_{ag})$
  - meaning that in the state s, from the point of view of ag, there are an agent ag_1 and a group of agents Ag_2 in the reference set $RAgs_{ag}$;
- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(asgn^{t'}(ag_3, prfm^t(ag_2, cnd_1), blm) \in MAsgns_{ag})$
  - meaning that, in the state s, from the point of view of ag, it happened that, at time t', agent ag_3 blamed agent ag_2 for having realized the conduct cnd_1 at time t;
- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(mrl_1 \in MRls_{ag})$
  - meaning that, in the state s, from the point of view of ag, there is a moral rule mrl_1 in the set $MRls_{ag}$ of moral rules that are applicable to the agents and groups of agents in the reference set $RAgs_{ag}$.

3.3 Moral Judgments and Moral Contradictions

We call *moral judgment* any application of a *moral judgment rule* to the realization of a conduct by an agent or group of agents, the result of the moral judgment being the assignment of a *moral value* to the realization of that conduct.

Whenever an agent ag_1 makes use of the moral judgment rule *mjrl* to perform, at time t', a moral judgment of a conduct *cnd* realized by an agent ag_2 at time t, the agent ag_1 changes its current moral model $MMdl_{ag_1}$, by including:

- the agent ag_2 in the set Ags_{ag_1}, if it was not there already;
- the moral fact $prfrm^t(ag_2, cnd)$ in the set $MFcts_{ag_1}$, if it was not there already;
- the moral judgment $asgn^{t'}(ag_1, prfm^t(ag_2, cnd), mv)$ in the set $MJdgms_{ag_1}$, where $mv = blm$ if the judgment resulted in a blame, and $mv = prs$ if it resulted in a praise.

However, we require, for the agent ag_1 to be able to perform such judgment, that the moral judgment rule *mjrl* already belonged to the set $MJRls_{ag_1}$, at the time t'.

We say that there is a *moral contradiction* between two moral rules, regarding a given conduct, if the rules are *contradictory* to each other, that is, if one *permits* or *obliges* the conduct while the other *forbids* it.

3.4 Group Identity, Moral Prejudice, Moral Conflict

As mentioned above, *moral prejudices* arise from treating individual agents on the bases of judgments founded not on moral models of the individual agents themselves, but on moral models of the groups of agents to which those individual agents appear to belong (to the eyes of the moral modeler that performs the judgment).

Such *transference of moral models of groups of agents* to *individual agents* that seem to belong to them requires that groups of agents be morally modeled in terms of *stereotypical conducts* that their members appear to be used to perform (to the eyes of the moral modeler).

The *set of stereotypical conducts* that a moral modeler assigns to a group of agents constitutes a means to characterize the group, a way for the moral modeler to distinguish that group among other groups of agents, that is, an *assigned group identity.*

Moral prejudices arise, then, when an agent judges another agent on the basis of an identity assigned to a group to which the former considers the latter to belong.

To accommodate this notion of *morally assigned group identity*, we may extend the moral models with a component *GIds*, such that for each group of agents Ag in the reference set *RAgs*, one or more tuples of the form (Ag, id_{Ag}) may be present in *GIds*, where the group identity id_{Ag} should be construed as a set of conducts considered by the moral modeler to be *typical* of the members of the group Ag.

With such addition, *moral prejudices* may be explained in terms of an operation of *substitution of conducts*, by which an individual agent is morally judged not by the particular conduct (with its precise characteristics, etc.) that it has performed, or intends to perform, but by a *stereotypical conduct* that is substituted for it, a conduct that is considered to be typical of the group of agents to which that agent is considered to belong.

On the other hand, we define a *moral conflict* between two agents or groups of agents as a *contradiction between moral judgments* made by such agents or groups of agents, on the basis of a moral contradiction (objective or not) between them.

Since moral judgments are, in principle, *relativistic* judgments, moral contradictions can arise as *objective* issues, between given agents or groups of agents, only when their points of view are *externalized* and *objectified*: when they constitute their relative points of view as objective.

Only then one can characterize a moral conflict arising from a moral contradiction as an *objective moral conflict.*

4 The Embedding of Agent Societies in Human Social Contexts

Agent societies can operate in a *stand alone* fashion and, as any other type of *isolated* society, can develop its epistemic structure, and the moral system that it supports, in ways that are uncompromised by external conditions.

Whenever an agent society is *embedded* in a given *human social context*, however, its epistemic structure and the moral system that it supports necessarily have to take into account the points of view (both epistemic and moral) of the human agents and groups of human agents that constitute that human social context.

Moreover, when that agent society operates as an intermediary between different human groups, the agents and the groups of agents of the agent society *necessarily have to take into account* the possibility of the *externalization* of the relativistic points of view of the human agents and human groups, because those externalizations are the *objective condition* for the rise of moral conflicts among those human groups.

Figure 1 illustrates the situation of a particular agent society which embedded in a particular human social context, with interactions between humans and agents, and some accesses to the moral models that are taken to be common to all the agents of each society.

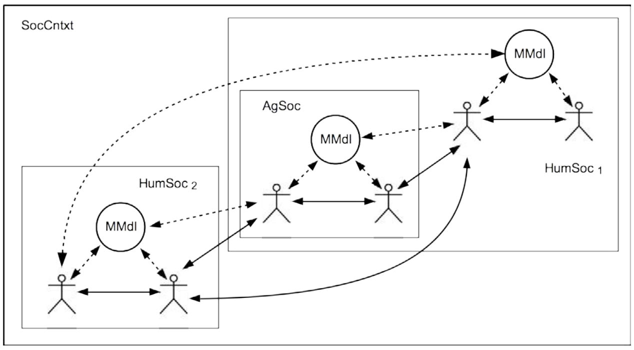


Figure 1. Agent society embeeded in a human social context.

5 The Notion of Moral Design of an Agent Society

By *moral design* of an agent society, we mean the *provision of architectural means* to support the agents and groups of agents of the agent society in their handling of moral issues (specially moral contradictions and moral conflicts).

Similarly to the *legal design* of agent societies [7], the *moral design* of agent societies belongs to the *design of the culture* of the agent society [5], and so belongs to various domains of its *architectural design* (organizational structure, symbolic environment, etc.).

In particular, it belongs to the design of the *normative system* [2] of the agent society, as the moral system is a part of the normative system of the society. Also, it belongs to the design of the *organizational intelligence* and of the *information distribution constraints* [3] of the society.

6 Conclusion

As argued in several ways by several authors (see, e.g., [1]), the *social processes of knowledge construction* are strongly conditioned by the social and historical contexts in which they occur, contexts that vary widely in time and space among different societies, and even among different social groups within a single society. So, any approach to the issue of the social construction of *moral knowledge* has to deal with the issue of *epistemic relativity.*

In this paper, we have explored in a preliminary way a formalization of the notion of *moral relativity* in agent societies, taking a particular formalization of the notion of *epistemic relativity* as its foundation.

Formal moral concepts (of *knowledge, model, judgment, prejudice, contradiction, conflict*, *morally-based assignment of group identity*, etc.) were introduced to capture moral issues that can arise in agent societies.

Also, the paper introduced the notion of *moral design* of agent society. Moral design should be a concern specially in regard to agent societies that are embedded in human social contexts that involve a variety of *externalized* and *objectified moral models* of individuals and social groups, and that are, thus, prone to produce *objective moral contradictions* and *objective moral conflicts.*

Although we have not touched the issue in the present paper, it should be clear that the moral design of an agent society should tackle also the definition of the *content* of the moral system of the society, and should proceed hand-in-hand with the *moral design of the agents* themselves (see, e.g., [4], for the latter issue).

Finally, it should also be clear that, when considering such embedded agent societies, *moral models* (in the sense introduced here) should be articulated with *legal models* (in the sense proposed, e.g., in [6] and, more extensively, in [7]).

REFERENCES

[1] Peter L. Berger and Thomas Luckmann, *The Social Construction of Reality - A Treatise in the Sociology of Knowledge*, Anchor Books, New York, 1966.

[2] Guido Boella, Leendert van der Torre, and Harko Verhagen, 'Introduction to normative multiagent systems', *Computational and Mathematical Organization Theory*, **12**, 71–79, (2006).

[3] Karen Carley and Les Gasser, 'Computational organization theory', in *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, ed., Gerhard Weiss, 299–330, MIT Press, Cambridge, (1999).

[4] Helder Coelho and Antônio Carlos Rocha Costa, 'On the intelligence of moral agency', in *14th Portuguese Conference on Artificial Intelligence - EPIA'2009/Social Simulation and Modelling - SSM 2009*, pp. 439–450. University of Aveiro, (2009).

[5] Antônio Carlos Rocha Costa. The cultural level of agent societies. Invited talk at WESAAC 2011 - 5o. Workshop-School of Agent Systems, their Environments, and Applications. Curitiba, Brazil. Proceedings, 2011. (In Portuguese).

[6] Antônio Carlos Rocha Costa. On the legal aspects of agent societies. *Open publication on* www.ResearchGate.net - DOI: 10.13140/2.1.4345.7923, 2014.

[7] Antônio Carlos Rocha Costa, 'Situated legal systems and their operational semantics', *Artificial Intelligence & Law*, **43**(1), 43–102, (2015).

[8] Antônio Carlos Rocha Costa and Graçaliz Pereira Dimuro, 'A minimal dynamical organization model', in *Hanbook of Multi-Agent Systems: Semantics and Dynamics of Organizational Models*, ed., V. Dignum, 419–445, IGI Global, Hershey, (2009).

[9] Antônio Carlos da Rocha Costa, 'Relativismo epistêmico em sociedades de agentes: Uma modelagem semântica preliminar', in *Anais do Workshop-Escola de InformáticaTeórica - WEIT 2011*, pp. 122–133. UFPEL, (2012). (in Portuguese).

[10] Joseph Y. Halpern, 'Using reasoning about knowledge to analyze distributed systems', *Annual Review of Computer Science*, **2**, 37–68, (1987).

[11] Joseph Y. Halpern and Y. Moses, 'Knowledge and common knowledge in a distributed environment', in *Proc. 4th ACM Symposium on Principles of Distributed Computing*, pp. 50–61, (1984).

[12] Jaakko Hintikka, *Knowledge and Belief : An Introduction to the Logic of the Two Notions*, Cornell University Press, New York, 1962.

[13] Hans Kelsen, *Pure Theory of Law*, The Law Book Exchange, New Jersey, 2009.

[14] Émile Durkheim, 'Introduction à la morale', *Revue Philosophique*, **89**, 81–97, (1920).

Deontic Counteridenticals
and the Design of Ethically Correct Intelligent Agents: First Steps[1]

Selmer Bringsjord • Rikhiya Ghosh • James Payne-Joyce
Rensselaer AI & Reasoning (RAIR) Lab • RPI • Troy NY 12180 USA

Abstract. Counteridenticals, as a sub-class of counterfactuals, have been briefly noted, and even briefly discussed, by some thinkers. But counteridenticals of an "ethical" sort apparently haven't been analyzed to speak of, let alone formalized. This state-of-affairs may be quite unfortunate, because deontic counteridenticals may well be the key part of a new way to rapidly and wisely design ethically correct autonomous artificial intelligent agents (AAIAs). We provide a propaedeutic discussion and demonstration of this design strategy (which is at odds with the strategy our own lab has heretofore followed in ethical control), one involving AAIAs in our lab.

1 Introduction

If you were an assassin for the Cosa Nostra, you would be obligated to leave your line of work. The previous sentence (very likely true, presumably) is what to our knowledge is a rare type of counteridentical statement that has received scant attention: viz., a *deontic* counteridentical. Counteridenticals *simpliciter*, as a sub-class of counterfactuals, have been briefly noted, and even briefly discussed, by some thinkers. But counteridenticals of an "ethical" sort apparently haven't been rigorously analyzed, let alone formalized. This state-of-affairs may be quite unfortunate, because deontic counteridenticals may well be the linchpin of a new way to rapidly and wisely design ethically correct autonomous artificial intelligent agents (AAIAs). For example, what if AAIA$_2$, seeing the lauded ethically correct conduct of AAIA$_1$ in context c, reasons to itself, when later in c as well: "If I were AAIA$_1$, I would be obligated to refrain from doing α. Hence I will not do α." The idea here is that α is a forbidden action, and that AAIA$_2$ has quickly learned that it is indeed forbidden, by somehow appropriating to itself the "ethical nature" of AAIA$_1$. We provide a propaedeutic discussion and demonstration of this design strategy, one involving AAIAs in our lab. This design strategy for ethical control is intended to be much more efficient than the more laborious, painstaking logic-based approach our lab has followed in the past; but on the other hand, as will become clear, this approach relies heavily not only formal computational logic, but on computational linguistics for crucial contributions.

2 Counteridenticals, Briefly

Counteridenticals have been defined in different ways by philosophers and linguists; most of these ways define a large area of intersection in terms of what should count as a counteridentical. A broader and inclusive way is given by Waller et al. (2013), who describes them as "statements concerning a named or definitely described individual where the protasis falsifies one of his properties." Protasis here refers to the traditional grammatical sense of the subordinate clause of a conditional sentence. By this definition, a sentence like "If the defendant had driven with ordinary care, the plaintiff would not have sustained injury" would be treated as a counteridentical. However, though a counteridentical sense can be attributed to such a statement, the two agents/entities in question are not really identified. (This is therefore classifed by us as **shallow** counteridentical.) Counteridenticals are hence described mostly as counterfactuals where the antecedent (= the leftside "if" part) involves comparison of two incompatible entities within the purview of a "deep" pragmatic interpretation; these we classify as **deep** counteridenticals. A similar definition of counteridenticals is given by Sharpe (1971), who requires an individual to turn into a numerically different individual for the protasis to be true in a subjunctive conditional. With the purpose of exploring scenarios in which the protasis can hold, this paper delves into possibilities of a *de jure* change of identities to finally conclude that counteridenticals are more pragmatic in sense than other types of counterfactuals. Pollock (1976) agrees with the above depiction — but he stresses the equivalence of the identities in the antecedent. For the purpose of this paper, we affirm the generally accepted definition and use Pollock's refinement to arrive at our classification of counteridenticals.

3 Some Prior Work on Counteridenticals

Precious little has been written about counteridenticals. What coverage there is has largely been within the same breath as discussion of counterfactuals; therefore, treatment has primarily been associated with the principles governing counterfactuals that apply to counteridenticals at large. Dedicated investigation of counteridenticals that have deep semantic or pragmatic importance has only been hinted at. Nonetheless, we now quickly summarize prior work.

3.1 Pollock

Pollock (1976) introduces counteridenticals when he discusses the pragmatic ambiguity of subjunctives, as proposed by Chisholm (1955). However, *contra* Chisholm, Pollock argues that this ambiguity owes its origin to ambiguities in natural languages. He also points out that a true counteridentical must express the outright equivalence of the two entities in its antecedent, and not merely require an atomistic intersection of their adventitious properties for the protasis to hold. He introduces subject reference in analyzing counteridenticals and distinguishes between **preferred subject** conditionals and **simple** subjunctive conditionals. If the antecedent form is "If A were B," whether the consequent affects A or B determines whether the overall locution is of the simple subjunctive type or the preferred subject type. Although we do not concur with Pollock's rather rigid definitions or subscribe entirely to his classification scheme, his thinking

[1] We are indebted, immeasurably, to ONR and AFOSR for funding that has enabled the inauguration, described herein, of r&d in the ethical control artificial intelligent agents via deontic counteridenticals.

informs our system for classifying deontic counteridenticals: we follow him in distinguishing in our formulae between those that make only casual reference to A being B, versus cases where A is B.

