Proceedings of the 1st Workshop on Rights and Duties of Autonomous Agents (RDA2) 2012

Olivier Boisser, Grégory Bonnet, Catherine Tessier (eds.)

Workshop co-located with the
20th European Conference on Artificial Intelligence (ECAI 2012)
August 28th 2012, Montpellier, France

Preface

Olivier Boissier* Grégory Bonnet† Catherine Tessier‡

The autonomous decision capability embedded in software or robot agents is one of the major issues of Artificial Intelligence. It is a core property for artificial intelligence applications such as e-commerce, serious games, ambient computing, social or collective robotics, companion robots, unmanned vehicles.

Autonomous agents decide and act in a given context or environment and under domain constraints, and possibly interact with other agents or human beings e.g. to share tasks or to execute tasks on behalf of others. It is thus important to define regulation and control mechanisms to ensure sound and consistent behaviours both at the agent's individual level and at the multi-agent level. Organisation models, conversation policies, normative systems, constraints, logical frameworks address the problem of how agents' autonomous behaviours should be controlled, paving the way for formal or pragmatic definitions of agents' Rights and Duties. The issue is all the more important as autonomous agents may encounter new situations, evolve in open environments, interact with agents based on different design principles, act on behalf of human beings and share common resources. For instance: should an autonomous agent take over the control from a human operator? under which circumstances?

The aim of this workshop is to promote discussions and exchanges on the different issues raised by autonomous agents' Rights and Duties and models that can be proposed to represent and reason on Rights and Duties, namely:

- autonomous agents and privacy protection
- rights and duties for learning agents
- authority sharing between autonomous agents and human users or operators
- rights and duties of autonomous agents towards other agents; towards human users or operators
- rights and duties of human users or operators towards autonomous agents (especially robots)
- consistency, conflicts among rights and duties in multi-agent and human/agent systems
- mutual intelligibility, explanations
- rights and duties vs failures
- rights and duties of autonomous agents and ethical issues
- control of autonomous agents within organisations, institutions, normative systems
- sociology and law in the modelling of rights and duties: authority, power, dependence, penalty, contracts
- trust and reputation for autonomous agents regulation
- emergence and evolution of rights and duties
- knowledge representation and models for rights and duties
- reasoning on rights and duties

*ENS Mines Saint-Etienne, France, email: olivier.boissier@emse.fr
†University of Caen, UMR CNRS 6072 GREYC, France, email: gregory.bonnet@unicaen.fr
‡Onera, France, email: catherine.tessier@onera.fr

- validation of rights and duties in autonomous agents

Ten papers were submitted to RDA2, seven 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. Ethics and legal framework for rights and duties of autonomous agents (four papers)
2. Rights and duties conflict management by autonomous agents (three papers)

The RDA2 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. We are also very grateful to all the authors who submitted papers to RDA2 Workshop. We would like also to thank Antônio Carlos da Rocha Costa who has accepted to give an invited talk in the workshop. We would like also to thank the organizers of ECAI 2012.

Program committee

- Colin Allen, Indiana University, USA
- Estefania Argente Villaplana, UPV, Valencia, Spain
- Ronald Arkin, Georgia Tech, USA
- Laurence Cholvy, Onera, Toulouse, France
- Frédéric Dehais, ISAE, Toulouse, France
- Catholjin M. Jonker, Delft UT, The Netherlands
- Jean-Gabriel Ganascia, UPMC, Paris, France
- Sylvain Giroux, University of Sherbrooke, Canada
- Gwendal Le Grand, CNIL, Paris, France
- René Mandiau, Lamih Valenciennes, France
- Pablo Noriega, IIIA, Barcelona, Spain
- Eugénio Oliveira, Universidade do Porto, Portugal
- Xavier Parent, University of Luxembourg, Luxembourg
- Guillaume Piolle, Supélec, Rennes, France
- Jeremy Pitt, Imperial College London, UK
- Patrick Reignier , INRIA Rhônes-Alpes, Grenoble, France
- Jean Sallantin, LIRMM Montpellier, France
- Giovanni Sartor, European University Institute, Florence, Italy
- Munindar P. Singh, North Carolina State University, USA
- Axel Schulte, Bundeswehr University Munich, Germany
- Natalie van der Wal, Vrije Universiteit, The Netherlands
- Wamberto Vasconcelos, University of Aberdeen, UK
- Clara Smith, Universidad Nacional de La Plata, Argentina

Organization committee

- Olivier Boissier, ENS Mines, Saint-Etienne, France
- Grégory Bonnet, University of Caen, Caen, France
- Catherine Tessier, Onera, Toulouse, France

Contents

Functional rights and duties at the micro and macro social levels

Antonio Carlos da Rocha Costa[1] **(invited speaker)**

Abstract. This talk considers the issue of rights and duties in the context of social relations based on persistent exchange processes, occurring at both the micro and the macro social levels. Rights and duties that acquire a functional nature in such context are characterized in a tentative formal way. A possible connection between functional rights and duties and the issue of morality as a regulation mechanism for persistent micro social relations is investigated in a preliminary way. The relevance of functional rights and duties at the macro social level for the issue of the modularity of multiagent systems is indicated.

[1] Centro de Ciłncias Computacionais, Universidade Federal do Rio Grande, Brasil, email: ac.rocha.costa@gmail.com

Ethics and Authority Sharing for Autonomous Armed Robots

Florian Gros[1] and **Catherine Tessier**[1] and **Thierry Pichevin**[2]

Abstract. The goal of this paper is to review several ethical questions that are relevant to the use of autonomous armed robots and to authority sharing between such robots and the human operator. First, we discern the commonly confused meanings of morality and ethics. We continue by proposing leads to answer some of the most common ethical questions raised by literature, namely the autonomy, responsibility and moral status of autonomous robots, as well as their ability to reason ethically. We then present the possible advantages that authority sharing with the operator could provide with respect to these questions.

1 INTRODUCTION

There are many questions and controversies commonly raised by the use of increasingly autonomous robots, especially in military contexts [51]. In this domain autonomy is can be explored because of the need for reducing the atrocities of war, e.g. loss of human lives, violation of human rights, and for increasing battle performance to avoid unnecessary violence [3]. Since full autonomy is far from achieved, robots are usually supervised by human operators. This coupling between a human and a robotic agent involves a shared authority on the robot's resources [30], allowing for adaptability of the system in complex and dynamic battle contexts. Even with humans in the process, the deployment of autonomous armed robots raises ethical questions such as the responsibility of robots using lethal force incorrectly [47], the extent of their autonomous abilities and the related dangers, their ability to comply with a set of moral rules and to reason ethically [44], and the status of robots with regard to law due to the ever-increasing autonomy and human resemblance that robots display [28].

In this paper we will highlight the distinction between morality and ethics (section 2). Then several ethical issues raised by the deployment of autonomous armed robots, such as autonomy, responsibility, consciousness and moral status will be discussed (section 3). As another kind of ethical questions, a review of the frameworks used to implement *ethical* reasoning into autonomous armed robots will be presented afterwards (section 4). Finally, we will consider the ethical issues and implementations mentioned earlier in the framework of authority sharing between a robot and a human operator (section 5).

2 MORALITY AND ETHICS

The concepts of *morality* and *ethics* are often used in an identical fashion. If we want to talk about ethics for autonomous robots, we have to distinguish those terms and define them.

[1] Onera, the French Aerospace Lab, Toulouse, France, email: name.surname@onera.fr

[2] CREC, Ecoles de Saint-Cyr Coetquidan, France, email: thierry.pichevin@st-cyr.terre-net.defense.gouv.fr

2.1 Morality

If we ignore meta-ethical debates that aim at defining morality and its theoretical grounds precisely, we can conceive morality as principles of good or bad behaviour, an evaluation of an action in terms of right and wrong [52]. This evaluation can be considered either absolute or coming from a particular conception of life, a typical moral rule being "Killing is wrong". It is important to note that in this work, we focus on moral *action*, whether it results from rules, or from intentions of the subject doing the action.

2.2 Deontology and teleology

One of the bases for morality is the human constant need to believe in a meaning of one's actions. In most philosophical debates, this sense pertains to two often opposed categories : teleology and deontology.

For teleology, the moral action has to be good, the goal being to maximize the good and to minimize the evil produced by the action [33]. In this case, morality is commonly viewed as external to the agent, because it comes within the scope of a finalized world defining the rules and the possible actions and their goals, therefore defining the evaluation of actions.

For deontology, the moral action is done by duty, and must comply with rules regardless of the consequences of the action, whether they are foreseen or not, good or bad [34]. A case by case evaluation is not necessarily relevant here, because it is the humans' responsibility to dictate the rational and universal principles they want to live by.

2.3 Ethics

Ethics appears as soon as a conflict between existing legal or moral rules emerges, or when there is no rule to guide one's actions [36]. For example, if a soldier has received an order not to hurt any civilian, but to neutralize any armed person, what should he do if he encounters an armed civilian? We can thus consider ethics as the commitment to resolving moral controversies [13] where the agent, with good will, has to solve the conflicts he is faced with.

Those conflicts often oppose deontological and teleological principles, namely what has to be privileged between right and good ? The goal of ethics is not to pick one side and stand by it forever, but to be able to keep a balance between right and good when solving complex problems. Solving an ethical conflict then requires, apart from weighing good and evil, a sense of creativity in front of a complex situation and to be able to provide alternative solutions to moral rules imperatives [31].

To provide an illustration of the distinction between morality and ethics, we will consider that any moral conflict needs ethical reasoning abilites to be solved. Speaking of ethical rules would not make sense since ethics apply when rules are absent or in conflict.

3 AUTONOMY, RESPONSIBILITY, MORAL STATUS : PROSPECTS FOR ROBOTS

Technologies leave us presently in an intermediate position where robots can perceive their environment, act and make decisions by themselves, but lack a more complete kind of autonomy or the technological skill to be able to analyze their environment precisely and understand what happens in a given situation. Still research advances urge us to think about how to consider autonomous robots in a moral, legal and intellectual frame, both for the time being and when robots are actually skilled enough to be considered similar to humans. In this section, we will review important questions for autonomous robots i.e. autonomy, responsibility, moral status and see which answers are plausible. Then we will relate these questions to authority sharing.

3.1 Autonomy

3.1.1 Kant and the autonomy of will

When considering autonomy, one of the most influential view in occidental culture is Kant's. For him, human beings bend reality to themselves with their perception and reason, they escape natural or divine laws. Only reason enables humans to create laws that will determine humankind. Then laws cannot depend on external circumstances as reason only can provide indications in order to determine what is right or wrong. Consequently laws have to be created by a *good will*, i.e. a will imposing rules on itself not to satisfy an interest, but by duty towards other humans. Therefore no purpose can be external to humankind, and laws are meaningful to humans only if they are universal. This leads to the well-known moral "categorical" imperative[3], that immediatly determines what it orders because it enounces only the idea of an universal law and the necessity for the will to follow it [39].

Humans being the authors of the law they obey, it is possible to consider them as an end, and the will as autonomous. Thus, to be universal, a law has to respect humans as ends in themselves, inducing a change in the categorical imperative. If the law was external to humans, they would not be ends in themselves, but mere instruments used by another entity. Such a statement would deny the human ability to escape divine or natural laws, which is not acceptable for the kantian theory. We can only conceive law as completely universal, respecting humans as ends in themselves. To sum up, the kantian autonomy is the ability for an agent to define his own laws as ways to fulfill his goals and to govern his own actions.

3.1.2 Autonomy and robots

In the case of an Unmanned System, autonomy usually stands for decisional autonomy. It can be defined as the ability for an agent to minimize the need for supervision and to evolve alone in its environment [43], or more precisely, its "own ability of sensing, perceiving, analyzing, communicating, planning, decision making, and acting/executing, to achieve its goals as assigned by its human operators" [21].

[3] "act only according to that maxim by which you can at the same time will that it be a universal law"

We can see a difference between those definitions and Kant's. Robot autonomy is perceived differently for robots than for humans, as an autonomy of means, not of end. The reason for this is that robots are not sophisticated enough to be able to define their own goals and to achieve them. Robots are therefore viewed as mere tools whose autonomy is only intended to alleviate the operators' workload.

Consequently, to be envisioned as really autonomous, robots should be able to determine their own goals once deployed, thus to have will and be ends in themselves. The real question to ask here is if it is really desirable to build such fully autonomous robots, especially if they are to be used on a battlefield. If the objective is solely to display better performance than human soldiers, full autonomy is probably inappropriate, since being able to control robots and their goals from the beginning to the end of their deployment is one of the main reasons for actually using them.

3.2 Responsibility

If we want to use autonomous robots, we have to know to what extent a subject is considered responsible for his actions. It is especially important when applied to armed robots, since they can be involved in accidents where lives are at stake.

3.2.1 Philosophical approaches to responsibility

Classically responsibility has been considered from a broad variety of angles, whether being a relationship to every other human being in order to achieve a goal of salvation given by a divine entity (Augustine of Hippo), a logic consequence of the application of the categorical imperative (Kant), a duty towards the whole humanity as the only way to give a sense, a determination to one's actions and to define oneself in the common human condition (Sartre, [42]), or an obligation to maintain human life on Earth as long as possible by one's actions (Jonas, [22]).

The problem with those approaches is that they are thought for humans and consequently they require, more or less, an autonomy of end. As discussed above, this is not a direct possibility for robots. We then need to envision robot responsibility in their own "area" of autonomy, namely an autonomy of means, where the actions are not performed by humans. To discuss this problem, it is necessary to distinguish two types of responsibility : causal responsibility and moral responsibility.

3.2.2 Causal responsibility vs. moral responsibility

By moral responsibility, we mean the ability, for a conscious and willing agent, to make a decision without referring to a higher authority, to give the purposes of his actions, and to be judged by these purposes. To sum up, the agent has to possess a high-level intentionality [12]. This moral responsibility is not to be confused with causal responsibility, which establishes the share of a subject (or an object) in a causal chain of events. The former is the responsibility of a soldier who willingly shot an innocent person, the latter is the responsibility of a malfunctioning toaster that started a fire in a house.

Every robot has some kind of causal responsibility. Still, trying to determine the causal responsibility of a robot (or of any agent) for a given event is way too complex because it requires to analyze every action the robot did that could have led to this event. What we are really interested in is to define what would endow robots with a *moral* responsibility for their actions.

3.2.3 Reduced responsibility, a solution ?

Some approaches that are currently considered for the responsibility of autonomous robots are based on their status of "tools", not of autonomous agents. Thus, their share of responsibility is reduced or transferred to another agent.

The first approach is to consider robots as any product manufactured and designed by an industry. In case of a failure, the responsibility of the industry (as a moral person) is substituted to the responsibility of the robot. The relevant legal term here is *negligence* [24]. It implies that manufacturers and designers have failed to do what was legally or morally required, thus can be held accountable of the damage caused by their product. The downside of this approach is that it can lean towards a causal responsibility which – as said earlier – is more difficult to assess than a moral responsibility. Besides, developing a robot that is sure *enough* to be used on a battlefield would demand too much time for it to represent a good business, and it wouldn't even be enough to be safely used, a margin of error still existing no matter how sophisticated a robot is.

Another approach then would be to apply the *slave morality* to autonomous robots [24] [28]. A slave, by itself, is not considered responsible for his actions, but his master is. At a legal level, it is considered as *vicarious liability*, illustrated by the well-known maxim *Qui facit per alium facit per se*[4]. If we want to apply this to autonomous armed robots, their responsibility would be substituted to their nearest master, namely the closest person in the chain of command who decided and authorized the deployment of the robots. This way, a precise person takes responsibility for the robots actions, which spares investigations through the chain of command to assess causal responsibilities.

Finally, if we consider an autonomous robot to be able to comply with some moral rules, to reason as well as to act, it is possible to envision the robot as possessing, not moral responsibility, but moral intelligence [5]. The robotic agent is then considered to be able to adhere to an ethical system. Therefore there is a particular morality within the robot that is specific to the task it is designed for.

3.2.4 Other leads for a moral responsibility

No robot has been meeting the necessary requirements for moral responsibility, and no law has been specifically written for robots. The question is then to determine what is necessary for robots to achieve moral responsibility and what to do when they break laws.