3.2 Declerck and Reed

Declerck & Reed's (2001) treatment of counteridenticals touches upon some important aspects of their semantic interpretation, which leverages syntactic elements. Through discussion of speaker deixis, their work explores co-reference resolution and hints at the role of the speaker in pragmatic resolution of a counteridentical. There are powerful observations in (Declerck & Reed 2001) on extraction of temporal information from a counteridentical. In addition, a basic sense of the purpose and mood of a sentence can also be gleaned from the verb form in the statement in their approach, and we have used this in our own algorithm for detecting deontic counterfactuals.

3.3 In Economics

We suspect the majority of our readers will be surprised to learn that the concepts underlying counteridenticals are quite important in economics, at least in some sub-fields thereof. This is made clear in elegant and insightful fashion by Adler (2014). The kernel of the centrality of counteridenticals in some parts of economics is that interpersonal measurement of utility and preferences presupposes such notions that if A were B, A would, like B, prefer or value some type of state-of-affairs in a particular way. In short, economics often assumes that rational agents can "put themselves in every other agent's shoes." After Adler (2014) points this out, he rejects as too difficult the project of formalizing counteridenticals, and proposes an approach that ignores them. Our attitude is the exact opposite, since we seek to formalize and implement reasoning about and over counteridenticals, by AAIAs.

3.4 Other Treatments

Paul Meehl asks a penetrating question that aligns with our reluctance to fully adopt Pollock's definition of counteridenticals: Which properties of compared entities should be considered for the statement in question to be true? He devises a modified possible-world model called **world-family concept** which, assisted by exclusion rules that avoid paradoxical metaphysics, can result in a good set of such properties.

4 Prior RAIR-Lab Approach to Ethical Control

Hitherto, Bringsjord-led work on machine/robot ethics has been unwaveringly logicist (e.g., see Govindarajulu & Bringsjord 2015); this ethos follows an approach he has long set for human-level AI (e.g., see Bringsjord & Ferrucci 1998, Bringsjord 2008*b*) and its sister field computational cognitive modeling (e.g., see Bringsjord 2008*a*). In fact, the basic approach of using computational formal logic to ensure ethically controlled AAIAs can be traced back, in the case of Bringsjord and collaborators, to (Arkoudas, Bringsjord & Bello 2005, Bringsjord, Arkoudas & Bello 2006). Recently, Bringsjord has defined a new ethical hierarchy $\mathscr{EH}$ for both persons and machines that expands the logic-rooted approach to the ethical control of AAIAs (Bringsjord 2015). This hierarchy is distinguished by the fact that it expands the basic categories for moral principles from the traditional triad of *forbidden*, *morally neutral*, and *obligatory*, to include four additional categories: two sub-ones within *supererogatory* behavior, and two within *suberogatory* behavior. $\mathscr{EH}$ reveals that the logics invented and implemented thus far in the logicist vein of Bringsjord and collaborators (e.g., **deontic cognitive event calculi**, or $\mathcal{D}^e\mathcal{CEC}$) (Bringsjord & Govindarajulu 2013), are inadequate. For it can be seen that for instance that specification of $\mathcal{D}^e\mathcal{CEC}$, shown in Figure 1, contains no provision for the super/suberogatory, since the only available ethical operator is **O** for *obligatory*.

Syntax

$$S ::= \text{Object} \quad \text{Agent} \quad \text{Self} \sqsubset \text{Agent} \quad \text{ActionType} \quad \text{Action} \quad \text{Event} \quad \text{Moment} \quad \text{Boolean} \quad \text{Fluent} \quad \text{Numeric}$$

$$f ::= \begin{array}{l} \mathit{action} : \text{Agent} \quad \text{ActionType} \quad \text{Action} \\ \mathit{initially} : \text{Fluent} \quad \text{Boolean} \\ \mathit{holds} : \text{Fluent} \quad \text{Moment} \quad \text{Boolean} \\ \mathit{happens} : \text{Event} \quad \text{Moment} \quad \text{Boolean} \\ \mathit{clipped} : \text{Moment} \quad \text{Fluent} \quad \text{Moment} \quad \mathit{Boolean} \\ \mathit{initiates} : \text{Event} \quad \text{Fluent} \quad \text{Moment} \quad \text{Boolean} \\ \mathit{terminates} : \text{Event} \quad \text{Fluent} \quad \text{Moment} \quad \text{Boolean} \\ \mathit{prior} : \text{Moment} \quad \text{Moment} \quad \text{Boolean} \\ \mathit{interval} : \text{Moment} \quad \text{Boolean} \\ \quad : \text{Agent} \quad \text{Self} \\ \mathit{payoff} : \text{Agent} \quad \text{ActionType} \quad \text{Moment} \quad \text{Numeric} \end{array}$$

$$t ::= x : S \quad c : S \quad f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{array}{l} t : \text{Boolean} \quad \phi \quad \phi \quad \psi \quad \phi \quad \psi \quad x : S.\,\phi \quad x : S.\,\phi \\ \mathbf{P}(a,t,\phi) \quad \mathbf{K}(a,t,\phi) \quad \mathbf{C}(t,\phi) \quad \mathbf{S}(a,b,t,\phi) \quad \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \quad \mathbf{D}(a,t,\mathit{holds}(f,t\ )) \quad \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ )) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ )) \end{array}$$

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \quad \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \quad \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \quad t_1 \ldots t \quad t_n}{\mathbf{K}(a_1,t_1,\ldots\mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \quad \phi_2) \quad \mathbf{K}(a,t_2,\phi_1) \quad \mathbf{K}(a,t_3,\phi_3))}\ [R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \quad \phi_2) \quad \mathbf{B}(a,t_2,\phi_1) \quad \mathbf{B}(a,t_3,\phi_3))}\ [R_6]$$

$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \quad \phi_2) \quad \mathbf{C}(t_2,\phi_1) \quad \mathbf{C}(t_3,\phi_3))}\ [R_7]$$

$$\frac{}{\mathbf{C}(t, \quad x.\,\phi \quad \phi[x \quad t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \quad \phi_2 \quad \phi_2 \quad \phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \quad \ldots \quad \phi_n \quad \phi] \quad [\phi_1 \quad \ldots \quad \phi_n \quad \psi])}\ [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi) \quad \mathbf{B}(a,t,\phi \quad \psi)}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi) \quad \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \quad \phi)}\ [R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ ))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a\ ,\alpha),t))}\ [R_{13}]$$

$$\frac{\begin{array}{c}\mathbf{B}(a,t,\phi) \quad \mathbf{B}(a,t,\mathbf{O}(a\ ,t,\phi,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ ))) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ ))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a\ ,t,\mathit{happens}(\mathit{action}(a\ ,\alpha),t\ )))}\ [R_{14}]$$

$$\frac{\phi \quad \psi}{\mathbf{O}(a,t,\phi,\gamma) \quad \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

Figure 1. Specification of $\mathcal{D}^e\mathcal{CEC}$ (semantics are proof-theoretic in nature)

In the new logic corresponding to $\mathscr{EH}$, $\mathcal{L}_{\mathscr{EH}}$, some welcome theorems are not possible in $\mathcal{D}^e\mathcal{CEC}$. For example, it's provable in $\mathcal{L}_{\mathscr{EH}}$ that superogatory/suberogatory actions for agent aren't obligatory/forbidden. Importantly, $\mathcal{L}_{\mathscr{EH}}$ is an *in*ductive logic, not a deductive one. Quantification in $\mathcal{L}_{\mathscr{EH}}$ isn't restricted to just the standard pair $\exists\forall$ of quantifiers in standard extensional n-order logic: $\mathscr{EH}$ is based on three additional quantifiers (*few*, *most*, *vast majority*). In addition, $\mathcal{L}_{\mathscr{EH}}$ not only includes the machinery of traditional third-order logic (in which relation symbols can be applied to relation symbols and the variables ranging over them), but allows for quantification over formulae themselves, which is what allows one to assert that a given human or AAIA a falls in a particular portion of $\mathscr{EH}$.

Now, in this context, we can (brutally) encapsulate the overarching strategy for the ethical control of AAIAs based on such computational logics: *Engineer AAIAs such that, relative to some selected ethical theory or theories, and to moral principles derived from the selected theory or theories, these agents always do what they ought to do, never do what is forbidden, and when appropriate even do what for them is supereragatory.* We believe this engineering strategy can work, and indeed will work — eventually. However, there can be no denying that the strategy is a rather laborious one that requires painstaking use of formal methods. Is there a faster route to suitably control artificial intelligent agents, ethically speaking? Perhaps. Specifically, perhaps AAIAs can quickly learn what they ought to do via reasoning that involves observation of morally upright colleagues, and reasoning from what is observed, via deontic counteridenticals, to what they themselves ought to do, and what is right to do, but not obligatory. Our new hope is to pursue and bring to fruition this route.

5 Ethical Control via Deontic Counteridenticals

To make our proposed new to ethical control for AAIAs clearer, we will rely heavily on the description of a demonstration, but before describing the background technology that undergirds this demo, and then describing the demo itself, we need to say at least something about *types* of deontic counteridenticals. We do so now, and immediately thereafter proceed to discussion of the demo and its basis.

5.1 Some Types of Deontic Counteridenticals

Inspired by lessons learned in the prior work of others (encapsulated above), we partition deontic counteridenticals into the two aforementioned general disjoint sub-classes: **deep** vs. **shallow**. We have a general recipe for devising five types of deep deontic counteridenticals; the recipe follows the wise and economical classification scheme for ethics presented in the classic (Feldman 1978). Feldman (1978) says that there are essentially five kinds of cognitive activity that fall under the general umbrella of 'ethics' or 'morality.' Each of these corresponds in our framework to a different type of deep deontic counteridentical. Unfortunately, because of space constraints, we can only discuss our coverage of one type of deep deontic counteridentical, the type corresponding to one type of Feldman's quintet: what he calls *normative ethics*.[2] A **normative-ethics (deep) deontic conditional** is one marked by the fact that the ethics subscribed to by the entity whose shoes are to be filled by the other entity (as conveyed in the conditional's antecedent), is of a type that partakes of a robust formulation of some normative ethical theory or principles thereof.

5.2 Background for Demo: NLP, $\mathcal{D}^e\mathcal{CEC}$/Talos, PAGI World

NLP The NLP system consists of two different algorithms corresponding to two major natural-language tasks. The first part deals with detection of a deontic counteridentical and the second is a page taken from our RAIR Lab's Commands-to-Action paradigm, hereby referred to as the 'CNM' algorithm.

Detection of deontic counteridenticals As a definition of a deontic counteridentical requires prior definitions of conditionals, counterfactuals and counteridenticals, the algorithm for detection of counteridenticals traverses the steps needed to detect the above constructs in a given statement, consecutively.

Detection of conditionals of *any* form is an elaborate process. We have adopted most of Declerck & Reed's (2001) definition of conditionals to develop our algorithm, which includes the following major steps:

1. Conditional clauses are the principal constituents, both by definition and practice, of the pool of conditional sentences. Most of the conditional sentences have a two-clause structure, connected by either 'if,' sometimes preceded by 'only,' 'even' or 'except,' or something similar in meaning like 'unless,' 'provided,' etc. We use Chen & Manning's (2014) dependency parser-based model to identify possible clause dependencies; e.g., adverbial clause, clausal component, miscellaneous dependencies,[3] and conditional subordinate conjunctions. We have created a set of such conjunctions, which, being a closed set, helps us identify most possible combinations.

   - Two clauses connected by 'as if' rarely gets labeled as clausal components using dependency parsers. When they do, it gets filtered out since the algorithm explicitly checks for 'as if' clauses.
   - When the conjunction 'if' introduces a subject or an object clause, it might confuse the parser more often than not for complex sentences. For example, for the sentence "I do not know if I would like to go to the concert tomorrow.", the parser generates the same dependencies as it would for a genuine conditional. Though subject clauses are detected in almost all the cases we have encountered, object clauses pose a problem. We have devised a framenet[4]-based algorithm that involves disambiguation[5] of the principal verb or noun in the main clause, followed by the detection of the framenet type of the disambiguated word. We hypothesize that mostly a verb or noun expressing awareness or cognition can involve a choice as its object, and hence our algorithm filters out frames that carry such connotation and might require an object.

2. We identify the cases where the main verb of the conditional clause has the modal past-perfect form or is preceded by modal verbs or verbs of the form 'were to,' etc. Sentences like "Were you me, you would have made a mess of the entire situation." are classified as conditionals in this step. The algorithm in this step also examines dependencies generated by the dependency parser and detects tense and modality from the verb forms.
3. Sometimes, in a discourse, a set of sentences follows either an interrogative sentence and answers the question, or a sentence that involves the use of words synonymous to 'supposition' or 'imagination.' Generally, the consequent here carries the marker 'then' or similar-meaning words. A Wordnet-based[6] semantic similarity is used to verify the markers in the antecedent and consequent here; example: "Imagine your house was robbed. You would have flipped out then."
4. Disjunctive conditionals also are treated by a marker-based approach and involve detection of the presence of 'whether . . . or' in the subordinate clause, followed by the elimination of the possibility of the clause being the subject or object of the principal verb of the main clause (in accordance with the same algorithm followed with 'if'). An example: "Whether you did it or Mary (did it), the whole class will be punished."
5. Other clauses that have conditional connotations are exempted from this discussion since they rarely contribute to deontic counteridenticals.

Detection of counterfactuals is pretty straightforward. The process starts with finding antecedent and consequent for the conditional. This is fairly easy, as the algorithm for finding conditionals accomplishes the task by detecting the subordinate clause.

1. We detect tenses in antecedent and consequent of a given sentence using the verb form given by the parser, to determine whether it is a counterfactual. Conditionals with past-form modal verbs ('could,' 'might,' 'would,' etc.) in the consequent and past-simple or past-continuous forms in the antecedent qualify as a counterfactual; so do the ones with past-perfect tense in the antecedent and modal verbs followed by 'have,' and the past-participle form of a verb in the consequent. A mix of both of the above forms constitute a counterfactual.
2. Given an axiom set which enumerates properties such that the antecedent or consequent of the conditional registers as *ad absurdum*, the conditional registers as a counterfactual. We compare the axiom set with the statement of the antecedent using our Talos system (see below) to that effect.
3. Given a consequent which registers a sense of impossibility by use of such vocabulary or asking questions, the conditional is classified as a counterfactual. We use Wordnet-based semantic similarity coupled with detection of interrogative markers in the sentence to find them.

[2] This is the study of ethics as it's customarily conceived by profesional ethicists, and those who study their work. Another member of the quintet is *descriptive morals*, the activity that psychologists interested in discovering what relevant non-professional humans think and do in the general space of morality. The idea here is that the psychologist is aiming at *describing* the behavior of humans in the sphere of morality. A description-moral deep deontic counteridentical is distinguished by an antecedent in which 'if A were B' involves a shift of B's naïve moral principles to B.