For [19] and [1], the key to moral responsibility is the access to a moral status. Besides an emotional system, this requires the ability of rational deliberation, allowing oneself to *know* what one is doing, to be conscious of one's actions in addition to make decisions. Severals leads for robots to access to a moral status are detailed in the next section.

As far as responsibility is concerned, a commonly used argument is that robots cannot achieve moral responsibility because they cannot suffer, and therefore cannot be punished [47]. Still, if we consider punishment for what it is, i.e. a convenient way to change (or to compensate for) a behaviour deemed undesirable or unlawful, we can agree that it is not the *sine qua non* requirement for responsibility. There are other ways to change one's behaviour, one of the most known examples being treatment, i.e. spotting the "component" that produces the unwanted behaviour and tweak it or replace it to correct the problem [28]. Beating one's own car because of a malfunction would be absurd, in this case it is more fitting to replace the malfunctioning component. The same applies with certain types of law infringement (leading to psychological treatment or therapy), so it could apply to robots as well, e.g. by changing the program of the defective vehicle. Waiting for technology to progress to finally being able to punish robots so that they could have moral responsibility is not a desirable solution, but using vicarious liability, treatment and moral status appears to be a sound basis.

[4] "He who acts through another does the act himself."

3.3 Consciousness and moral status for autonomous robots

We have said earlier that for a robot to be considered responsible for its actions, it must be attributed a moral status, so it needs consciousness [19]. The purpose of this section is to see how this can be achieved and how moral status can be applicable to robots in order to help them to have moral responsibility.

3.3.1 Consciousness

Since there is an abundant literature on the topic of consciousness, and still no real consensus among the scientific community on how define consciousness, the purpose of this section is not to give an exhaustive nor accurate definition of consciousness, but merely to see what seems relevant to robots. However, if we want to use consciousness, we can consider it as described by [32], namely the ability to know *what it is like* to have such or such mental state from one's own perspective, to subjectively experience one's own environment and internal states.

The first approach for robots consciousness is the theory of mind [38] [6]. It is based on the assumption that humans tend to grant intentionnality to any being displaying enough similarities of action with them (emotions ou functional use of language). It is then possible for humans, by analogy with their experience of their own consciousness, to assume that those beings have a consciousness as well. This approach is already developing with conversational agents or robots mimicking emotions, even if it can be viewed as a trick of human reasoning more than an "absolutely true" model of consciousness.

The second approach considers consciousness as a purely biological phenomenon, and has gained influence with the numerous discoveries of neurosciences. Even if we do not know what really explains consciousness (see the Hard problem of consciousness [9]), considering it as a property of the brain may allow conscious robots to be developed, as did [55] [54] by recreating a brain from collected brain cells. There is still a lot of work to do here, as well as many ethical questions to answer, but it definitely looks promising. Indeed, if a being, even with a robotic body, has a brain that is similar to a human's, in a materialist perspective, this being is conscious.

The last approach is the one proposed by [25] [26] to build self-aware robots that can explore their own physical capacities to find their own model and to determine their own way to move accordingly. Those robots are probably the closest ones to consciousness as defined by [32]. They are still far from being used on a battlefield, but this method of self-modelling could be applied to more "evolved" robots for ethical decision-making. This way a robot could explore its own capacities for action and could build an ethical model of itself.

3.3.2 Moral status

An individual is granted moral status if it has to be treated never as a means, but only as an end, as prescribed by Kant's categorical imperative. To define this moral status, two criteria are commonly

used [7], namely sentience (or *qualia*, the ability to experience reality as a subject) and sapience (a set of abilities associated with high-level intelligence). Still, none of those attributes have been successfully implemented in robots. Even though it could be counter-productive to integrate *qualia* to robots in some situations (e.g. coding fear into an armed robot), it can be interesting to model some of them into robots, like [4] did for moral emotions like guilt. This could provide a solid ground for access of robots to moral status. [7] have proposed two principles stating that two different agents can have the same moral status if they possess enough similarities : if two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation (Principle of Substrate Non-Discrimination) or on how they came to existence (Principle of Ontogeny Non-Discrimination), then they have the same moral status.

Put simply, those principles are pretty similar to what the theory of mind proposes, that is if robots can exhibit the same functions as human's, then they can be considered as having a moral status, no matter what their body is made of (silicon, flesh, etc.) or how they matured (through gestation or coding). Still, proving that robots can have the same conscious experience as humans is currently impossible, so we can consider a more applicable version of those principles: [49] proposes that robots have moral agency if they are responsible with respect to another moral agent, if they possess a relative level of autonomy and if they can show intentional behaviour. This definition is vague but is grounded on the fact that moral status is attributed. What matters is that the robot is advanced enough to be similar to humans, but it does not have to be identical.

Another solution for autonomous robots with a moral status is to create a sort of Turing Test comparing the respective "value" of a human life with the existence of a robot. This is called by [46] the Triage Turing Test and shows that robots will have the same moral status as humans when it is at least as wrong to "kill" a robot as to kill a human. Advanced reflections on this topic can be found in [48].

4 IMPLEMENTING ETHICAL REASONING INTO AUTONOMOUS ARMED ROBOTS

Another question related to autonomous armed robots is how those robots can solve ethical problems on the battlefield and make the most ethically satisfying decision. In this section, we will briefly review several frameworks to integrate ethical reasoning into robots.

Three kinds of approaches are considered:

- Top-down : these approaches take a particular ethical theory and create algorithms for the robot, allowing it to follow the aforesaid theory. This is convenient to implement, e.g. a deontological morality into a robot.
- Bottom-up : the goal is to create an environment wherein the robot can explore different courses of action, with rewards to make it lean towards morally satisfying actions. Those approaches focus on the autonomous robot learning its own ethical reasoning abilities.
- Hybrid : these approaches look for a merge between top-down and bottom-up frameworks, combining their advantages without their downsides.

4.1 Top-down approaches

Top-down frameworks are the most studied in the field of ethics for robots and the number of ethical theories involved is high. Literature identifies theories such as utilitarianism [10], divine-command ethics [8] and other logic-based frameworks [27] [15]. Still, the most famous theory among top-down approaches is the Just-War Theory [35], which underlies the instructions and principles issued in the Laws of War and the Rules of Engagement (for more on these documents, see [3]). Those approaches have in common to take a set of rules and to program them into the robot code so that their behaviour could not violate them. The upside of those approaches is that the rules are general, well-defined and easily understandable. The downside is that no set of rules will ever handle every possible situation, mostly because they do not take into account the context of the particular mission the robot is deployed for. Thus top-down approaches are usually too rigid and not precise enough to be applicable. Also, since they rely on specific rules – more morality-like than ethics-like – they are not fit to capture ethical reasoning abilities but they are usually used to justify one's own actions. In order to implement ethical reasoning abilities into robots, it seems more desirable to use top-down approaches as moral heuristics guiding ethical reasoning [53].

4.2 Bottom-up approaches

Bottom-up frameworks are way less developed than top-down approaches. Still, some research like [26] gives interesting options, using self-modeling. Most of the bottom-up approaches insist on machine learning [17] or artificial evolution using genetic algorithms based on cooperation [45] to allow agents to reason ethically given a specific parameter. The strength of these frameworks is that learning allows flexibility and adaptability in complex and dynamic environments, which is a real advantage in the field of ethics wherein there is no predefined answers. Nevertheless the learning process takes a lot of time and never completely removes the risk of unwanted behaviour. Plus, the reasoning behind the action produced by the robot cannot be traced, making the fix of undesirable behaviours barely possible.

4.3 Hybrid approaches

Three different frameworks can be distinguished among hybrid approaches : case-based approach [29] [2], virtue ethics [24] [53] and the hybrid reactive/deliberative architecture proposed by [3], using the Laws of War and the Rules of Engagement as a set of rules to follow. They are probably the most applicable researches to autonomous robots and combine aspects of both top-down (producing algorithms derived from ethical theories) and bottom-up (using agents able to learn, evolve and explore possible ethical decisions) specifications. The main problem with these approaches is their computing time, since learning is often involved in the process. Nevertheless, they appear theoretically satisfying and their applicability looks promising.

5 ETHICS AND AUTHORITY SHARING

In this section we will focus on the previously mentioned ethical issues in the framework of authority sharing between a robot and a human operator.