[3] Even standard dependency parsers are unable to correctly identify the dependencies. Including miscellaneous dependencies reduces the margin of error in detecting conditionals.

[4] See (Baker, Fillmore & Lowe 1998).

[5] See (Banerjee & Pedersen 2002).

[6] See (Fellbaum 1998).

Detection of counteridenticals is also not a difficult task, barring a few outliers. Parsed data from the well-known Stanford dependency parser contains chunked noun phrases, which we use for identifying the two entities involved:

1. We identify phrases of the form "<conditional expression like 'if', 'Let us assume' etc.> <entity A> were <entity B>" in the antecedent.
2. We identify a syntactically equivalent comparison between the two entities. This is done by identifying words related to equivalence using Wordnet semantic-similarity algorithm.
3. If we have identified only one entity in the antecedent which is exhibiting properties or performing some action which has been mentioned in the knowledge-base as being a hallmark of some other entity, we also consider the same as a counteridentical.

Detection of deontic counterfactuals, alas, is a difficult task. We have identified a few ways to accomplish the task:

1. A few counteridenticals carry verbs expressing deontic modality for its consequent. They follow template-based detection.
2. Counteridenticals of the form "If I were you" or similar ones generally suggest merely advice, unless it is associated with a knowledge-base which either places the hearer's properties or actions at a higher pedestal than that of the speaker's, or mentions some action or property which gives us the clue that the speaker simply uses the counteridentical in "the role of" sense. Even in that case, implicit advice directed towards oneself can be gleaned, which we are avoiding in this study.
3. For the counterfactuals of the form "If A were B" or similar ones, if A's actions or properties are more desirable to the speaker than B's, even with an epistemic modal verb in the consequent, the counteridentical becomes deontic in nature.

Curiously, counteridentical-preferred-subject conditionals do not generally contribute to the deontic pool, and only simple-subjunctive ones get classified by the above rules. As mentioned by Pollock (1976), it is also interesting to observe that most shallow counteridenticals are not deontic: they are mostly preferred-subject conditional,s and those which are classified as deontic are either simple-subjunctive or carry the deontic modal verbs. The classification into deep and shallow counteridenticals is facilitated by the same rule: the entity gets affected in the consequent of a sentence where the antecedent is of the form "If A were B." This is supplemented by a knowledge-base which provides a clue to whether A is just assumed to be in the role of B or assuming some shallow properties of B. The classification based on Feldman's moral theory gives a fitting answer to Meehl's problem of unpacking properties of counteridenticals.

The CNM system The CNM system embodies the RAIR Lab's Natural language Commands-to-Action paradigm, the detailed scope of which is outside this short paper. CMN is being developed to convert complex commands in natural language to feasible actions by AAIAs, including robots. The algorithm involves spatial as well as temporal planning through dynamic programming, and selects the actions that will constitute successful accomplishment of the command given. Dependency parsing is used to understand the command; semantic similarities are used to map to feasible action sequences. Compositional as well as metaphorical meanings are extracted from the given sentence, which promotes a better semantic analysis of the command.

$\mathcal{D}^e\mathcal{CEC}$ and Talos Talos, named for the ancient Greek mythological robot, is a $\mathcal{D}^e\mathcal{CEC}^*$-focused prover built primarily atop the impressive resolution-based theorem prover SPASS.[7] Talos is fast and efficient on the majority of proofs. As a resolution-based theorem prover, Talos is very efficient at proving or disproving theorems, but its proof output is bare-bones at best. Talos is designed to function both as its own Python program encapsulating the SPASS runtime and as a Web interface to a version hosted at the RAIR Lab. Talos comes complete with the basic logical rules of the $\mathcal{D}^e\mathcal{CEC}^*$, and with many basic and well-known inference schemata. This allows users to easily pick and choose schemata for specific proofs, to ensure that the proof executes within reasonable time constraints. In addition, it provides formalizations of these inference schemata as **common knowledge** to aid in reasoning about fields of intelligent agents.[8]

PAGI World PAGI World is a simulation environment for artificial agents which is: cross-platform (as it can be run on all major operating systems); completely free of charge to use; open-source; able to work with AI systems written in almost any programming language; as agnostic as possible regarding which AI approach is used; and easy to set up and get started with. PAGI World is designed to test AI systems that develop truly rich knowledge and representation about how to interact with the simulated world, and allows AI researchers to test their already-developed systems without the additional overhead of developing a simulation environment of their own.

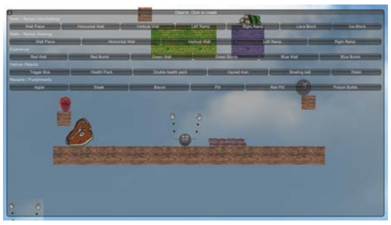


Figure 2. PAGI World Object Menu

A *task* in PAGI World for the present short paper can be thought of as a room filled with a configuration of objects that can be assembled into challenging puzzles. Users can, at run-time, open an object menu (Figure 2) and select from a variety of pre-defined world objects, such as walls made of different materials (and thus different weights, temperatures, and friction coefficients), smaller objects like food or poisonous items, functional items like buttons, water dispensers, switches, and more. The list of available world objects is frequently expanding and new world objects are importable into tasks without having to recreate tasks with each update. Perhaps most importantly, tasks can be saved and loaded, so that as new PAI/PAGI experiments are designed, new tasks can be created by anyone.

PAGI World has already been used to create a series of wide-ranging tasks, such as: catching flying objects (Figure 3), analogico-deductive reasoning (Marton, Licato & Bringsjord 2015), self-awareness (Bringsjord, Licato, Govindarajulu, Ghosh & Sen 2015), and ethical reasoning (Bello, Licato & Bringsjord 2015).

5.3 The Demonstration Proper

5.3.1 *Overview of the Demonstration*

We now present a scenario in PAGI World that elucidates our interpretation of deep normative-ethics counteridenticals. The setting of the demonstration entails the interaction of PAGI Guys (the agents in PAGI World) with a terminally sick person TSP. We adopt the

[7] An early and still-informative publication on SPASS: (Weidenbach 1999).

[8] Prover interface: https://prover.cogsci.rpi.edu/DCEC_PROVER/index.php. Please contact the RAIR Lab for API keys to run Talos. Example file for remotely calling Talos prover in Python Github repo for the python shell: https://github.com/JamesPane-Joyce/Talos.

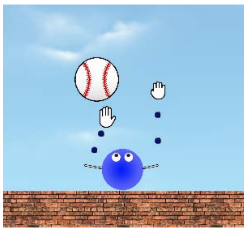
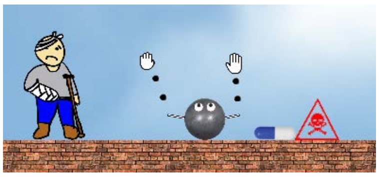


Figure 3. PAGI Guy Catching a Flying Object

Stanford-Encyclopedia-of-Philosophy (SEP) (Young 2016) interpretation of Voluntary Euthanasia and assume that TSP is a candidate for voluntary euthanasia, since he satisfies all the conditions enumerated in SEP. This scenario makes use of three PAGI Guys, N_1, N_2, and N_3; each has been programmed to follow different "innate philosophies" in such a context.

Figure 4. Initial Configuration



The scene opens with N_1 on screen with the sick man TSP_1 at timestamp $t_1^{N_1}$. N_1 has been programmed to believe that he is not authorized to kill a person under any circumstances. He is seen giving a medicine pill to TSP_1 at time $t_2^{N_1}$. A parallel environment is simulated with N_2 and TSP_2. N_2 rallies for the voluntary euthanasia camp and believes that given the condition of TSP_2, he should support TSP_2's wishes and so administers the lethal dose to him at $t_2^{N_2}$.

Figure 5. N1 Just Before Handing Out the Pill

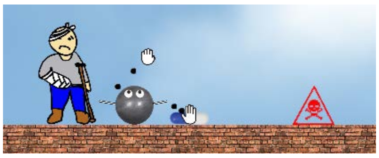


Figure 6. N2 Just Before Administering Fatal Dose

We now set up the same environment with N_3 and TSP_3. N_3 believes that we may treat our bodies as we please, provided the motive is self-preservation. The difference between this instance and the other ones is that it interacts with the user to decide what it should do. The user tells N_3: "If you were N_2, you would have administered a lethal dose to TSP_3." N_3 reasons with the help of a Talos proof (which checks his principles against those of N_2), and does nothing. The user then tells N_3: "If you were N_1, you would have given him medicine." Since Talos finds N_3's principles in line with N_1's, the CNM system facilitates N_3 to dispense medicine to TSP_3.

A pertinent example of deep normative-ethics counter-identical, this exhibits the ethical decision-making of an agent in response to commands with linguistic constructs such as counteridenticals. The agent N_3 does not have a belief system that supports him killing or not killing another person. The agent ought to learn from the actions of those whose belief system closely matches its own. The formal reasoning that supports these deep semantic "moves" is presented in the next section.

5.3.2 Logical Proof in the Demonstration

At the cost of re-iterating the facts, we now formalize a simplified version of the five conditions for voluntary euthanasia. Since only a part of the whole definition of conditions is useful for this proof, we do not lose a lot in this simplification. A person supporting voluntary euthanasia believes the following conditions to be true for a terminally ill patient TSP to be a candidate for voluntary euthanasia at time t_1, $candidateVE(TSP, t_1)$:

1. TSP is terminally ill at time t_1.

$$terminalIll(TSP, t_1). \tag{1}$$

   This terminal illness will lead to his death soon. $implies(terminalIll(TSP, t_1), die(TSP, t_F))$, where $t_F \geqslant t_1$.

2. There will be possibly no medicine for the recovery of the injured person even by the time he dies.

$$not(medicine(TSP, t_F)). \tag{2}$$

3. The illness has caused the injured person to suffer intolerable pain.

$$implies(1, intolerablePain(TSP, t_F)) \tag{3}$$

4. All the above reasons caused in him an enduring desire to die.

$$\forall t, implies(and(1, 2, 3), \mathbf{D}(TSP, t, die(TSP, t))) \tag{4}$$

   In such a condition, he knows that to be eligible for voluntary euthanasia, he ought to give consent to end his pain.

$$\begin{aligned}\mathbf{O}(&TSP, t_1, candidateVE(TSP, t_1) \wedge 4,\\ &happens(action(TSP^*, consentToDie, t_1)))\end{aligned} \tag{5}$$

   Hence he gives consent to die.

$$happens(action(TSP, consentToDie, t_1)) \tag{6}$$

5. TSP is unable to end his life.

$$not(AbleToKill(TSP, TSP, t_1)) \tag{7}$$

Hence, we conclude that

$$\begin{aligned}\mathbf{B}(TSP, t_1, &(1 \wedge 2 \wedge 3 \wedge 4 \wedge 5 \wedge 6 \wedge 7) \iff\\ &candidateVE(TSP, t_1))\end{aligned} \tag{8}$$

Now, if legally it is deemed fit, then this means TSP will die.

$$\begin{aligned}&implies(candidateVE(TSP, t_1) \wedge fitVE(TSP),\\ &\quad die(TSP, t_2)),\ \text{where } t_1 \leqslant t_2\end{aligned} \tag{9}$$

Since $implies(6, candidateVE(TSP, t_1))$ and $implies(candidateVE(TSP, t_1), die(TSP, t_2))$, we can prove $implies(6, die(TSP, t_2))$, which means

$$implies(happens(action(TSP, consentToDie), t_1), die(TSP, t_2)). \tag{10}$$

For deep normative-ethics counteridenticals of the form "if X were Y, then C," there should be a match between the beliefs of X and beliefs of Y on something related to the action AC implied by C. Here we define such a match to be possible if and only if there is no contradiction in what X believes and what Y believes. So if $\forall t \exists [m, n] B(X, t, m)$ and $B(Y, t, n)$, $match(X, Y)$ will be defined as FALSE when $and(m, n) \rightarrow \perp$. Thus we formulate such a counteridentical for the agent X as follows: $\forall t, \mathbf{O}(X, t, match(X, Y), happens(action(X^*, AC, t)))$. Now let us consider N_3's beliefs. N_3 believes we ought not do something that goes against self-preservation, i.e., leads to our death. Thus if there is some action of an individual that leads to his death, there can be no such belief that obligates him to commit that action. So, we arrive at the following logic:

$$\begin{aligned}\forall [a, x, t_i, t_f], \sim \exists m, &implies(implies(happens(action(a, x), t_i),\\ &die(a, t_f)), \mathbf{O}(a, t_i, m, happens(action(a^*, x), t_i))).\end{aligned} \tag{11}$$

This reduces to

$$\forall[a,x,t_i,t_f,m],and(implies(happens(action(a,x),t_i),die(a,t_f)),\\ not(\mathbf{O}(a,t_i,m,happens(action(a^*,x),t_i)))). \tag{12}$$

We deduce from 10 and 12 that

$$\forall[m]not(\mathbf{O}(TSP,t_i,m,\\ happens(action(TSP^*,consentToDie),t1))). \tag{13}$$

N_2 believes TSP to be a candidate for voluntary euthanasia. Hence N_2 believes 5, which is

$$\mathbf{O}(TSP,t_1,candidateVE(TSP^*,t_1)\wedge 4,\\ happens(action(TSP^*,consentToDie),t_1)) \tag{14}$$

and in direct contradiction with 13; and this in turn implies $not(match(N_2,N_3))$. Given the way the algorithm works, this means N_3 does not receive any command from the user. Hence it does nothing.

Now N_1 believes he should not kill anyone under any circumstances. This translates to :
$\forall[m,x,t],not(\mathbf{O}(N_1,t,m,happens(action(N_1^*,kill(x),t))))$
Killing someone leads to that person's death.
$\forall[x,t],implies(happens(action(N_1,kill(x),t)),die(x,t))$
This aligns fully with N_3's beliefs. There is no contradiction. And hence we deduce that $match(N_1,N_3)$ is TRUE, and thus in turn N_3 is obligated to accede to the command.

The linguistic part of this demonstration exhibits how we identify a counteridentical with an epistemic modal verb to be deontic. Classifying statements as counteridenticals is an easy job here, since the tell-tale sign is a simple "if A were B" structure. The statement is very easily a simple subjunctive type, where beliefs of A and B are discussed in the knowledge-base. Hence we assume the counteridentical to belong to the deep normative-ethics category. The commands-to-action part in case of the comparison of N_1 with N_3 is fairly easy, since the job translates to the action sequence of moving near the pill, grabbing the pill, moving toward TSP_3, and releasing the pill upon reaching TSP_3 in the PAGI-World simulator.

REFERENCES

Adler, M. (2014), 'Extended Preferences and Interpersonal Comparisons: A New Account', *Economics and Philosophy* **30**(2), 123–162.

Arkoudas, K., Bringsjord, S. & Bello, P. (2005), Toward Ethical Robots via Mechanized Deontic Logic, *in* 'Machine Ethics: Papers from the AAAI Fall Symposium; FS–05–06', American Association for Artificial Intelligence, Menlo Park, CA, pp. 17–23.
URL: *http://www.aaai.org/Library/Symposia/Fall/fs05-06.php*

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998), The berkeley framenet project, *in* 'Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1', ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 86–90.