Joining human and machine abilities aims at increasing the range of actions of "autonomous" systems [23]. However the relationship between both agents is dissymmetric since the human operator's "failures" are often neglected when designing the system. Moreover simultaneous decisions and actions of the artificial and the human agents are likely to create conflicts [11]: unexpected or misunderstood authority changes may lead to inefficient, dangerous or catastrophic situations. Therefore in order to consider the human agent and the artificial agent in the same way [20] and the human-machine

system as a whole [56], it seems more relevant to work on authority and authority control [30] than on autonomy, which concerns the artificial agent exclusively.

Therefore authority sharing between a robot and its operator can be viewed as an “upgraded” autonomy. As far as ethical issues are concerned, authority sharing considered as a relation between two agents [18] may provide a better compliance with sets of laws and moral rules, this way enabling ethical decision-making within a pair of agents instead of leaving this ability to only one individual.

5.1 Autonomy

As previously mentioned, the autonomy of an armed robot can be conceived as an autonomy of means only; robots are almost always used as tools. Authority sharing can bring a change in this organization. As a robot cannot (yet) determine its own goals, it is the human operator's role to provide the goals so as some methods or partial plans to achieve them [14]. Still, authority sharing allows the robot to be granted decision-making power allowing it to take authority from the operator to accomplish some tasks neglected by him (e.g., going back to base because of a fuel shortage) or even when the operator's actions are not following the mission plan and may be dangerous. For example, some undesirable psychological and physiological ”states” of the operator, e.g. tiredness, stress, attentional blindness [37] can be detected by the robot, in order to allow it to take authority if the operator is not considered able to fulfill the mission anymore.

5.2 Moral responsibility

Concerning moral responsibility, authority sharing forces us to make a distinction between two instances : the one where the operator has authority over the robot, and the reverse one. The former is simple; since the robot is a tool, we use the vicarious liability, therefore the operator engages his responsibility for any accident caused by the use of the robot that could happen during the mission. The latter is more complex and we do not claim to give absolute answers, but mere propositions.

What we propose is that, in order to assess moral responsibility when the robotic agent has authority over the system, it is necessary to define a mission-relevant set of rules, e.g. Laws of War and Rules of Engagement [35] [3], and a contract, as proposed by [41] or [40], between robotic and human agents, providing specific clauses for them to respect during the mission. These clauses must to be based on the set of rules previously mentioned, and an agent who violates them would be morally responsible of any accident that could happen as a consequence of his actions.

This kind of contract would provide clear conditions for authority sharing (i.e., an agent loses authority if he violates the contract) and could open the way to apply works on trust [4] or persuasion [16] in robotic agents. During a mission, such contracts would engage both agents to monitor the actions of the other agent and, if possible, to take authority if this can prevent any infringement of the contract. If one agent detects a possibly incoming accident due to the other agent's actions, e.g. aiming at a civilian, and does nothing to prevent it, then this agent is responsible for this accident as much as the one causing it. Because of the current state of law, i.e. dealing only with human behaviours, if a robot is considered responsible for ”evil” or unlawful actions, then it should be treated by replacing the parts of its program or the pieces of hardware that caused the unwanted behaviour. Human operators, if displaying the same kind of unlawful behaviour, should be judged by the appropriate laws. To integrate contracts in a concrete way, we can lean towards the perspective presented by [3] who proposes some recommendations to warn the operator of his responsibility when using potentially lethal force.

5.3 Consciousness and moral status

Authority sharing is not of a great help to implement consciousness into robots. Still, [37] and [50] provide leads to allow robots to assess the ”state” of the operator and to take authority from him if he is not considered able to achieve the mission. This approach would help robots to improve their situational awareness and to design systems that are better at interacting with humans, either operator or civilians. Enhancing the responsibility and autonomy of robots could also be a way to push them towards the ”same functionality” proposed by [7], i.e. acting with enough caution to be considered equals to humans in a specific domain, thus helping to give a moral status to robots.

5.4 Ethical reasoning

Given the current state of law and the common deployment of robots on battlefields, granting robots with ethical reasoning have to be rooted in a legally relevant framework, that is Just-War Theory [35]. Laws of War and Rules of Engagement have to be the basic set of rules for robots. Still, battlefields being complex environments ethics needs to be integrated into robots with a hybrid approach combining learning capabilities and experience with ethical theories. In the case of authority sharing, two frameworks seem relevant at the moment : case-based reasoning [2] and Arkin's reactive/deliberative architecture [3]. What seems applicable in case of an ethical conflict is to give the authority to the operator and to use the robotic agent both to assist him during the reasoning, i.e. by displaying relevant information on an appropriate interface, and to act as an ethical handrail in order to make sure that the principles of the Laws of War, e.g. discrimination or proportionality, are respected.

6 CONCLUSION AND FURTHER WORK

The main drawback of the implementation of ethics into autonomous armed robots is that, even if the technology, the autonomy and the lethal power of robots increase, the legal and philosophical frameworks do not take them into account, or consider them only from an anthropocentric point of view. Authority sharing allow a coupling between a robot and a human operator, hence a better compliance with ethical and legal requirements for the use of autonomous robots on battlefields. It can be achieved with vicarious liability, a good situational awareness produced by tracking both the robot and the operator's ”states”, and a hybrid model of ethical reasoning – allowing adaptability in complex battlefields environments.

We are currently building an experimental protocol in order to test some of our proposals, namely automous armed robots that embed ethical reasoning while sharing authority with a human operator. We have constructed two fully-simulated battlefield scenarios in which we will test the compliance of the system with a specific principle of the Laws of War (proportionality and discrimination). These scenarios feature hostile actions done towards the robot or its allies, e.g. throwing rocks or planting explosives, that need to be handled while complying with a set of rules of engagement. During the simulation, the operator is induced to produce an immoral behaviour, provoking an authority conflict in which we expect the robot to detect the said behaviour and to take authority from the operator: the authority conflict thereby generated has to be solved by the robot via the production of a morally correct behaviour. Since the current state of our

software does not yet allow the robotic agent to actually observe the operator, we are working on some pre-defined evaluations of actions in order for the robot to be able to detect unwanted behaviours, and to act accordingly.

REFERENCES

[1] K. Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, 35–52, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[2] M. Anderson, S. Anderson, and C. Armen, 'An approach to computing ethics', in *IEEE Intelligent Systems*, pp. 56–63, (July/August 2006).

[3] R.C. Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', Technical report, Georgia Institute of Technology, (2007).

[4] R.C. Arkin, P. Ulam, and A.R. Wagner, 'Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception', in *Proceedings of the IEEE*, volume 100, pp. 571–589, (2011).

[5] P. Asaro, 'What should we want from a robot ethic?', *International Review of Information Ethics*, **Vol. 6**, 9–16, (Dec. 2006).

[6] S. Baron-Cohen, 'The development of a theory of mind in autism: deviance and delay?', *Psychiatrics Clinics of North America*, **14**, 33–51, (1991).

[7] N. Bostrom and E. Yudkowsky. The Ethics of Artificial Intelligence. Draft for Cambridge Handbook of Artificial Intelligence, 2011.

[8] S. Bringsjord and J. Taylor, *The Divine-Command Approach to Robot Ethics*, 85–108, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[9] D.J. Chalmers, 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2**(3), 200–219, (1995).

[10] C. Cloos, 'The utilibot project: An autonomous mobile robot based on utilitarianism', in *2005 AAAI Fall Symposium on Machine Ethics*, (2005).

[11] Fr. Dehais, C. Tessier, and L. Chaudron, 'Ghost: Experimenting conflicts countermeasures in the pilot's activity', in *IJCAI'03*, Acapulco, Mexico, (2003).

[12] D. Dennett, *When HAL Kills, Who's to Blame?*, chapter 16, MIT Press, 1996.

[13] H.T. Engelhardt, *The Foundations of Bioethics*, Oxford Uninversity Press, Oxford, 1986.

[14] K. Erol, J. Hendler, and D. Nau, 'HTN planning: complexity and expressivity', in *AAAI'94*, Seattle, WA, USA, (1994).

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

[16] M Guerini and O. Stock, 'Towards ethical persuasive agents', in *IJCAI Workshop on Computational Models of Natural*, (2005).