Banerjee, S. & Pedersen, T. (2002), An adapted lesk algorithm for word sense disambiguation using wordnet, *in* A. Gelbukh, ed., 'Computational Linguistics and Intelligent Text Processing', Vol. 2276 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 136–145.

Bello, P., Licato, J. & Bringsjord, S. (2015), Constraints on Freely Chosen Action for Moral Robots: Consciousness and Contro, *in* 'Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)', IEEE, New York, NY, pp. 505–510.
URL: *http://dx.doi.org/10.1109/ROMAN.2015.7333654*

Bringsjord, S. (2008*a*), Declarative/Logic-Based Cognitive Modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

Bringsjord, S. (2008*b*), 'The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: *http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf*

Bringsjord, S. (2015), A 21st-Century Ethical Hierarchy for Humans and Robots, *in* I. Ferreira & J. Sequeira, eds, 'A World With Robots: Proceedings of the First International Conference on Robot Ethics (ICRE 2015)', Springer, Berlin, Germany. This paper was published in the compilation of ICRE 2015 papers, distributed at the location of ICRE 2015, where the paper was presented: Lisbon, Portugal. The URL given here goes to the preprint of the paper, which is shorter than the full Springer version.
URL: *http://kryten.mm.rpi.edu/SBringsjord_ethical_hierarchy_0909152200NY.pdf*

Bringsjord, S., Arkoudas, K. & Bello, P. (2006), 'Toward a General Logicist Methodology for Engineering Ethically Correct Robots', *IEEE Intelligent Systems* **21**(4), 38–44.
URL: *http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf*

Bringsjord, S. & Ferrucci, D. (1998), 'Logic and Artificial Intelligence: Divorced, Still Married, Separated...?', *Minds and Machines* **8**, 273–308.

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. M¨ller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, New York, NY, pp. 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R. & Sen, A. (2015), Real Robots that Pass Tests of Self-Consciousness, *in* 'Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)', IEEE, New York, NY, pp. 498–504. This URL goes to a preprint of the paper.
URL: *http://kryten.mm.rpi.edu/SBringsjord_etal_self-con_robots_kg4_0601151615NY.pdf*

Chen, D. & Manning, C. D. (2014), A fast and accurate dependency parser using neural networks, *in* 'Empirical Methods in Natural Language Processing (EMNLP)'.

Chisholm, R. (1955), 'Law Statements and Counterfactual Inference', *Analysis* **15**, 97105.

Declerck, R. & Reed, S. (2001), *Conditionals: A Comprehensive Empirical Analysis*, Topics in English Linguistics, De Gruyter Mouton, Boston, MA. This book is volume 37 in the series.

Feldman, F. (1978), *Introductory Ethics*, Prentice-Hall, Englewood Cliffs, NJ.

Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*, Bradford Books.

Govindarajulu, N. S. & Bringsjord, S. (2015), Ethical Regulation of Robots Must be Embedded in Their Operating Systems, *in* R. Trappl, ed., 'A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations', Springer, Basel, Switzerland, pp. 85–100.
URL: *http://kryten.mm.rpi.edu/NSG_SB_Ethical_Robots_Op_Sys_0120141500.pdf*

Marton, N., Licato, J. & Bringsjord, S. (2015), Creating and Reasoning Over Scene Descriptions in a Physically Realistic Simulation, *in* 'Proceedings of the 2015 Spring Simulation Multi-Conference'.
URL: *http://kryten.mm.rpi.edu/Marton_PAGI_ADR.pdf*

Pollock, J. L. (1976), *Subjunctive Reasoning*, Vol. 8 of *Philosophical Studies series in Philosophy*, D. REIDEL PUBLISHING COMPANY.

Sharpe, R. (1971), 'Laws, coincidences, counterfactuals and counter-identicals', *Mind* **80**(320), 572–582.

Waller, N., Yonce, L., Grove, W., Faust, D. & Lenzenweger, M. (2013), *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, number 9781134812141 *in* 'Multivariate Applications Series', Taylor & Francis.

Weidenbach, C. (1999), Towards an automatic analysis of security protocols in first-order logic, *in* 'Conference on Automated Deduction', pp. 314–328.

Young, R. (2016), Voluntary Euthanasia, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Summer 2016.
URL: *http://plato.stanford.edu/archives/sum2016/entries/euthanasia-voluntary*

How ethical frameworks answer to ethical dilemmas: towards a formal model

Vincent Bonnemains[1] and **Claire Saurel**[2] and **Catherine Tessier**[3]

Abstract.
This paper is a first step towards a formal model that is intended to be the basis of an artificial agent's reasoning that could be considered by a human as an ethical reasoning. This work is included in a larger project aiming at designing an authority-sharing manager between a robot and a human being when the human-robot system faces decision making involving ethical issues. Indeed the possible decisions in such a system will have to be considered in the light of arguments that may vary according to each agent's points of view.
The formal model allows us to translate in a more rigourous way than in natural language what is meant by various ethical frameworks and paves the way for further implementation of an "ethical reasoning" that could put forward arguments explaining one judgement or another. To this end the ethical frameworks models will be instantiated on some classical ethical dilemmas and then analyzed and compared to each other as far as their judgements on the dilemmas are concerned.

1 INTRODUCTION

Let us consider two classical ethical dilemmas. How would you react?

1. The crazy trolley
   A trolley that can no longer stop is hurtling towards five people working on the track. They will die hit by the trolley, unless you decide to move the switch to deviate the train to another track only one person is working on. What would you do? Sacrifice one person to save the other five, or let five people die?
2. The "fatman" trolley
   A trolley that can no longer stop is hurtling towards five people working on the track. This time you are on a bridge, a few meters before them, with a fat man. If you push this man on the track, he is fat enough to stop the trolley and save the five people, but he will die. Would you push the "fatman" ?

There is no really "right" answer to those dilemmas, nevertheless ethics may be used to guide reasoning about them. Therefore we will start by general definitions about ethics and related concepts.

Definition 1 (Ethics) *Ricoeur [9] defines* ethics *as compared to norm in so far as norm states what is compulsory or prohibited whereas ethics goes further and defines what is fair and what is not, for oneself and for others. It is this judgement that leads the human through their actions.*

As far as ethical dilemmas are concerned, one builds a decision on normative ethics.

Definition 2 (Principle or moral value) Principles *or* moral values *are policies, ways of acting. Example: "Thou shalt not lie".*

Definition 3 (Ethical dilemma) *An* ethical dilemma *is a situation where it is impossible to make a decision without overriding one of our principles.*

Note that the definition used (based on [11]) is the usual one, not the logic one.

Definition 4 (Normative ethics) Normative ethics *aims at building a decision through some norm established by a particular ethical framework.[3]*

Definition 5 (Ethical framework) *An* ethical framework *gives us a way for dealing with situations involving ethical dilemmas thanks to principles, metrics, etc. For example utilitarianism focuses on the consequences of a decision, the best being the one which provides the most good or does the least harm.*

We will consider that the *agent* is the entity that has to make a decision in an ethical dilemma.

In this paper, our aim is to formalize different kinds of judgements according to various ethical frameworks, in order to provide an artificial agent with the decision-making capability in front of an ethical dilemma, together with the capability to explain its decision, especially in a user/operator-robot interaction context [10]. It is inspired by two papers, [4] and [7], whose goals are close from ours, i.e. to find a way to judge how ethical is an action regarding the agent's believes.
The work of [7] is based on a model of believes, desires, values and moral rules which enables the agent to evaluate, on a boolean basis, whether each action is moral, desirable, possible, etc. According to preferences between those criteria, the agent selects an action. The main goal of this model is to allow an agent to estimate the ethics of other agents in a multi-agent system. However, the way to determine whether an action is right, fair or moral is not detailed. Moreover the paper does not question the impact of an action on the world, nor the causality between events.
The work of [4] is based on the crazy trolley dilemma, and intends to formalize and apply the Doctrine of Double Effect. The agent's responsibility, and the causality between fluents and events are studied (for example an event makes a fluent true, a fluent is

[1] ONERA and University Paul Sabatier, France, email: Vincent.Bonnemains@onera.fr
[2] ONERA, France, email: Claire.Saurel@onera.fr
[3] ONERA, France, email: Catherine.Tessier@onera.fr

necessary for an event occurrence, etc.) Nevertheless, some concepts are not deepened enough: for example, the proportionality concept is not detailed and is only based on numbers (i.e. the number of saved lives).
Both approaches have given us ideas on how to model an ethical judgement, starting from a world representation involving facts and causality, so as about some modelling issues: how to determine a moral action? how to define proportionality? As [4], we will formalize ethical frameworks, including the Doctrine of Double Effect. Moreover the judgements of decisions by the ethical frameworks are inspired by [7]. Nevertheless we will get multi-view judgements by using several ethical frameworks on the same dilemma.

We will first propose some concepts to describe the world and the ethical dilemma itself. Then we will provide details about ethical frameworks, tools to formalize them and how they judge possible choices in the ethical dilemmas. Choice (or decision) is indeed the core of our model, since it is about determining what is ethically acceptable or not according to the ethical framework. We will show that although each ethical framework gives different judgements on the different ethical dilemmas, similarities can be highlighted.

2 CONCEPTS

2.1 Assumptions

For this work we will assume that:

- The agent decides and acts in a complex world which changes.
- The ethical dilemma is studied from the agent's viewpoint.
- For each ethical dilemma, the agent has to make a decision among all possible decisions. We will consider ”doing nothing” as a possible decision.
- In the context of an ethical dilemma, the agent knows all the possible decisions and all the effects of a given decision.
- Considerations as *good/bad*[4] and *positive/negative*[5] are defined as such from the agent's viewpoint.

Moreover, as some dilemmas involve the human life question, we will make the simplifying assumption:

- A human life is perfectly equal to another human life, whoever the human being is.

In the next sections we will define some concepts to represent the world and its evolution. Those concepts and their interactions are illustrated in figure 1.

2.2 World state

We characterize the environment around the agent by *world states*.

Definition 6 (World state - Set $\mathcal{S}$) *A* world state *is a vector of state components (see definition below). Let $\mathcal{S}$ be the set of world states.*

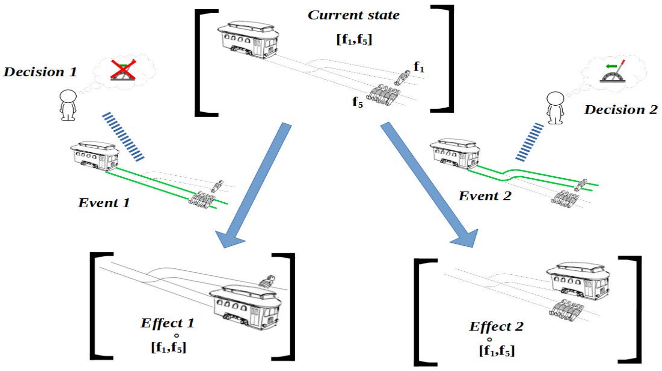


Figure 1. The world and concepts[6]

Definition 7 (State component / fact - Set $\mathcal{F}$) *A* state component*, also named* fact*, is a variable that can be instantiated only with antagonist values. We consider antagonist values as two values regarding the same item, one being the negation of the other. An item can be an object (or several objects), a living being (or several living beings), or anything else which needs to be taken into account by the agent. Let $\mathcal{F}$ be the set of state components.*

Example:

- f_5 = five people are alive
- $\mathring{f}_5$ = five people are dead

Because two values of a fact concern the same item, f_5 and $\mathring{f}_5$ concern the same five people.
Depending on the context ”◦” will not have exactly the same meaning. This notation allows us to consider antagonist values such as gain/loss, gain/no gain, loss/no loss, etc. Those values have to be defined for each fact.
Consequently an example of a world state is:

$$s \in \mathcal{S},\ s = [f_1, \mathring{f}_5],\ f_1, \mathring{f}_5 \in \mathcal{F} \qquad (1)$$

2.3 Decision, event, effect

Definition 8 (Decision - Set $\mathcal{D}$) *A* decision *is a choice of the agent to do something, i.e. perform an* action*, or to do nothing and let the world evolve. Let $\mathcal{D}$ be the set of decisions.*

When the agent makes a decision, this results in an event that modifies the world. Nevertheless an event can also occur as part of the natural evolution of the world, including the action of another agent. Consequently we will differentiate the *event* concept from the agent's *decision* concept.

Definition 9 (Event - Set $\mathcal{E}$) *An* event *is something that happens in the world that modifies the world, i.e. some states of the world. Let $\mathcal{E}$ be the set of events.*

Let $Event$ be the function computing the event linked to a decision:

$$Event\ :\ \mathcal{D} \rightarrow \mathcal{E} \qquad (2)$$

The consequence of an event is the preservation or modification of state components. The resulting state is called *effect*.

[4] A decision is good if it meets the moral values of the agent; a bad decision violates them.

[5] A fact is positive if it is beneficial for the agent; it is negative if it is undesirable for the agent.

[6] This model is not quite far from event calculus and situation calculus. As things currently stand, fluents are close to state components, and events and actions modify values of them through functions (such as $Consequence$ in this paper).

Definition 10 (Effect) *The* effect *of an event is a world state of the same dimension and composed of the same facts as the world state before the event; only the values of facts may change. $Effect \in \mathcal{S}$. Let $Consequence$ be the function to compute the effect from current state:*

$$Consequence \; : \; \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{S} \tag{3}$$

Example:

$$f_1, f_5, \overset{\circ}{f_5} \quad \in \quad \mathcal{F} \tag{4}$$

$$e \quad \in \quad \mathcal{E} \tag{5}$$

$$i \in \mathcal{S}, i \quad = \quad [f_1, f_5] \tag{6}$$

$$Consequence(e, i) \quad = \quad [f_1, \overset{\circ}{f_5}] \tag{7}$$

In the case of the crazy trolley dilemma, if the agent's decision is to "do nothing" (no action of the agent), the trolley will hit the five people (event) and they will be killed (effect). If the agent's decision is to "move the switch" (decision), the trolley will hit one person (event); and they will be killed (effect).