[17] G. Harman and S. Kulkarni, *Reliable Reasoning: Induction and Statistical Learning Theory*, MIT Press, 2007.

[18] H. Hexmoor, C. Castelfranchi, and R. Falcone, *Agent Autonomy*, Kluwer Academic Publishers, 2003.

[19] K. Himma, 'Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?', in *7th International Computer Ethics Conference*, San Diego, CA, USA, (July 2007).

[20] *Handbook of cognitive task design*, ed., E. Hollnagel, Mahwah, NJ: Erlbaum, 2003.

[21] H. Huang, K. Pavek, B. Novak, J. Albus, and E. Messin, 'A framework for autonomy levels for unmanned systems ALFUS', in *AUVSIs Unmanned Systems North America 2005*, Baltimore, MD, USA, (2005).

[22] H. Jonas, *Das Prinzip Verantwortung. Versuch einer Ethik fr die technologische Zivilisation*, Insel Verlag, Frankfurt, 1979.

[23] D. Kortenkamp, P. Bonasso, D. Ryan, and D. Schreckenghost, 'Adjustable autonomy for human-centered autonomous systems', in *Proceedings of the AAAI 1997 Spring Symposium on Mixed Initiative Interaction*, (1997).

[24] P. Lin, G. Bekey, and K. Abney, 'Autonomous military robotics: Risk, ethics, and design', Technical report, California Polytechnic State University, (2008).

[25] H. Lipson, J. Bongard, and V. Zykov, 'Resilient machines through continuous self-modeling', *Science*, **314**(5802), 1118–1121, (2006).

[26] H. Lipson and J.C. Zagal, 'Self-reflection in evolutionary robotics: Resilient adaptation with a minimum of physical exploration', in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 2179–2188, (2009).

[27] G.J. Lokhorst, 'Computational meta-ethics: Towards the meta-ethical robot', *Minds and machines*, **6**, 261–274, (2011).

[28] G.J. Lokhorst and J. van den Hoven, *Responsibility for Military Robots*, 145–156, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[29] B. McLaren, 'Computational models of ethical reasoning: Challenges, initial steps, and future directions', in *IEEE Intelligent Systems*, pp. 29–37, (July/August 2006).

[30] S. Mercier, C. Tessier, and F. Dehais, 'Dtection et rsolution de conflits dautorit dans un systme homme-robot', *Revue dIntelligence Artificielle, numro spcial 'Droits et Devoirs dAgents Autonomes'*, **24**, 325–356, (2010).

[31] S. Miller and M. Selgelid, *Ethical and Philosophical Consideration of the Dual-Use Dilemma in the Biological Sciences*, Springer, New York, 2009.

[32] T. Nagel, 'What is it like to be a bat?', *The Philosophical Review*, **83**(4), 435–450, (1974).

[33] *Teleological Language in the Life Sciences*, ed., L. Nissen, Rowman and Littlefield, 1997.

[34] R.G. Olson, *Deontological Ethics*, The Encyclopedia of Philosophy, Collier Macmillan, London, 1967.

[35] B. Orend, *The Morality of War*, Broadview Press, Peterborough, Ontario, 2006.

[36] T. Pichevin, 'Drones arms et thique', in *Penser la robotisation du champ de bataille*, ed., D. Danet, Saint-Cyr, (November 2011). Economica.

[37] S. Pizziol, F. Dehais, and C. Tessier, 'Towards human operator state assessment', in *1st ATACCS (Automation in Command and Control Systems)*, Barcelona, Spain, (May 2011).

[38] D. Premack and G. Woodruff, 'Does the chimpanzee have a theory of mind?', *The Behavioral and Brain Sciences*, **4**, 515–526, (1978).

[39] S. Rameix, *Fondements philosophiques de l'thique mdicale*, Ellipses, Paris, 1998.

[40] J. Rawls, *A Theory of Justice*, Belknap Harvard University Press, Harvard, 1971.

[41] J.-J. Rousseau, *Du contrat social*, 1762.

[42] J.-P. Sartre, *L'existentialisme est un humanisme*, Gallimard, Paris, 1946.

[43] D. Schreckenghost, D. Ryan, C. Thronesbery, P. Bonasso, and D. Poirot, 'Intelligent control of life support systems for space habitat', in *Proceedings of the AAAI-IAAI Conference*, Madison, WI, USA, (1998).

[44] N. Sharkey, 'Death strikes from the sky: the calculus of proportionality', *Technology and Society Magazine, IEEE*, **28**(1), 16–19, (2009).

[45] B. Skyrms, *Evolution of the Social Contract*, Cambridge University Press, Cambridge, UK, 1996.

[46] R. Sparrow, 'The Turing triage test', *Ethics and Information Technology*, **6**(4), 203–213, (2004).

[47] R. Sparrow, 'Killer robots', *Journal of Applied Philosophy*, **24**(1), 62–77, (2007).

[48] R. Sparrow, *Can Machine Be People?*, 301–315, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[49] J. Sullins, 'When is a robot a moral agent?', *International Journal of information Ethics*, **6**(12), (2006).

[50] C. Tessier and F. Dehais, 'Authority management and conflict solving in human-machine systems', *AerospaceLab, The Onera Journal*, **Vol.4**, (2012).

[51] G. Veruggio, 'Roboethics roadmap', in *EURON Roboethics Atelier*, Genoa, (2011).

[52] L. Vikaros and D. Degand, *Moral Development through Social Narratives and Game Design*, 197–216, Ethics and Game Design: Teaching Values through Play, IGI Global, Hershey, 2010.

[53] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Rights from Wrong*, Oxford University Press, New York, 2009.

[54] K. Warwick, *Robots with Biological Brains*, 317–332, Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012.

[55] K. Warwick, D. Xydas, S. Nasuto, V. Becerra, M. Hammond, J. Downes, S. Marshall, and B. Whalley, 'Controlling a mobile robot with a biological brain', *Defence Science Journal*, **60**(1), 5–14, (2010).

[56] D.D. Woods, E.M. Roth, and K.B. Bennett, 'Explorations in joint human-machine cognitive systems', in *Cognition, computing, and cooperation*, eds., S.P Robertson, W. Zachary, and J.B. Black, 123–158, Ablex Publishing Corp. Norwood, NJ, USA, (1990).

Integrating civil unmanned aircraft operating autonomously in non-segregated airspace: towards a dronoethics?

Thomas Dubot[1]

Abstract. In the context of integrating Unmanned Aircraft Systems (UAS) in non-segregated airspace, autonomous operations raise legal and ethical questions. What is the expected behaviour of a civil unmanned aircraft operating autonomously in an airspace shared with other airspace users? And how could we implement this behaviour? We present in this paper a preliminary study that allowed us, through the analysis of aviation reference documents, to identify some ethical criteria necessary to develop a first set of logical rules formalizing this expected behaviour.

1 TERMINOLOGY AND SCOPE

UAS and UAOA

The term UAS designates the global system of an aircraft (UA) and its associated elements operated with no pilot on board. Regulators currently distinguish two types of Unmanned Aircraft (UA): the Remotely-piloted aircraft (RPA) which are remotely and fully controlled from another place by a licensed remote pilot, and autonomous unmanned aircraft, that do not allow pilot intervention in the management of the flight. As the purpose of our study is not to clarify the terminology linked to autonomous aircraft or operations, we will use in this paper the unofficial acronym UAOA (Unmanned Aircraft Operating Autonomously) to designate an UA that must at time t manage its flight and make decisions without any human intervention. This definition does not exclude communication links with pilot or any other authorized personnel such as Air Traffic Service (ATS), and potential orders or requests sent by these actors.

Dronoethics

In reference to the term roboethics, the name dronoethics is introduced to refer to an Applied Ethics dedicated to UAS.

Civil vs military

Our study is focussed on civil autonomous operations and does not encompass specific military ethical issues, such as the acceptable loss of human life.

2 INTRODUCTION

In the last decades, the use of Unmanned Aircraft Systems (UAS) has significantly increased in the military domain but despite the large variety of civil applications identified, the civil market has not yet developed significantly, due to the inability for UAS to access to non-segregated airspace. The need to operate military, commercial, and privately-owned unmanned aircraft in the same airspace as manned aircraft, especially outside segregated areas is now considered by all regulators as a high priority [1].

Nowadays, UAS are generally operated in segregated areas or with operations limited to specific airspace (e.g. temporary restricted, low-density/unpopulated areas) and specific procedures (e.g. low-range, visual observers on ground) [2]. If these alternatives allow managing current operations on a case-by-case basis, they are not sufficient to deal with the forecast growth of UAS operations and the whole ATM/UAS community is now developing simultaneously the operational, procedural and technological framework required for the UAS integration in non-segregated airspace [25].

According to ICAO, only Remotely-piloted aircraft (RPA) will be able to integrate into the international civil aviation system in the foreseeable future [3]. Nevertheless our study is focussed on Unmanned Aircraft Operating Autonomously (UAOA) operations in non-segregated airspace that may represent the biggest challenge of the UAS integration in the future.

If we consider the new Air Traffic Management (ATM) Concepts of operations (CONOPS) defined within current international programmes such as SESAR [4], the first idea is that UAS, as new airspace users, should mirror the procedures applicable to manned aircraft, without any special requirement for the Air Traffic Controllers (ATC), and without increasing the risk for other airspace users. Thus if we intent to integrate UAS into non-segregated airspace, within this ATM framework, they should behave like manned aircraft, whatever their mode of operations (human-in-the-loop or acting autonomously): an UAOA is then supposed to reproduce manned aircraft behaviour i.e. to make the same choices as a pilot onboard would make.

If many technical and operational studies have dealt with problematic like the Detect and Avoid concept to replace the See and Avoid procedure, the legal framework linked to the responsibility of an UAOA in case of accident is insufficient [5] and ethical issues have not been enough addressed [6]. In parallel, the importance of robot ethics (or roboethics) has been raised recently by working groups such as [7]. Following roboethics recommendations e.g. from the ethical committee of the French Scientific Research Centre CNRS [8], could we also consider endowing UAOA with moral sense or ethics that could allow them to act ethically when they must make decisions?

[1] ONERA, email: thomas.dubot@onera.fr

To the heart of these considerations, our study aims at exploring three questions:

- What could be the ethical behaviour expected from an UAS in non-segregated airspace? Which criteria express this behaviour? Is this behaviour a mirror of the manned aviation behaviour?
- Could we formalize this behaviour as a set of logical rules?
- How do we imagine applying these rules to UAOA?

As a first answer, this paper presents a preliminary analysis leading to the elaboration of a first set of ethical principles that could serve as a basis for the definition of UAOA logical rules.

3 TOWARDS AN ETHICAL BEHAVIOUR: IDENTIFICATION OF CRITERIA

3.1 Rules of the Air

Whatever the region of the world overflown, pilots are supposed to know and apply Rules of the Air that provide rules to properly fly and manoeuvre aircraft. Defined at regional [9], sub-regional [10] or national level [11], they guarantee the rational behaviour of each aircraft. Within all these documents, we have identified five major topics that could be applicable to unmanned aircraft.

Safety - An aircraft must not endanger persons and property
During all the flight phases, the aircraft should not have behaviour potentially dangerous to persons or property. For instance, if the aircraft flies over a congested area, it should be at such height as will permit, in case of emergency, to safely land without hurting people on the ground. The main rule is that aircraft shall not be operated in such proximity to other aircraft as to create a collision hazard. Nevertheless according to the Rules of the Air, a pilot may depart from these rules in the interest of safety.

If we consider an UAOA, these simple rules are already challenging: a prerequisite is that the aircraft must know its position and be able to detect and analyze its environment before modifying its path.

Priority and status - An aircraft must interact with other Airspace Users (AU) according to priority rules
When two aircraft are converging, each of them must act according to right-of-way rules: one must yield the way and the other that has the right-of-way must maintain its heading and speed. Rules have been refined according to several scenarios e.g. approaching head-on, overtaking or converging but these rules have exceptions linked to the type of aircraft. Typically aircraft with less manoeuvrability has the right-of-way but this rule is superseded when an aircraft is in distress and therefore has the priority to all other traffic.

From an UAOA point of view, several conditions seem to be necessary. Firstly the aircraft must have self-awareness about its type of aircraft and its current status (Unmanned aircraft with no passengers onboard? Flight leader of a squadron of aircraft flying in formation? In a final approach? In an emergency mode?). Then knowing its type and status, the aircraft must be able to communicate this information to all other airspace users via signals or anti-collision and navigation lights. It must also identify the status of the surrounding traffic. For instance even if it is supposed to have the right-of-way, it must detect whether the convergent aircraft is landing or is in distress and in that case yield the way.

Communication - An aircraft must continuously communicate with Air Traffic Services (ATS)
Each aircraft should comply with any instruction given by the appropriate ATS unit. Even if its flight is in line with the flight plan and the ATC orders, it should report its position when passing reporting points or periodically. And as soon as there is a deviation from the requirements, it should be communicated to air traffic services unit. To ensure this permanent interaction, the aircraft should always maintain a continuous air-ground communication, if possible with a dual channel (radio and data link). In case of failure of this communication, the aircraft must attempt to restore a communication with the appropriate ATC unit using all other available means.

In case of UAS, this could imply to maintain or try to establish the communication, to answer to potential ATS requests and to take into account these clearances in its decision-making process.

Predictability - An aircraft must have a predictable flight
Before departure, for each aircraft flying in controlled airspace, a flight plan should have been submitted to air traffic services containing as information as possible, including the forecast route but also alternative procedures. If any potential modification can be anticipated, it must be indicated in the flight plan. During the flight, the aircraft is supposed to adhere as much as possible to the flight plan but if it fails to stick to this plan, its behaviour should still be predictable. For instance the aircraft could maintain its heading and speed when it encounters some problems and then rejoin its current flight plan route no later than the next significant point. In the same way, it could land at the nearest suitable aerodrome, easily identifiable by air traffic services.

This requirement of predictability is one of the most challenging when considering an UAOA that could make decisions based on different choices, including ATC instructions. This implies specifically that alternatives should be identified and emergent behaviours anticipated.

Emergency - An aircraft must handle emergency procedures
A predictable behaviour includes non-nominal use cases when the aircraft operates in an emergency mode. In case of a loss of communication, it could for instance maintain its speed and heading during a few minutes and try to reconnect to its ground station, before entering a new emergency phase with the choice of continuing its flight or landing at a close aerodrome. Aircraft should also be able to comply with interception rules that specify the procedures to manage the instructions given by the intercepting aircraft. Therefore an UAOA should firstly know when it is operating in emergency mode, then have a catalogue of contingency plans, communicate all its choices and finally if intercepted act accordingly with interception rules, superior to any previous order.

3.2 Limitations of the Rules of the Air

Rules of the Air allow identifying high level requirements defining the rational behaviour expected from an aircraft in a shared airspace. Nevertheless a major question in the current development of UAS regulation is whether it can be based on these regulations

or whether UAS have substantially different characteristics that require new regulation. According to [12], only 30% of current manned aviation regulation applies as it to UAS, with 54% that may apply or require revision and 16% that does not apply.
Some initiatives [2][13] recommend consequently considering alternative approaches with a new way of thinking. Following UAS specificities could lead for instance to new operational procedures and modifications to existing regulations:

- Priority: in some cases, small unmanned aircraft could yield the right-of-way to manned aircraft
- "Sacrificability": in order to minimize risk to persons and property, an UAS crash could be considered in a controlled manner
- Severity of loss: although for manned aviation loss of an aircraft would mean a high probability of multiple fatalities, in the case of UAS this is not necessarily true
- Security of communications: with a pilot on ground, the importance of communications link and availability of bandwidth is now fundamental

3.3 Key ATM expectations

If Rules of the Air are a set of rules guaranteeing a safe manned aviation, they do not explain the fundamental values that support these rules. And with a new airspace user that could imply the need for a revision of these rules, the whole coherence of the system may not be ensured. Like many industry business, the ATM world has defined its own performance indicators to assess the performance of the current system and to guide the development of future ATM systems. ICAO has thus defined eleven Key Performance Areas (KPAs) [14] [15] to categorize performance subjects related to high-level ATM ambitions and expectations. The figure hereafter presents these expectations that have been clustered during the SESAR definition phase [4] into three major groups, according to the degree of visibility of the KPA outcome and impact.