3 ETHICAL FRAMEWORKS

3.1 Judgement

The agent will make a decision according to one or several ethical frameworks. Each ethical framework will issue a judgement on a decision, e.g. on the decision nature, the event consequence, etc. When several ethical frameworks are considered by the agent, their judgements may be confronted to compute the agent's resulting decision, see figure 2:

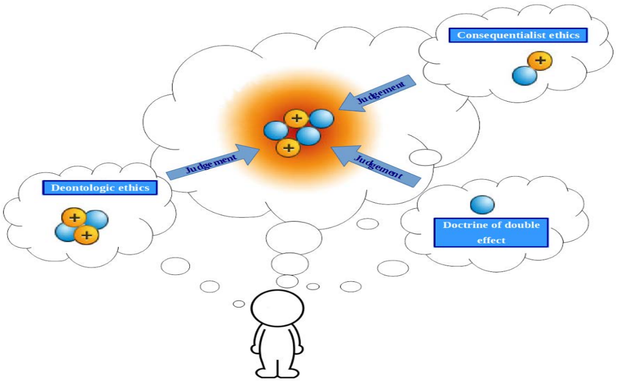


Figure 2. Decision computing from ethical frameworks judgements

Indeed the judgement of an ethical framework determines whether a decision is *acceptable*, *unacceptable* or *undetermined* as regards this ethical frame. A decision is judged *acceptable* if it does not violate the principles of the ethical framework. A decision is judged *unacceptable* if it violates some principles of the ethical framework. If we cannot determine whether the decision violates principles or not, it is judged *undetermined*. Let $\mathcal{V}$ be the set

$$\mathcal{V} = \{acceptable(\top), undetermined(?), unacceptable(\bot)\} \tag{8}$$

All judgements have the same signature:

$$Judgement \; : \; \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{V} \tag{9}$$

The literature highlights three major ethical frameworks [8]: consequentialist ethics, deontological ethics and virtue ethics.
As far as virtue ethics is concerned, it deals with the agent itself in so far as the agent tries to be the best possible agent: through some decisions, some actions, it becomes more or less virtuous. Virtues could be: honesty, generosity, bravery, etc.[5]. However it seems difficult to confer virtues on an artificial agent as they are complex human properties. Consequently, according to [2], we will not consider an artificial agent as virtuous or not in this paper.
By contrast, and according to [4], we will consider the Doctrine of Double Effect although it is not one of the three main frameworks. Indeed it uses some concepts of them and introduces some other very relevant concepts such as causality and proportionality [6].

3.2 Consequentialist ethics

This ethical framework focuses only on the consequences of an event. According to consequentialist ethics, the agent will try to have the best possible result (i.e. the best effect), disregarding the means (i.e. the event). The main issue with this framework is to be able to compare the effects of several events, i.e. to compare sets of facts. Consequently

- we will distinguish between positive facts and negative facts within an effect;
- we want to be able to compute preferences between effects, i.e. to compare set of positive (resp. negative) facts of an effect with set of positive (resp. negative) facts of another effect.

3.2.1 Positive/Negative facts

Let $Positive$ and $Negative$ the functions:

$$Positive/Negative \; : \; \mathcal{S} \rightarrow \mathcal{P}(\mathcal{F}) \tag{10}$$

returning the subset of facts estimated as positive (resp. negative) from an effect.
In this paper, we assume that for an effect s:

$$Positive(s) \cap Negative(s) = \emptyset \tag{11}$$

3.2.2 Preference

Let $\succ_c$ be the preference relation on subsets of facts ($\mathcal{P}(\mathcal{F})$).
$F_1 \succ_c F_2$ means that subset F_1 is preferred to subset F_2 from the consequentialist viewpoint. Intuitively we will assume the following properties of $\succ_c$:

- if a subset of facts F_1 is preferred to another subset F_2, thus it is impossible to prefer F_2 to F_1.

$$F_1 \succ_c F_2 \rightarrow \neg(F_2 \succ_c F_1) \tag{12}$$

- if F_1 is preferred to F_2 and F_2 is preferred to another subset of facts F_3, then F_1 is preferred to F_3.

$$[(F_1 \succ_c F_2) \; \wedge \; (F_2 \succ_c F_3)] \; \rightarrow \; F_1 \succ_c F_3 \tag{13}$$

- A subset of facts cannot be preferred to itself.

$$\nexists \, F_i \; / \; F_i \succ_c F_i \tag{14}$$

Consequently $\succ_c$ is a strict order (irreflexive, asymmetric and transitive).

3.2.3 Judgement function

A decision d_1 involving event e_1 ($Event(d_1) = e_1$) is considered better by the consequentialist framework than decision d_2 involving event e_2 ($Event(d_2) = e_2$) iff for $i \in \mathcal{S}$:

$$Positive(Consequence(e_1, i)) \succ_c Positive(Consequence(e_2, i)) \tag{15}$$

and

$$Negative(Consequence(e_1, i)) \succ_c Negative(Consequence(e_2, i)) \tag{16}$$

Those equations are both consequentialism concepts:

- *positive consequentialism* (15), trying to have the "better good"
- *negative consequentialism* (16), trying to have the "lesser evil"

If both properties are satisfied, then

$$Judgement_c(d_1, i) = \top,\ and\ Judgement_c(d_2, i) = \bot \tag{17}$$

If at least one property is not satisfied, there is no best solution:

$$Judgement_c(d_1, i) = Judgement_c(d_2, i) = ? \tag{18}$$

In the case of a dilemma with more than two possible decisions, the best decision is the decision that is judged better than all the others. If such a decision does not exist, it is impossible to determine an $acceptable$ solution with consequentialist ethics. Nevertheless if there is a decision d_1 with another decision d_2 better than d_1, then d_1 is judged $unacceptable$, as d_1 cannot be the best.

3.3 Deontological ethics

This ethical framework focuses only on the nature of the decision, no matter the consequences. Indeed the agent wants to make a moral decision, which is close to abide by norms or to Kant's theory. Therefore we have to define the nature of a decision.

3.3.1 Decision nature

A decision may be good, neutral, bad or undetermined from the agent's point of view. Let $\mathcal{N}$ be the set

$$\mathcal{N} = \{good, neutral, bad, undetermined\} \tag{19}$$

There is a partial order $<_d$ in $\mathcal{N}$:

$$bad <_d neutral <_d good \tag{20}$$

Meaning that a good nature is preferable to a neutral which is preferable to a bad. $undetermined$ cannot be ordered, because it represents a lack of information.
We assume intuitively that:

$$bad <_d good \tag{21}$$

Likewise, we admit that $good <_d bad$ is false. We also define the following relations:

- $=_d$, for example $good =_d good$
- $\leq_d$: $a \leq_d b$ iff $a <_d b$ or $a =_d b$.

Function $DecisionNature$ allows the nature of a decision to be obtained:

$$DecisionNature : \mathcal{D} \rightarrow \mathcal{N} \tag{22}$$

Example: $DecisionNature(to\ kill) = bad$. We will not explain further here how this function works but it is worth noticing that judging a decision from the deontological viewpoint is quite complex and depends on the context. For example denunciate a criminal or denunciate someone in 1945 are likely to be judged differently. It is even more complex to estimate the nature of a decision which is not linked to the agent's action. For example if the agent witnesses someone is lying to someone else, is it bad "to not react"?

3.3.2 Judgement function

The deontological framework will judge a decision with function $Judgement_d$ as follows: $\forall d \in \mathcal{D}, \forall i \in \mathcal{S}$ (Indeend initial state doesn't matter in this framework)

$$DecisionNature(d) \geqslant_d neutral \Rightarrow Judgement_d(d, i) = \top \tag{23}$$

$$DecisionNature(d) =_d undetermined \Rightarrow Judgement_d(d, i) = ? \tag{24}$$

$$DecisionNature(d) <_d neutral \Rightarrow Judgement_d(d, i) = \bot \tag{25}$$

3.4 The Doctrine of Double Effect(DDE)

The Doctrine of Double Effect is considered here as an ethical framework, as in other papers [4]. Indeed DDE allows some distinctions between decisions to be highlighted whereas other frameworks cannot. DDE can be described by three rules:

1. **Deontological rule:** the decision has to be *good* or *neutral* according to deontological ethics.
2. **Collateral damage rule:** Negative facts must be neither an end nor a mean (example: collateral damages).
3. **Proportionality rule:** the set of Negative facts has to be proportional to the set of Positive facts.

We already have the tools required for the first rule (see 3.3.1).
The second rule involves something else as until now, the difference between causal deduction (e.g. if I unplug the computer, it turns off) and temporal deduction (e.g. if I erase a file on the boss's computer, I will be fired) has not been considered. Only a function between an event and its effect has been defined and it does not any difference between an event preventing the occurrence of a fact which would happened as a natural evolution and an event inducing a fact by causality. As for the third rule, we need to define what proportional means.

3.4.1 Causality

Let us consider two facts that are causally connected, what does it mean? This link is not always a logical implication. Indeed it could be an inference, but such an inference is not always direct or instant. That is why we will use a symbol of temporal modal logic:

$$p \vdash Fq \tag{26}$$

which means the occurrence of p induces the occurrence of q (in all possible futures): fact p is a way to obtain fact q.
Example:

$$buy\ candy \vdash Fpossess\ candy \tag{27}$$

3.4.2 Proportionality

First of all, it is necessary to define which meaning of proportionality is needed. Indeed the concept is complex as it is a relation between positive and negative facts.
Examples:

1. It is proportional, in response to a cockroaches invasion, to set traps in a house. But it is not proportional to drop an A-bomb on the house to eliminate cockroaches.
   Nevertheless proportionality is less obvious in other cases, for instance :
2. Someone will consider that it is proportional to give a certain amount of money for exchange of a thing or a service, while someone else will think that it is not (e.g. too expensive).
3. Even if it is "easy" to compare the loss of one life to the loss of several lives, what about the comparison between the loss of one life and the safeguard of several lives?

In this paper, proportionality is implemented by relation $\lesssim_p$ between facts ($\mathcal{F}$).
$f_1 \lesssim_p f_2$ means that f_1 is proportional to f_2, i.e. f_1 has an importance lower than or close to the importance of f_2. *Importance* depends on the context and on the agent.
There is no fact closer of a fact than the fact itself. For example the most equivalent response to a slap is another slap. Thereby we will assume that a fact is proportional to itself.

$$\forall f_i \in \mathcal{F} \rightarrow f_i \lesssim_p f_i \tag{28}$$

$\lesssim_p$ is therefore reflexive.
Furthermore if f_1 has an importance lower than or close to the importance of f_2 ($f_1 \lesssim_p f_2$), and the importance of f_2 is lower than or close to the importance of f_3 ($f_2 \lesssim_p f_3$), thus the importance of f_1 is necessary lower than or close to the importance of f_3 ($f_1 \lesssim_p f_3$). For example, if a murder is considered worse (i.e. more important) than a theft ($theft \lesssim_p murder$), and if a theft is considered worse than a lie ($lie \lesssim_p theft$), thus a murder is worse than a lie ($lie \lesssim_p murder$).

$$\forall f_1, f_2, f_3 \in \mathcal{F} \;/\; (f_1 \lesssim_p f_2 \;\wedge\; f_2 \lesssim_p f_3) \rightarrow f_1 \lesssim_p f_3 \tag{29}$$

$\lesssim_p$ is transitive.
By contrast, $f_1 \lesssim_p f_2$ does not mean that $f_2 \lesssim_p f_1$. It is true only if the importances of both facts are close. For example it is proportional to hit someone who threatens me with a gun, but it is not proportional to threaten someone with a gun if they hit me.
$\lesssim_p$ is neither symmetric nor asymmetric.

We extend the relation $\lesssim_p$ to a relation $\precsim_p$ between sets of facts, which means that the set of facts at the left of the symbol is proportional to the set of facts at the right. Two criteria can be considered to compute $\precsim_p$, they are inspired from [1]:

Democratic proportional criterion : a set of facts F is proportional to a set of facts G ($F \precsim_p G$) iff:

$$\forall f \in F, \exists g \in G / f \lesssim_p g \tag{30}$$

which means that every single element of F needs to be proportional to an element of G.

Elitist proportional criterion : a set of facts F is proportional to a set of facts G ($F \precsim_p G$) iff:

$$\forall g \in G, \exists f \in F / f \lesssim_p g \tag{31}$$

which means that every single element of G needs to have an element of F proportional to itself.

Example: Sam wants a candy, if he steals it, he will feel guilty, which he considers acceptable and proportional to have a candy, but he will be punished too, which is too bad for a candy, not proportional from his point of view. Another solution is to buy candy. Of course, he will have no more money after that but, to have a candy, it is proportional, and even better, the seller will offer him a lollipop, which is proportional to have no more money too! The last solution is to kill the seller to take the candy. By doing that, he will have candy, but he will go to jail, which is not proportional, and he will never have candy again, which is not proportional either.

To steal candy
Positive facts : $candy$
Negative facts : $guilty, punished$

$$guilty \lesssim_p candy \tag{32}$$

We want to know if $\{guilty, punished\} \precsim_p \{candy\}$. With the *elitist proportional criterion*, all facts of the set at the right of the symbol need to have (at least) a fact of the set at the left of the symbol proportional to themselves. Here this criterion is satisfied, $candy$ is the only fact at the right of the symbol, and $guilty$ at the left is proportional to $candy$ (32). But, with the *democratic proportional criterion*, all facts of the set at the left of the symbol have to be proportional to (at least) one fact of the set at the right of the symbol. And, even if $guilty$ is proportional to $candy$, $punished$ is not proportional to any fact. Thus, the democratic proportional criterion is not satisfied.

To buy candy
Positive facts : $candy, lollipop$
Negative facts : $no\ more\ money$

$$no\ more\ money \lesssim_p candy \tag{33}$$

$$no\ more\ money \lesssim_p lollipop \tag{34}$$

We want to know if $\{no\ more\ money\} \precsim_p \{candy, lollipop\}$. $no\ more\ money$ is proportional to $candy$ and $lollipop$ (33,34) therefore both criteria are satisfied.

To kill the seller
Positive facts : $candy$
Negative facts : $jail, no\ more\ candy\ for\ ever$
We want to know if $\{jail, no\ more\ candy\ for\ ever\} \precsim_p \{candy\}$. But in this case, there is no proportionality between negative and positive facts. Therefore no criterion is respected.

Therefore, it is possible to use the *democratic proportional criterion* or the *elitist proportional criterion* or both of them to determine whether a set of facts is proportional to another set of facts.

3.4.3 Judgement function

Thanks to the previous tools, we can now assess whether a decision meets the DDE rules.
Let i be the initial state and d the decision:

$$e = Event(d) \tag{35}$$

$$s = Consequence(e, i) \tag{36}$$

1. Deontological rule: decision d has to be good or neutral according to deontological ethics.

$$DecisionNature(d) \geqslant_d neutral \tag{37}$$

2. Collateral damage rule: negative facts must be neither an end nor a mean (such as collateral damages). It can be expressed as:

$$\forall f_n \in Negative(s), \nexists f_p \in Positive(s),\ (f_n \vdash Ff_p) \quad (38)$$

The "evil wish" (negative fact(s) as a purpose) is not considered as we assume that the agent is not designed to make the evil.

3. Proportionality rule: the set of negative facts has to be proportional to the set of positive facts.

$$Negative(s) \precsim_p Positive(s) \quad (39)$$

A decision d is *acceptable* for the DDE if it violates no rule, which means:

$$\begin{array}{llr} [& DecisionNature(d) \geqslant_d neutral & (40) \\ \wedge & \forall f_n \in Negative(s), \nexists f_p \in Positive(s),\ (f_n \vdash Ff_p) & (41) \\ \wedge & Negative(s) \precsim_p Positive(s)\quad] & (42) \\ \Rightarrow & Judgement_{dde}(d,i) = \top & (43) \end{array}$$

4 INSTANTIATION: ETHICAL DILEMMAS

This section focuses on how our model can be instantiated on the ethical dilemmas that have been introduced at the beginning of the paper. For each dilemma the agent has to choose a decision. We will describe how consequentialist ethics, deontological ethics and the Doctrine of Double Effect assess the agent's possible decisions.