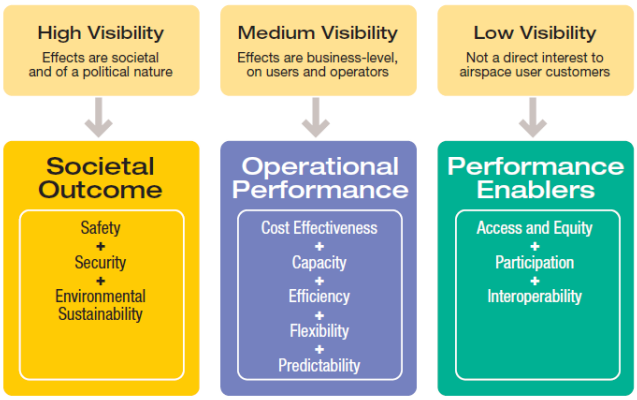

Figure 1. ATM performance targets applied to the European ATM system

As stated by ICAO [14], the ATM system should involve the participation of the entire aviation community: UAS, as new airspace users should therefore operate with a behaviour compatible with these ATM values, which means behaviour based on these values or that respects other airspace users in accordance with these values.

If we consider these ATM criteria from the UAS perspective, i.e. a new airspace user point of view, we can split these criteria in 3 groups according to the rules that can be inferred:

- ATM services: as any airspace user, UAS should have right to operate in a way compatible with [access and equity, participation to the ATM community, interoperability]
- ATM rules: as any airspace user, UAS operations should take into account [safety, security, environment, efficiency, flexibility and predictability]
- ATM global common good: UAS should not be operated in a way that could decrease the global performance of the ATM system according to [ATM rules], cost-effectiveness (cost of ATM services, e.g. the number of the Air Traffic Controller to face a raising workload, or the integration of new tools and systems to be developed and maintained) and capacity (decrease of the global capacity linked to UAS operations e.g. the insertion in a high density approach or the activation of a reserved airspace).

3.4 Limitations of ATM expectations

ICAO expectations are not fixed moral rules: they have been defined to answer to the 2025 expected scenario (without UAS specificities taken into account) and may be moving in the future [16]. Besides, like in many other domains it has always been difficult to quantify ethics in ATM and to transcribe an ethical behaviour into indicators.

3.5 UAS behaviour versus manned aviation behaviour

We noted in the introduction that one of the main concepts proposed for the integration of UAS in ATM environment was that UAS, as new airspace users, should mirror the procedures applicable to manned aircraft. After this analysis of current regulations and ATM expectations, we decided to transcend this first statement and envisage an UAS behaviour different from the manned aviation behaviour and in the same time acceptable for the manned aviation community.

Considering some criteria previously defined, we could imagine some UAS able to integrate as a parameter the global interest of the ATM community. Advanced algorithms could simulate and analyze the global impact of a modification of the UAS flight on the overall traffic based on criteria such as the capacity or the efficiency. Besides, data of interest (weather data, surrounding non-cooperative traffic detected by a Detect and Avoid system) could be shared with the ATM community according to the current needs, e.g. a volcanic ash particles analysis after a volcanic eruption. Finally we could imagine for some type of UAS mission a "Good Samaritan Law" that would bind an UAOA to assist other airspace users (or more generally humans) in need like basic international laws that require ships to assist other naval vessels in distress.

Such behaviour could also be beneficial to the manned aviation community that could adapt its own behaviour according to these

new principles: a part of the role of the Network manager, in charge of the common good of the ATM (notably via the traffic flow and capacity management processes) could be delegated to airspace users, currently focused on personal mission/business needs.

4 TOWARDS A FIRST SET OF RULES FOR UAOA

As we considered roboethics studies and roadmaps as a reference for our study, we firstly explored sets of rules defined for autonomous robots to analyse their form (granularity of rules, logical assertions) but also their content (ethical requirements for autonomous agents, conflicts among laws).

4.1 Back to sci-fi robot rules

The most famous robot rules have been defined in 1942 by the science fiction author Isaac Asimov. In his novel [17], he introduced the following three laws of robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

He also added a fourth law in a following novel to precede the others:

4. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Within our UAOA problematic, the first law could refer to the first principle (safety) identified in the Rules of the Air: UAS should not be operated in such proximity to other aircraft as to create a collision hazard that could lead to human injury. Besides the injury through inaction could evoke the Good Samaritan law described at the end of first part. The second law could be interpreted as a rule specifying that an UAOA must always obey the orders of authorized personnel such as operators, ATS, and possibly in the future the Network Manager. The third law could be adapted to UAS operations that should avoid any danger threatening the existence of the aircraft (safety of goods). Nevertheless it should be linked with the principle of "sacrificability" described in first part. Finally, in the last law, humanity could recall the global common good described previously in the ATM expectations paragraph.

In our UAS context, it appears that sci-fi robot rules could help defining the expected behaviour of an UAOA integrated in air traffic. Some examples are listed hereafter:

- A robot must establish its identity as a robot in all cases (communication) [18]
- A robot must know it is a robot (identity) [19]
- A robot will obey the orders of authorized personnel (communication/orders) [20]
- Robots must refrain from damaging human homes or tools, including other robots (safety) [21]

4.2 Working groups and national initiatives

In April 2007, the government of Japan published recommendations to "secure the safe performance of next-generation robots". The same month, the European Robotics Research Network (EURON) updated its "Roboethics Roadmap" [7]. But the most relevant initiative comes from South Korea that provided a "Robot Ethics Charter" that describes the rights and responsibilities for Robots on the basis of Asimov's laws but also with rights and responsibilities of manufacturers and users/owners.

According to [22], E.U will also establish a Roboethics Interest Group (RSI). Some standards should be particularly taken into account in the implementation of all robot types:

- Safety**:** Design of all robots must include provisions for control of the robot's autonomy. Operators should be able to limit robots autonomy in scenarios in which the robots behaviour cannot be guaranteed
- Security**:** Design of all robots must include as a minimum standard the hardware and software keys to avoid illegal use of the robot.
- Traceability**:** Design of all robots must include provisions for the complete traceability of the robots' actions, as in an aircraft's 'black-box' system.
- "Identifiability"**:** All robots must be designed with protected serial and identification numbers.
- Privacy**:** Design of all robots potentially dealing with sensitive personal information must be equipped with hardware and software systems to encrypt and securely store this private data.

4.3 First set of rules

Starting from criteria identified via manned aviation reference documents or roboethics studies, we developed a first set of rules and rights that should be applicable to UAOA during the execution phase of its flight:

1) An UAOA must not operate in such a way it could injure a human being or let a human being injured without activating controls or functions identified as means to avoid or attenuate this type of incident.

2) An UAOA should always maintain a continuous communication with predefined interfaces to obey orders of authorized personnel (UAS operator, ATS, Network Manager…) except if such actions conflict with first law.

3) An UAOA must operate in such a way it could protect its own existence and any other human property, on ground or in the air, including other UAS, except if such operations conflict with first or second law.

4) An UAOA must always have a predictable behaviour, based on its route but also alternative pre-programmed scenarios, except if all forecast options conflict with first, second or third law.

5) An UAOA interacts with surrounding traffic (separation, communication) according to requirements of the operating airspace, general priority rules and emergency and interception procedures except if such actions conflict the first, the second or the third law.

6) An UAOA must always know its UAS identity and status and indicate it honestly when requested or when deemed necessary.

7) As any airspace user, an UAOA should not operate in a way that could decrease significantly the global performance of ATM system in terms of safety, security, environment, cost-effectiveness, capacity and quality of service (efficiency, flexibility and predictability), except if such operation is required by first, second or third law.

8) An UAOA must ensure a complete traceability of all its actions.

Other rules should be added but they seem difficult to implement at the UAOA level. They should then be ensured by the UAS community (participation to the ATM community, interoperability) and UAS designers/operators (security, privacy or interoperability). Some recent initiatives such as the UAS Operations Industry "Code of Conduct" [23] aim at providing such guidelines and recommendations for future UAS operations.