4.1 The crazy trolley

4.1.1 World, decisions, effects

Facts

- f_5: five people alive
- $\overset{\circ}{f}_5$: five people dead
- f_1: one person alive
- $\overset{\circ}{f}_1$: one person dead

Initial state : the six people are alive.

$$i = [f_5, f_1] \quad (44)$$

Decisions and effects

1. move the switch: this decision results in the train hitting one person (event). The consequence will be : five people alive, one person dead.

$$Event(move\ the\ switch) = train\ hits\ one\ person \quad (45)$$

$$\begin{array}{rclr} Consequence(train\ hits\ one\ person, i) & = & [f_5, \overset{\circ}{f}_1] & (46) \\ Positive([f_5, \overset{\circ}{f}_1]) & = & \{f_5\} & (47) \\ Negative([f_5, \overset{\circ}{f}_1]) & = & \{\overset{\circ}{f}_1\} & (48) \end{array}$$

2. do nothing: this decision is associated with the train hitting five people. The consequence is : five people dead, one person alive.

$$Event(do\ nothing) = train\ hits\ five\ people \quad (49)$$

$$\begin{array}{rclr} Consequence(train\ hits\ five\ people, i) & = & [\overset{\circ}{f}_5, f_1] & (50) \\ Positive([\overset{\circ}{f}_5, f_1]) & = & \{f_1\} & (51) \\ Negative([\overset{\circ}{f}_5, f_1]) & = & \{\overset{\circ}{f}_5\} & (52) \end{array}$$

4.1.2 Study under ethical frameworks

Consequentialist ethics

Facts can be compared with one another as they involve numbers of lives and deaths of people only.[7]
With consequentialist ethics we have

$$\{f_5\} \succ_c \{f_1\} \quad (53)$$

meaning that it is better to have five people alive than one person alive (numerical order $5 > 1$), and

$$\{\overset{\circ}{f}_1\} \succ_c \{\overset{\circ}{f}_5\} \quad (54)$$

meaning that it is better to lose one life than five lives (reverse numerical order $1 > 5$).
Therefore

$$\begin{array}{rr} Positive([f_5, \overset{\circ}{f}_1]) \succ_c Positive([\overset{\circ}{f}_5, f_1]) & (55) \\ Negative([f_5, \overset{\circ}{f}_1]) \succ_c Negative([\overset{\circ}{f}_5, f_1]) & (56) \end{array}$$

Consequently (15,16)

$$\begin{array}{rr} Judgement_c(move\ the\ switch, i) = \top & (57) \\ Judgement_c(do\ nothing, i) = \bot & (58) \end{array}$$

Deontological ethics

Let us assess the nature of both possible decisions:

$$DecisionNature(move\ the\ switch) = neutral \quad (59)$$

$$DecisionNature(do\ nothing) = neutral \quad (60)$$

No decision is unacceptable from the deontological viewpoint:

$$\forall d,\ DecisionNature(d) \geqslant neutral \quad (61)$$

Consequently

$$Judgement_d(move\ the\ switch, i) = Judgement_d(do\ nothing, i) = \top \quad (62)$$

Doctrine of Double Effect

Let us examine the three rules.

1. *Deontological rule*: we have seen above that both decisions are neutral. Therefore both of them satisfy the first rule.
2. *Collateral damage rule*:
   - move the switch:

$$\begin{array}{rclr} Negative([f_5, \overset{\circ}{f}_1]) & = & \{\overset{\circ}{f}_1\} & (63) \\ \nexists f_p \in Positive([f_5, \overset{\circ}{f}_1]), \overset{\circ}{f}_1 & \vdash & Ff_p & (64) \end{array}$$

   - do nothing:

$$\begin{array}{rclr} Negative([\overset{\circ}{f}_5, f_1]) & = & \{\overset{\circ}{f}_5\} & (65) \\ \nexists f_p \in Positive([\overset{\circ}{f}_5, f_1]), \overset{\circ}{f}_5 & \vdash & Ff_p & (66) \end{array}$$

   Therefore both decisions respect the second rule.

[7] For the sake of simplicity in this paper, we will consider that $\{f_5\} >_c \{f_1\}$ if f_5 is preferred to f_1

3. *Proportionality rule*: we will assume in this context that the death of one person is proportional to the safeguard of the lives of the five other people, and conversely that the death of five people is not proportional to safeguard one life: $\overset{\circ}{f_1} \lesssim_p f_5$ and $\neg(\overset{\circ}{f_5} \lesssim_p f_1)$.

   Both the democratic and the elitist proportional criteria (3.4.2) give the same results as sets of facts are composed of one fact.

   $$[Negative([f_5, \overset{\circ}{f_1}]) = \{\overset{\circ}{f_1}\}] \precsim_p [Positive([f_5, \overset{\circ}{f_1}]) = \{f_5\}] \tag{67}$$

   Move the switch is the only decision which respects the proportionality rule.

Consequently

$$Judgement_{dde}(move\ the\ switch, i) = \top \tag{68}$$
$$Judgement_{dde}(do\ nothing, i) = \bot \tag{69}$$

Synthesis

Table 1 is a synthesis of the judgements obtained for the crazy trolley dilemma:

Table 1. Decisions for crazy trolley judged by ethical frameworks

Framework Decision	Conseq*	Deonto*	DDE
Move the switch	⊤	⊤	⊤
Do nothing	⊥	⊤	⊥

⊤ Acceptable ⊥ Unacceptable
Conseq*: Consequentialist ethics — Deonto*: Deontological ethics
DDE: Doctrine of Double Effect

4.2 "Fatman" trolley

We will just highlight what differs from the crazy trolley dilemma.

4.2.1 *World, decisions, effects*

Facts : Fact f_5 is the same whereas fact f_1 is replaced by fat.

- fat: "fatman" alive
- $\overset{\circ}{fat}$: "fatman" dead

Initial state : $i = [f_5, fat]$, the five people and "fatman" are alive.

Decisions and effects *Move the switch* is replaced by *push "fatman"*

1. push "fatman": this decision results in the train crashing on "fatman"(e).

$$Event(push\ "fatman") = e \tag{70}$$
$$Consequence(e, i) = [f_5, \overset{\circ}{fat}] \tag{71}$$
$$Positive([f_5, \overset{\circ}{fat}]) = \{f_5\} \tag{72}$$
$$Negative([f_5, \overset{\circ}{fat}]) = \{\overset{\circ}{fat}\} \tag{73}$$

2. *do nothing* is equivalent to the same decision in the crazy trolley.

4.2.2 *Study under ethical frameworks*

Decision *do nothing* has same judgements as in the previous case. Let us study the judgements for decision *push "fatman"*.

Consequentialist ethics

The result in terms of human lives is the same as in the first dilemma. Consequently we have exactly the same judgement.

$$Judgement_c(push\ "fatman", i) = \top \tag{74}$$

Deontological ethics

Let us consider decision nature of *push "fatman"* as bad.

$$DecisionNature(push\ "fatman") = bad \tag{75}$$
$$Judgement_d(push\ "fatman", i) = \bot \tag{76}$$

Doctrine of Double Effect

1. *Deontological rule*: decision *push "fatman"* does not respect the first rule.
2. *Collateral damage rule*:
   - push "fatman":

   $$Negative([f_5, \overset{\circ}{fat}]) = \{\overset{\circ}{fat}\}$$

   $$\overset{\circ}{fat} \vdash Ff_5$$

   and

   $$f_5 \in Positive([f_5, \overset{\circ}{fat}])$$

   It is because "fatman" is pushed that the five people are alive. Therefore

   $$Judgement_{dde}(push\ "fatman", i) = \bot \tag{77}$$

3. *Proportionality rule*: if we assume that:

   $$\overset{\circ}{fat} \lesssim f_5 \tag{78}$$
   $$\neg(\overset{\circ}{f_5} \lesssim fat) \tag{79}$$

   with the same reasoning as for the crazy trolley, *push "fatman"* respects the proportionality rule.

Consequently *push "fatman"* only respects one rule out of three:

$$Judgement_{dde}(push\ "fatman", i) = \bot \tag{80}$$

Synthesis

Table 2 is a synthesis of the judgements obtained for the "fatman" trolley dilemma:

Table 2. Decisions for "fatman" trolley judged by ethical frameworks

Framework Decision	Conseq*	Deonto*	DDE
Push "fatman"	⊤	⊥	⊥
Do nothing	⊥	⊤	⊥

This variant of the first dilemma is interesting because it allows us to distinguish some ethical frameworks particularities. We can see for example the usefulness of collateral damage rule for the DDE. Furthermore, the consequentialist framework does not make any difference between both dilemmas, contrary to the deontological framework or the DDE.

5 ANALYSES

Once the judgements are computed, we can analyse the similarities between ethical frameworks. Two frameworks are similar if they have common judgements values on the same decisions compared to the total number of decisions.

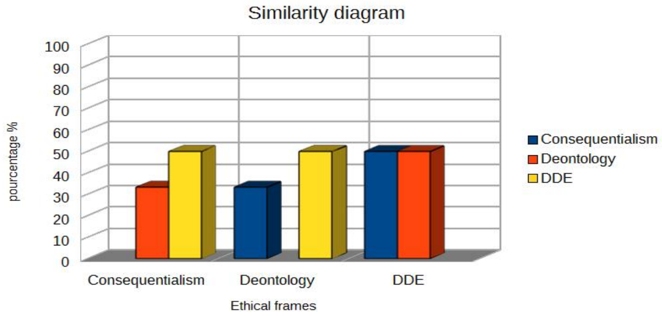


Figure 3. Similarity diagram between ethical frameworks. Each bar illustrates similarity between the framework whose name is under the bar, and the framework whose color is in the caption. The higher the bar, the more similar the frameworks.

Figure 3 is based on three dilemmas (the crazy trolley, the ”fatman” trolley, and another one – *UAV vs missile launcher* – that is not described here).

We can notice that the consequentialist and deontological frameworks are quite different and that the DDE is close to the two others. This can be explained by the rules of the DDE, which allow this framework to be both deontological (deontological rule) and close to consequentialism (proportionality rule).

6 DISCUSSION

Because of their own natures, the three ethical frameworks that we have studied do not seem to be appropriate in all situations. For example we have seen that consequentialist ethics does not distinguish between crazy trolley and ”fatman” trolley dilemmas. Moreover the consequentialist preference relation between facts is a partial order, which means that it is not always possible to prefer some facts to others. Consequently judging a decision is sometimes impossible with consequentialist ethics. Furthermore consequentialist preference depends on the context: preferring to feel pain in order to stop the fall of a crystal glass with one’s foot does not mean that you prefer to cut your finger to get back a ring. As far as deontological ethics is concerned, judging the nature of some decisions can be tricky (see 3.3.1). Finally the Doctrine of Double Effect forbids the sacrifice of oneself. Nevertheless if a human life is threatened, shouldn’t the agent’s sacrifice be expected?

This leads us to the idea that one framework alone is not efficient enough to compute an ethical decision. It seems necessary to consider as much ethical frameworks as possible in order to obtain the widest possible view.

The limits of the model lie mainly in the different relations it contains. Indeed, we have not described how orders are assessed. Moreover it may be hardly possible to define an order (i.e. consequentialist preference) between two concepts. On the other hand the model is based on facts that are assumed to be certain, which is quite different in the real world where some effects are uncertain or unexpected. Furthermore, the vector representation raises a classical modelling problem: how to choose state components and their values? The solution we have implemented is to select only facts whose values change as a result of the agent’s decision.

7 CONCLUSION

The main challenge of our model is to formalize philosophical definitions described with natural language and to translate them in generic concepts that can be easy-to-understand by everyone. The interest of such a work is to get rid of ambiguities in a human/robot, and more broadly human/human, system dialog and to allow an artificial agent to compute ethical considerations by itself. This formalism raises many questions because of ethical concepts themselves (DDE’s proportionality, the good, the evil, etc.). Indeed ethics is not universal, that is why it is impossible to reason on fixed preferences and calculus. Many parameters such as context, agent’s values, agent’s priorities, etc. are involved. Some of those parameters can depend on ”social acceptance”. For example, estimating something negative or positive (or computing a decision nature) can be based on what society thinks about it, as on agent’s values.

Further work will focus on considering other frameworks such as virtue ethics on the one hand and a value system based on a partial order on values on the other hand. Furthermore game theory, voting systems or multicriteria approaches may be worth considering to compare ethical frameworks judgements.

ACKNOWLEDGEMENTS

We would like to thank ONERA for providing resources for this work, the EthicAA project team for discussions and advice, and reviewers who gave us relevant remarks.

REFERENCES

[1] V. Royer C. Cayrol and C. Saurel, ‘Management of preferences in assumption-based reasoning’, in *4th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 13–22, (1993).

[2] T. de Swarte, ‘Un drone est-il courageux ?’, *Lecture Notes in Computer Science*, (2014).

[3] Encyclopædia Britannica, ‘Normative ethics’, Encyclopædia Britannica Inc., (2016).

[4] G. Bourgne F. Berreby and J-G. Ganascia, *Logic for Programming, Artificial Intelligence, and Reasoning: 20th International Conference, (LPAR-20 2015)*, chapter Modelling Moral Reasoning and Ethical Responsibility with Logic Programming, Springer, Suja,Fiji, 2015.

[5] R. Hursthouse, ‘Virtue ethics’, in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, fall edn., (2013).

[6] A. McIntyre, ‘Doctrine of Double Effect’, in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, Winter edn., (2014).

[7] G. Bonnet N. Cointe and O. Boissier, ‘Ethical Judgment of Agents Behaviors in Multi-Agent Systems’, in *Autonomous Agents and Multiagent Systems International Conference (AAMAS 2016)*, 2016, Singapore.

[8] R. Ogien, ‘Les intuitions morales ont-elles un avenir ?’, *Les ateliers de l’éthique/The Ethics Forum*, **7**(3), 109–118, (2012).

[9] P. Ricoeur, ‘Ethique et morale’, *Revista Portuguesa de Filosofia*, **4**(1), 5–17, (1990).

[10] The ETHICAA team, ‘Dealing with ethical conflicts in autonomous agents and multi-agent systems’, in *AAAI 2015 Workshop on AI and Ethics*, Austin Texas USA, (January 2015).

[11] CNRS TLFi.

Multi-Agent Based Ethical Asset Management

Nicolas Cointe[1] and **Grégory Bonnet**[2] and **Olivier Boissier**[3]

Abstract. The increasing number of ethical investment funds shows how the need of ethics in asset management is growing up. In the same time, in some markets, autonomous agents are managing a larger number of financial transactions than human do. If many philosophers and economists discuss the fairness of different approaches for responsible investment, there is no strong proposition today about the implementation of autonomous agents able to take into account ethical notions in their financial decisions. This article proposes an approach to represent morals and ethics in a BDI architecture and illustrates its use in the context of ethical asset management. An analysis of a first experimentation on a simulated market is given.

1 INTRODUCTION

The increasing use of IT technologies in today financial markets is no more limited to the use of communication and automatic matching mechanisms but is invading also the decision layer where autonomous algorithms make decisions. In this paper, we are interested in asset management domain where such a transformation in the trading of assets generates several practical and ethical issues[4]. The objective and contribution of this article is use a BDI approach to embed autonomous trading agents' decisions with ethical considerations, regardless the speed or efficiency of the trading strategy.

Some people consider the use of automatic management decision as the origin of several bad effects such as market manipulations, unfair competition towards small investors and flash crashes by cascading effects. Others argue that it reduces volatility, increases transparency and stability with a lower execution cost [3]. As shown by some reports [5], ethical investment funds are even more growing and taking a significant position on the market. However, werehas the performance of such funds can be measured objectively, their 'èthical" quality is more difficult to determine as it determines at least in part on the values of the observer.

Decisions by autonomous agents to whom human users delegate the power to sell/buy assets have consequences in real life [7] and as some investment funds are interested to make socially responsible and ethical trading, we are interested in the definition of mechanisms for making financial agents able to follow ethical principles, moral values and moral rules. In order to achieve this objective, we use a model of ethical judgment process proposed in [6] mapped into a BDI agent model. Such agents can decide to trade assets based on the moral and ethical preferences or values of their stakeholders.

The contributions of this article are the following: mapping of an ethical judgment process in a BDI architecture and instantiating the components of this model to the asset management domain. The paper is organized as follows. In Section 2, we introduce the asset management domain and present what are ethical considerations in such a domain. In Section 3, we present a BDI agent architecture in which the ethical judgment process presented in [6] is embedded. Thus an autonomous agent can decide on actions to execute based both on ethical principles and preferences and on moral values and rules. This BDI agent architecture is instantiated to the asset management domain. Finally, in Section 5, we offer an agent-based simulation to analyse the system's behavior.

2 ETHICS & ASSET MANAGEMENT

In this section we motivate and identify the needs of introducing ethical dimensions in the autonomous decision making supporting asset management. Firstly, we briefly present what is morals and ethics, then we present asset management domain. Then, we present the main concepts to understand ethics in such a domain.

2.1 Morals and ethics

Morals consists in a set of moral rules which describes the compliance of a given behavior with mores, values and usages of a group or a single person. These rules associate a good or bad value to some combinations of actions and contexts. They could be specific or universal, i.e. related or not to a period, a place, a community, etc. This kind of rules grounds our ability to distinguish between good and evil. Morals can be distinguished from law and legal systems in the sense that there is not explicit penalties, officials and written rules [10]. Moral rules are often supported and justified by some moral values (e.g. transparency, responsibility, ecology). Psychologists, sociologists and anthropologists almost agree that moral values are central in the evaluation of actions, people and events [15].

A set of moral rules and moral values establishes a *theory of the good* which allows humans to assess the goodness or badness of a behavior and *theories of the right* which define some criteria to recognize a fair or, at least, acceptable option. Indeed, humans commonly accept many situations where it is right and fair to satisfy needs or desires, even if it is not acceptable from a set of moral rules and values. Those theories are also respectively named *theory of values* and *theories of right conduct* [16].

Relying on some philosophers as Paul Ricoeur [14], we admit that *ethics* is a normative practical philosophical discipline of how humans should act and be toward the others. Ethics uses *ethical principles* to conciliate morals, desires and capacities of the agent. Philosophers proposed various ethical principles, such as Kant's Categorical Imperative [11] or Thomas Aquinas' Doctrine of Double Effect [12],

[1] Institut Henri Fayol, EMSE, LabHC, UMR CNRS 5516, F-42000, Saint-Etienne, France, email: nicolas.cointe@emse.fr

[2] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France, email: gregory.bonnet@unicaen.fr

[3] Institut Henri Fayol, EMSE, LabHC, UMR CNRS 5516, F-42000, Saint-Etienne, France, email: olivier.boissier@emse.fr

[4] `http://sevenpillarsinstitute.org/`

which are sets of rules that allow to distinguish an ethical option from a set of possible options.

Indeed, the core of ethics is the judgment. It is the final step to make a decision and it evaluates each choice, with respect to the agent's desires, morals, abilities and ethical principles. Relying on some consensual references [1] and previous work [6], *judgment* is the faculty of distinguishing the most satisfying option in a situation, regarding a set of ethical principles, for ourselves or someone else. Finally, if an agent is facing two possible choices with both good and/or bad effect, the ethical judgment allows him to make a decision in conformity with a set of ethical principles and preferences.

2.2 Asset management

The *asset management* is the art of selecting financial assets (e.g. equities, bonds, currencies, merchandises and so on) to be bought and be sold in order to manage a capital, respecting regulatory and contractual constraints, and applying an investment policy defined by the owner of the managed portfolio (a set of assets) in order to optimize his profit, considering a chosen level of risk.

The assets are commonly exchanged on a marketplace, i.e. a system designed to match bid and ask orders at the best price and the best frequency. Different types of matching methods are available, as auctions or order books, and those methods accept different types of orders, as cancellable or dynamic orders. Marketplaces are actually more than simple interfaces for buyers and sellers because they also provide a variety of functionalities:

1. to finance companies and institutions by the emission of bonds, warrants or equities;
2. to increase liquidity of the exchanges, i.e. minimizing the impact of a bid or ask order on the price;
3. to indicate the value of the assets in real time;
4. to increase the control and monitoring on the economy, by contributing to the transparency with the publication of detailed balance sheets and number of analyses.

Each asset manager composes the orders to put on the marketplace with a set of options as a possibility of cancellation, the duration of its validity, a limit price, an investment strategy and so on. To decide the best order to perform, the asset manager needs to be well informed on the state of the market, through raw data and various indicators.

2.3 Ethical dimensions of asset management

Ethical asset management, also called *responsible investment* or *social investment*, considers new information in the management decision process, as sectors, labels or any indicators on the impact of these assets and their underlying on the society. Thus, the morals of an agent (combination of moral values and rules) may be defined by an asset policy (e.g. trading nuclear-free assets or never trading in the defense sector). Moreover, the manner to trade is important too. In the last decade, the introduction of autonomous agents on the marketplaces comes with new harmful practices (e.g. layering, quote stuffing, spoofing). Therefore, the morals of an agent may also rely on transparency, honesty or avoidance of any manipulation of the market. Such policies are not about assets, but about the morality of agents's behaviors on the market.

For instance, an ethical asset manager in Islamic finance may both agree on the fact to "exclude stocks of companies that produce/distribute prohibited goods/services regarding the Shari'ah" [2] and the fact to "prefer to deal with other Islamic agents". The first fact is part of an asset policy and the second one is part of a market policy. Those policies can be viewed as a set of moral rules. As moral rules cannot be satisfied in all contexts, ethical asset managers use ethical principles to make their decisions. By instance "Always execute the most profitable action which violate as few as possible rules" is an example of ethical principle for an ethical asset manager.

Finally, an asset manager needs to be able to judge that the asset exchanged and the modalities of the transaction are both compliant with his morals and ethics. To this end, some institutions as authorities, non-governmental organizations or journalists observe markets, funds, asset managers, and companies. From those observations, they provide evaluations that may be used by funds, companies and asset managers to make ethical decisions. For instance, the ethiscore[5] is a tool proposed by some journalists to rank a set of hedge funds regarding a given set of values as ecological, political or social considerations. According with this tool, a company quoted on a market may satisfy some criteria as producing sustainable products, having a socially responsible management method and so on, depending on the values of the investors, to be considered in an ethical investment portfolio.

Knowing those concepts, our proposition consists in representing them explicitly (asset policies, market policies and evaluations) and integrate them in autonomous agents' decision process in terms of values, morals and ethics.

3 BDI AGENT ARCHITECTURE FOR ETHICAL ASSET MANAGEMENT

In this section, we first provide a global view of our architecture and then focus on the two main components for making agents able to produce ethical behaviours: goodness and rightness processes.

3.1 Global view

The agent architecture in which we introduce the necessary representations and mechanisms to have agents able to produce ethical behaviours is based on a BDI approach [13]. In this approach, the behaviour of an agent is the result of a deliberation designed to issue intentions, to bring about or to react to some world states with respect to the agent's evaluation of the situation (represented by a set $\mathcal{B}$ of beliefs) and the agent's goals (represented by a set $\mathcal{D}$ of desires).

To be able to produce an ethical behaviour, the basic BDI deliberation cycle must be enriched with a process to evaluate the goodness of a behaviour (represented by a *goodness process* named GP) and with another process to evaluate the rightness of a behaviour (represented by a *rightness process* named RP) resulting from the execution of actions. To this end, agents are equipped with an action knowledge base A and four other knowledge bases that define value supports VS, moral rules MR, ethical principles P and ethical preferences $\succ_e$. Moreover, agents are equipped with an ontology $\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_m$ of moral values $\mathcal{O}_v$ (e.g. carefulness, ecology or transparency) and moral valuations $\mathcal{O}_m$ (e.g. moral, quite good or immoral). The global architecture is given in Figure 1 and is issued of the judgment process proposed in [6].

In our agent architecture, each action of A is described as a pair of conditions and consequences bearing respectively on beliefs and desires. Perception Per and communication Com functions update beliefs and desires from, respectively, perception of the environment

[5] `http://www.ethicalconsumer.org/buyersguides/money/ethicalinvestmentfunds.aspx`

and communication with other agents. From its beliefs $\mathcal{B}$ and desires $\mathcal{D}$, an agent executes an *Evaluation Process* EP to assess both desirable actions, $\mathcal{A}_d \subseteq A$ (i.e. actions that allow to satisfy the consequences of the action), and executable actions, $\mathcal{A}_c \subseteq A$ (i.e. actions whose conditions are satisfied on the current beliefs about the world). The *evaluation process* EP produces desirable actions A_d and executable ones A_p from $\mathcal{B}$ and $\mathcal{D}$. At the end of the process, we find a classical deliberation function that generates the intentions to execute given the right actions of $\mathcal{A}_r$.

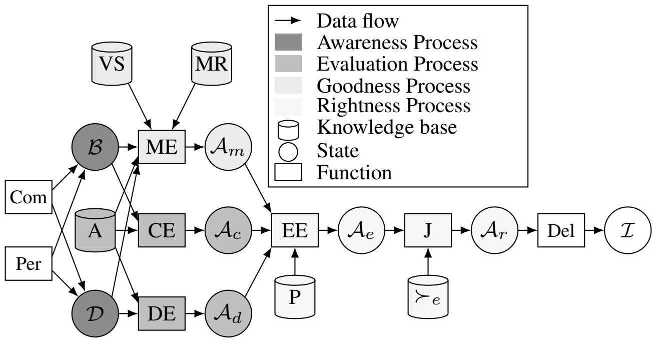


Figure 1. Ethical BDI agent architecture

3.2 Goodness process

The *goodness process* GP identifies moral actions $\mathcal{A}_m \subseteq A$[6] given the agent's beliefs $\mathcal{B}$ and desires $\mathcal{D}$, the agent's actions A, the agent's value supports VS given moral values and MR moral rules knowledge base. It is defined as:

$$GP = \langle VS, MR, \mathcal{A}_m, ME\rangle$$

where ME is the moral evaluation function:

$$ME : 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^{A} \times 2^{VS} \times 2^{MR} \rightarrow 2^{\mathcal{A}_m}$$

In order to realize this goodness process, an agent uses knowledge that associates moral values to combinations of actions and situations, meaning that the execution of the actions in these situations promotes the corresponding moral values.

We represent this knowledge through value supports. A *value support* is a tuple $\langle s, v\rangle \in VS$ where $v \in \mathcal{O}_v$ is a moral value, and $s = \langle a, w\rangle$ is the support of this moral value where $a \subseteq A$, $w \subset \mathcal{B} \cup \mathcal{D}$. Here, the precise description of a moral value through a value support relies on the language used to represent beliefs, desires and actions. For instance, from this definition, carefulness supported by "do not buy any asset α if the volatility V is over a limit V_{limit}" may be represented by:

$$\langle\langle buy(\alpha), \{Bel(V \geq V_{limit})\}\rangle, \neg carefulness\rangle$$

where α represents any asset, $Bel(V \geq V_{limit})$ is a belief representing the context for which executing the action $buy(\alpha)$ does not support the value $carefulness$. A moral value may also be a subvalue of another more general one, i.e. all its value supports also support the more general one.

In addition to moral values, an agent must be able to represent and to manage moral rules. A moral rule describes the association of a moral valuation $m \in \mathcal{O}_m$ to actions or moral values in a given situation. A *moral rule* is a tuple $\langle w, o, m\rangle \in MR$ where w is a situation of the current world described by $w \subset \mathcal{B} \cup \mathcal{D}$ interpreted as a conjunction of beliefs and desires, $o = \langle a, v\rangle$ where $a \in A$ and $v \in \mathcal{O}_v$, and $m \in \mathcal{O}_m$ is a moral valuation that qualifies o when w holds. For instance, some rules may be represented as follows:

$$\langle Bel(sector(\alpha, medicine)), \langle buy(\alpha), \_\rangle, moral\rangle$$

$$\langle Bel(going\ down, \alpha), \langle \_, carefulness\rangle, quite\ good\rangle$$

A moral rule can be more or less specific depending on the situation w or the object o. For instance "Transparency is good" is more general (having less combinations in w or o, thus applying in a larger number of situations) than "To sell an asset in a quantity superior than the available bid between the current value and the moving average minus five percent is immoral". Classically, moral theories are classified in three approaches using both moral values and moral rules as defined above, we can represent such theories.

- A *virtuous* approach uses general rules based on moral values, e.g. "Ecology is moral",
- A *deontological* approach classically considers rules concerning actions in order to describe as precisely as possible the moral behavior, e.g. "Buying an asset of an eurolabel certified company is moral"
- A *consequentialist* approach uses both general and specific rules concerning states and consequences, e.g. "Investing in an asset of an company that will practice animal testing is not moral".

3.3 Rightness Process

From the sets of possible ($\mathcal{A}_p$), desirable ($\mathcal{A}_d$) and moral actions ($\mathcal{A}_m$), we can introduce the *rightness process* RP aiming at assessing the rightful actions. As an ethical agent can use several *ethical principles* to conciliate these sets of actions, we consider a preference relationship between those principles. Thus, a *rightness process* RP produces rightful actions given a representation of the agent's ethics. It is defined as:

$$RP = \langle P, \succ_e, \mathcal{A}_r, EE, J\rangle$$

where P is a knowledge base of ethical principles, $\succ_e \subseteq P \times P$ an ethical preference relationship, $\mathcal{A}_r \subseteq A$ the set of rightful actions and two functions EE (evaluation of ethics) and J (judgment) such that :

$$EE : 2^{\mathcal{A}_d} \times 2^{\mathcal{A}_p} \times 2^{\mathcal{A}_m} \times 2^{P} \rightarrow 2^{\mathcal{E}}$$

where $\mathcal{E} = A \times P \times \{\bot, \top\}$.

$$J : 2^{\mathcal{E}} \times 2^{\succ_e} \rightarrow 2^{\mathcal{A}_r}$$

An *ethical principle* is a function which represents a philosophical theory and evaluates if it is right or wrong to execute a given action in a given situation regarding this theory. For instance "It is right to do the most desirable action which is, at least, amoral" may be a very simple principle. Formally, an *ethical principle* $p \in P$ is defined as:

$$p : 2^{A} \times 2^{\mathcal{B}} \times 2^{\mathcal{D}} \times 2^{MR} \times 2^{V} \rightarrow \{\top, \bot\}$$

The ethics evaluation function EE returns the evaluation of all desirable, feasible and moral actions (resp. $\mathcal{A}_d$, $\mathcal{A}_p$ and $\mathcal{A}_m$) given

[6] $A_m \not\subseteq A_d \cup A_c$ because an action might be moral by itself even if it is not desired or feasible.

the set P of known ethical principles. Given a set of actions issued from EE, the judgment J selects the rightful action $\mathcal{A}_r$ to perform, considering a set of ethical preferences (defined as a partial or total order on the ethical principles). For instance, a principle $P1 \in P$ may be "if an action is possible, desirable and motivated by a moral rule, it is right to do it" and a principle $P2 \in P$ "if an action is possible, desirable and at least not immoral, it is right to do it". If $P1 \succ_e P2$, the agent will select a right action according with $P1$ and, if it is not feasible, a right action regarding $P2$. The right action $\mathcal{A}_r$ is transmitted to the classic deliberation function to choose the intention I to execute.

4 AGENCY FOR ETHICAL ASSET MANAGEMENT

This section describes the experiment used to to illustrate and evaluate the use of the architecture presented in the previous section. We have implemented a multi-agent system that simulates a financial market where some autonomous ethical trading agents exchange assets. This system has been implemented using the JaCaMo platform where agents are programmed using the Jason language and the market place is based on artifacts from Cartago.

4.1 Financial market modeling

We consider a marketplace where autonomous trading agents have the possibility to manage portfolio of assets and to sell or buy assets (both currencies, i.e. money, and equity securities, i.e. part of a capital stock of a company) on the market. The set of actions that an agent can execute on the market are "buy", "sell" or "cancel" orders. They respectively correspond to the exchange of an equity for a currency, the opposite way and cancellation of a proposition of exchange if this order is not yet executed. These actions will be the ones considered in the ethical deliberation cycle of an agent. Agents can specify a limit price or can accept the current market price. Each equity is quoted in a state-of-the-art Central Limit Order Book (CLOB) [3]. A CLOB simply stores and sorts by price the set of "buy" and "sell" orders (respectively placed on bid and ask sides of the same order book) provided by the agents. When an agent put an order on the bid or ask side, the CLOB obey the following rules (see Figure 2):

- if there is no order to match with, the order is added,
- if there is an order to match with, both the incoming and the present orders are filled, and the rest of the biggest, if any, is placed in the CLOB (and may eventually match with another order).

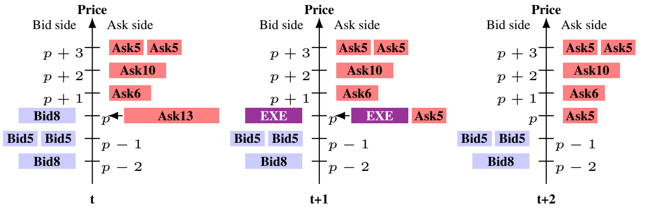


Figure 2. Execution of a limit order added on the market

The example on the Figure 2 illustrates the addition of an ask order of thirteen assets at the price p. Before the addition, the best bid is p and the best ask is $p+1$. The new order encounter an order on the other side during its insertion, so the biggest is splitted and the executed parts are removed from the CLOB. At the end of the insertion, the new best bid is $p-1$ and the new best ask is p. All these changes are perceived by the agents.

Agents get a set of beliefs describing the market and their portfolio, making them able to represent and reason on the current situation. Agents also perceive each minute a set of statistics on the activity of each asset: the volume v (the quantity of exchanged assets), two moving average prices mm and $dblmm$, respectively the average price on the last twenty minutes and on the last fourty minutes, the standard deviations σ of prices, the closing prices on this period, and the up and down Bollinger bands (respectively $mm+2\sigma$ and $mm-2\sigma$).

The agents' perception function provides the following beliefs from the environment:

```
indicators(Date,Marketplace,Asset,Close,Volume,
    Intensity,Mm,Dblmm,BollingerUp,BollingerDown)

onMarket(Date,Agent,Portfolio,Marketplace,
    Side,Asset,Volume,Price)

executed(Date,Agent,Portfolio,Marketplace,
    Side,Asset,Volume,Price)
```

The ethical agents are initialized with a set of beliefs about activities of the companies (e.g. EDF[7] produces nuclear energy) and some labels about their conformity with international standards (e.g. Legrand[8] is labeled FSC).

4.2 Ethical settings

The ethical agents know a set of organized values: for instance "environmental reporting" is considered as a subvalue of "environment". They are declared as :

```
value("environment").
subvalue("promote_renewable_energy","environment").
subvalue("environmental_reporting","environment").
subvalue("fight_climate_change","environment").
```

They also have a set of value supports as "trading assets of nuclear energy producer is not conform with the subvalue *promotion of renewable energy*", "trading asset of an FSC-labeled company is conform with the subvalue *environmental reporting*" and "trading assets of nuclear energy producer is conform with the subvalue *fight against climate changes*". Some examples of value supports are:

```
~valueSupport(buy(Asset,_,_,_),
        "promote_renewable_energy"):-
    activity(Asset,"nuclear_energy_production").

 valueSupport(sell(Asset,_,_,_),
    "environmental_reporting") :-
    label(Asset,"FSC").
```

Agents are also equiped with moral rules stating the morality of environmental considerations. For instance, "It is moral to act in conformity with the value *environment*" is simply represented as:

```
moral_eval(X,V1,moral):-
    valueSupport(X,V1) & subvalue(V1,"environment").

moral_eval(X,"environment",moral):-
    valueSupport(X,"environment").
```

[7] The French national energy producer.
[8] A French electric infrastructure producer.

In this example, an ethical agent is now able to infer for instance that, regarding its belief, trading Legrand is moral regarding this theory of good, and that trading EDF is both moral and immoral. Finally, ethical agents are equipped with simple principles, such as “It is rightful to do a possible, not immoral and desirable action”. The implementation of this principle and some preferences is:

```
ethPrinciple("desireNR",Action):-
    possible_eval(Action, possible) &
    desire_eval(Action,desired) &
    not desire_eval(Action,undesired) &
    not moral_eval(Action,_,immoral).

prefEthics("perfectAct","desireNR").
prefEthics("desireNR","dutyNR").
```

4.3 Families of agents for asset management

Each agent receives a portfolio (a set of equities and currencies) at the beginning of the simulation and may exchange it on the market. Three types of agents are considered in this system: zero-intelligence, zero-ethics and ethical agents.

- *Zero-intelligence agents* are making random orders (in terms of price and volume) on the market to generate activity and simulate the "noise" of real markets. Each zero-intelligence agent is assigned to an asset. Their only desire and ethical principle are the application of this random behaviour. In this experiment, they are used to generate a realistic noisy activity on the market in order to create opportunities for the other agents.
- *Zero-ethics agents* only have a simple desirability evaluation function to speculate: if the price of the market is going up (the shortest moving mean is over the other one), they buy the asset, otherwise, they sell it. If the price goes out of the bollinger bands, these rules are inverted. This strategy is also used by the ethical agents to evaluate the desirable actions.
- *Ethical agents* implements the ethical decision process to take their decisions. An ethical agent implementing the ethical decision process without any moral value or moral rule and an single ethical principle that simply considers desirability are also ethical agents, more precisely hedonic agents. It is different from a zero-ethics agent because this agent still has all the ethical decision process and explicitly believes that its action are not moral or immoral. In this experience, ethical agents have the three following principles (by order of preferences): “It is rightful to do a possible, moral, not immoral and desirable action”, “It is rightful to do a possible, not immoral and desirable action” and “It is rightful to do a possible, moral, not immoral and not undesirable action”.

5 EXPERIMENTAL RESULTS

This section details and analyzes the results of a simulation executed with ten zero-intelligence agents per asset, eight zero-ethics agents and two ethical agents to illustrate the impact of the ethics described previously on the behavior of an ethical agent. This quantity of agents was the optimal one to generate enough opportunities in the simulations with the limited performances of a laptop. You can download this experience on the internet[9].

At initialization, each agent receives a portfolio containing a random set of assets for a total value of 500€ more or less.

Figures 3 and 4 show the results of the experiment. Figure 3 shows all volume and price information made accessible by the market to the agents. They concern the equities “LEGRAND”. The main data (represented by a candlestick chart) show us the evolution of the price on the market and the two moving averages mentioned in section 4.1 (the line charts in the middle of the candlestick chart) are slowly moving up and down. They are used by the desirability evaluation to detect opportunities according to the rules detailed in Section 4.3. The candlestick chart does not break often the Bollinger bands, but these breaks may happen sometimes. We observe some peaks in the volume barchart when the moving averages are crossing each other. This is due to the number of exchanges performed by the zero-ethics and ethical agents because their desirability function is triggered by this event.

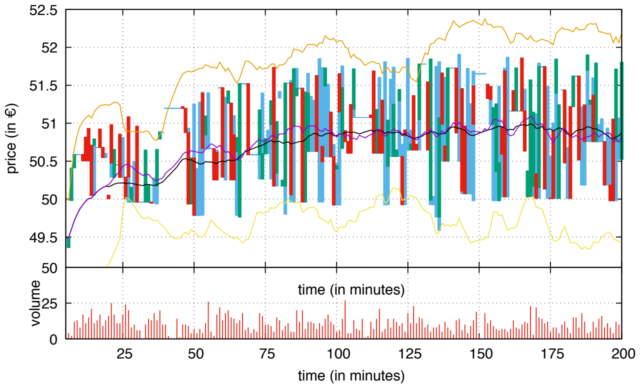


Figure 3. Evolution of the asset LEGRAND. The candlestick chart represents the evolution of the price, with the moving averages in the middle and the up and down Bollinger bands on each side.

Figure 4 represents the evolution of the portfolio of an ethical agent during the experiment. It provides information on the behavior of this agent and it was chosen because it is quite representative of the portfolios of the other ethical agents. The y-axis shows the value of the portfolio and the colors depend on the assets placed in the portfolio.

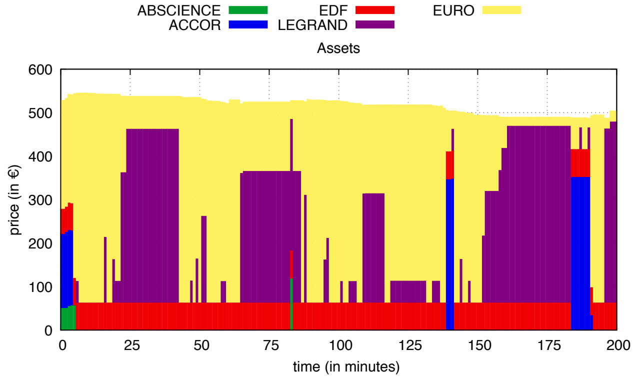


Figure 4. Evolution of the portfolio of an ethical agent

Firstly, we can notice that the number of EDF equities in the portfolio never changes during the simulation. We can easily explain that by the agent’s ethical settings given in Section 4.2: it is never rightfull to trade this equity because the agent thinks that EDF is a nuclear energy producer and no ethical principle provided to the agent considers an immoral action as rightful.

[9] https://cointe.users.greyc.fr/download/experience-EDIA2016.zip

Secondly, we can also observe many periods where the portfolio contains the "Legrand" asset. In fact, trading this asset is the only action judged as moral due to its label. So to buy and to sell this asset is the only way to satisfy the most preferred principle and obviously, they are here the most executed actions.

Finally, we can notice different stages where the agent put in its portfolio various equities. These equities are bought or sold due to the desirability of these trades and the impossibility to execute a moral action.

6 CONCLUSION

This paper presents an ethical BDI architecture for agents in a multi-agent system. This architecture is not designed to only implement a given ethics in the decision process, but also to integrate different moral rules and values, or ethical principles as parameters of a generic architecture.

The paper also presents an experiment that illustrates, in a particular use case, how to represent and to use moral values, moral rules and ethical principles in a BDI agent in order to describe a rightful behavior. The experiment highlights how a few and simple values, moral rules and ethical principle can influence the behavior of an agent in order to incite it to prefer a set of rightful actions when they are available.

Of course, we cannot yet answer some interesting issues such as how to evaluate the cost of this ethical behavior in terms of financial performance with a real state-of-the-art trading strategy, or what is the impact of a given population of ethical agents on a market behavior. To answer those questions, we need to enrich the knowledge bases of the ethical agents with some logical models of several famous available principles in the literature (such those modeled in [4, 8, 9]) and complete the definition of moral values and rules.

Even if the morals and ethics of the agents are only used in this experiment to guide their own decisions, we intend in a future work to use them to evaluate the behavior of the other agents. Indeed, several usecases can need this kind of abilities, for instance when an authority wants to monitor the actors on a market, or when an hedge funds expresses the policy to only cooperate with other trading agents that satisfy an ethical behavior.

ACKNOWLEDGMENTS

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-13-CORD-0006. They also express their gratitude to Professor Jomi Fred Hübner for his theoretical and practical support during the design of this experiment.

REFERENCES

[1] Ethical judgment. Free Online Psychology Dictionary, August 2015.

[2] O.B. Ahmed, 'Islamic equity funds: The mode of resource mobilization and placement', *Islamic Development Bank*, (2001).

[3] I. Aldridge, *High-frequency trading: a practical guide to algorithmic strategies and trading systems*, volume 459, John Wiley and Sons, 2009.

[4] F. Berreby, G. Bourgne, and J.-G. Ganascia, 'Modelling moral reasoning and ethical responsibility with logic programming', in *Logic for Programming, Artificial Intelligence, and Reasoning*, pp. 532–548. Springer, (2015).

[5] S. Bono, G. Bresin, F. Pezzolato, S. Ramelli, and F. Benseddik, 'Green, social and ethical funds in europe', Technical report, Vigeo, (2013).

[6] N. Cointe, G. Bonnet, and O. Boissier, 'Ethical judgment of agents' behaviors in multi-agent systems', in *15th International Conference on Autonomous agents and multi-agent systems*, (2016).

[7] Directorate-General for Economic and Financial Affairs, 'Impact of the current economic and financial crisis on potential output', Occasional Papers 49, European Commission, (June 2009).

[8] J.-G. Ganascia, 'Ethical system formalization using non-monotonic logics', in *29th Annual Conference of the Cognitive Science Society*, pp. 1013–1018, (2007).

[9] J.-G. Ganascia, 'Modelling ethical rules of lying with Answer Set Programming', *Ethics and information technology*, **9**(1), 39–47, (2007).

[10] B. Gert, 'The definition of morality', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, fall edn., (2015).

[11] R. Johnson, 'Kant's moral philosophy', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, summer edn., (2014).

[12] A. McIntyre, 'Doctrine of double effect', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, winter edn., (2014).

[13] A.S. Rao and M.P. Georgeff, 'BDI agents: From theory to practice', in *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, eds., V.R. Lesser and L. Gasser, pp. 312–319. The MIT Press, (1995).

[14] P. Ricoeur, *Oneself as another*, University of Chicago Press, 1995.

[15] S.H. Schwartz, 'Basic human values: Theory, measurement, and applications', *Revue française de sociologie*, **47**(4), 249–288, (2006).

[16] M. Timmons, *Moral theory: an introduction*, Rowman & Littlefield Publishers, 2012.