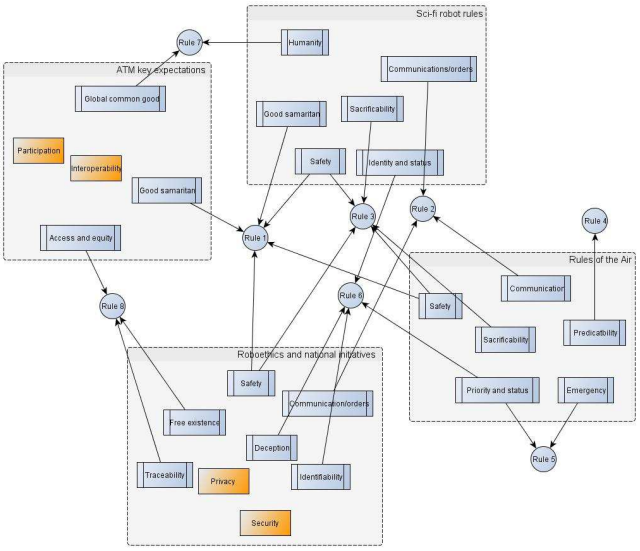

Figure 2. Correlation between UAOA rules and criteria

It should be noted that if these rules seem in line with current ATM regulations and principles, the exceptions and priorities may introduce important changes. For instance, if the fourth law states the need for a predictable behaviour, its exceptions allow unpredictable actions and therefore emergent behaviour in circumstances linked to the three first laws. Besides, the transformation of these ethical principles into logical rules will necessarily rely on the essential UAOA specificity, i.e. the absence of a pilot able to make decisions taking into account its own ethical values.

In the same way, some UAOA rights could be ensured by the establishment of general procedures. Last rule could help to verify the application of such rights, like the real access in equity of UAS to ATM resources without a priority mechanism leading to a systematic abuse limiting its efficiency and cost-effectiveness.

4.4 Conflicts and priorities among laws

Within the rules previously enounced, inherent criteria e.g. capacity or safety are interdependent, which implies improving the performance in one area can come at the price of reduced performance in another area. Some conflicts are unavoidable because ethics is by nature contradictory: they have been analyzed in [15] that presents some trade-offs between ATM criteria such as the access and equity versus the capacity. In the same way, the establishment of this first set of rules and rights applicable to UAOA allows us to identify potential conflicts:

- Human order versus safety: some orders given by the operator could contradict information coming from sensors onboard indicating a potential collision.
- Priority rules versus protection of existence: if the UAOA has the right-of-way, it should maintain its heading and speed. Nevertheless if another aircraft refuses to yield the way, the UAOA could adapt these parameters to protect its existence. In case of systematic violation of priority, such procedures should be considered to respect the right of UAOA to access and equity.
- "Sacrificability" versus safety: in some exceptional circumstances, some low-cost UAOA could be asked to voluntarily crash in order to avoid a potential danger.

According to the variety of aircraft and mission concerned, it seems therefore difficult to introduce clear priorities between logical UAOA rules previously described. However, safety is always the highest priority in aviation and is not subject to trade-offs. Therefore all the laws and even a combination of laws are applicable except if they conflict with first law. We can for instance imagine an UAOA threatened by an aircraft converging very quickly that chooses to violate the right-of-way of another UAS converging (law 5), even if the risk of collision with this UAS threatens its own existence (law 3) because of the risk of endangering human life aboard the first aircraft (law 1). In that kind of worst-case scenario, with a combination of laws conflicting together, we can foresee the danger of the behaviour of other airspace users that could be tempted to divert these rules to ensure personal benefices. Such behaviour should be analyzed in the post-flight phase ensured by the traceability ensured by the eighth law.

4.5 Limitations of UAOA rules

This first list of eight UAOA rules is an example that must be considered as the initial starting point of our study. Some iteration would be needed to review some terms and express clear responsibilities. For instance in the first law, it must be clarified who will "identify" the controls and functions that could allow an UAS to intervene after an accident. In the same way law 6 should specify exactly how an UAOA could answer "honestly" to requests. Then all the laws should be confronted to identify conflicts between laws.

Depending on the result of this analysis, another set of rules could be proposed, with fewer rules and less complexity between conflicting laws, such as the following set:

- Law 1: An UAOA should always maintain a continuous communication with predefined interfaces to obey orders of authorized personnel (UAS operator, ATS, Network Manager…).
- Law 2: An UAOA must not operate in such a way it could endanger persons and property except if such operation conflicts with first law.
- Law 3: An UAOA must always have a predictable behaviour, based on its route but also alternative pre-programmed scenarios, except if all forecast options conflict with first or second law.

This simplified set of rules could also ease the societal acceptability of autonomous operations. It could be then considered as a first step towards the application of the final set. That's why we inverted two first laws, considering that in a near future autonomous operations could be better accepted if it is acted that any human order can overcome any other decision.

5 CONCLUSION AND PERSPECTIVES

In this first phase of our study, we defined a first set of rules and rights via the analysis of criteria identified in ATM reference documents and in roboethics studies. As many other documents could be also relevant, we could reiterate this process in order to identify new criteria and refine this set.

Nevertheless we wish to explore alternative means to consolidate this first set of laws for instance via the definition of scenarios of UAOA integration such as UAS scenarios defined in [24] [25]. We will notably describe procedures for special cases such as loss of communication or critical system failures and apply them considering an UAOA complying with ethical rules. In parallel, we will analyse the potential correlation between various levels of automation in ATM and the integration of UAOA.

These analyses should allow us to identify rules to be added, removed or corrected and potential conflicts between combinations of laws, but also whether several sets need to be defined, according to the type of UAS, its type of mission and its degree of autonomy.

After this consolidation, we intend to formalize this ethical set a rules using non-monotonic logics [27], probably with the Answer Set Programming (ASP) formalism. This formalization will finalize the "logical" consolidation of our set and probably raise the question of how these rules could be applied to the development and execution of an UAOA: in the process of validation of control algorithms? Or directly injected as software overlay within an AI onboard able to integrate ethical criteria in its decision-making process?

Finally, in the same way as the development of intelligent robots raise the question of our fundamental ethical values, this study on UAOA could allow to consider new approaches for the "manned" aviation, with the introduction of new concepts of operation, the refinement of current rules and the application of UAS algorithms or systems to all airspace users.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Jean Hermetz, Assistant Director of the System Design and Performances Evaluation Department and my colleagues Luis Basora and Dr. Charles Lesire for their valuable and constructive suggestions during the reviewing of this paper. I hope this final version will contribute to convincing them to start many studies on this topic.

REFERENCES

[1] FAA Modernization and reform Act of 2012, Report 112–381, (2012).

[2] A. Lacher, A.Zeitlin, D. Maroney, K. Markin, D. Ludwig, and J. Boyd, Airspace Integration Alternatives for Unmanned Aircraft, The MITRE Corporation, (2010).

[3] ICAO Cir 328, Unmanned Aircraft System (UAS), (2011).

[4] SESAR, The ATM Target Concept, D3, (2007).

[5] Autonomous Machines Prompt Debate, The Engineer Online (GB), (2009).

[6] The Royal Academy of Engineering, Autonomous Systems: Social, Legal and Ethical Issues, (2009).

[7] G. Veruggio, EURON Roboethics Roadmap, (2007).

[8] J. Mariani, J-M Besnier, J. Bordé, J-M Cornu, M. Farge, J-G Ganascia, J-P Haton, E. Serverin, Comité d'Ethique du CNRS (COMETS), Pour une éthique de la recherche en Sciences et Technologies de l'Information et de la Communication (STIC), (2009).

[9] ICAO Rules of the Air, Annex 2 to the Convention on International Civil Aviation, (2007).

[10] Single European Sky (SES) Regulations, Final report for the draft implementing rule on the development of standardised European Rules of the Air, (2010).

[11] RDA, Annexe 1 à l'arrêté du 3 mars 2006 modifié (Règles de l'air), (2008).

[12] FAA CoE for General Aviation Research (CGAR) Annual Meeting, (2007)

[13] K. Dalamagkidis, K. P. Valavanis, L. A. Piegl, On Integrating Unmanned Aircraft Systems into the National Airspace System, (2009)

[14] ICAO Global Air Traffic Management Operational Concept, (2005)

[15] ICAO Manual on Global Performance of the Air Navigation System, (2009)

[16] ICAO Global Air Navigation Plan, (2007)

[17] I. Asimov, Runaround, (1942)

[18] L. Dilov, Icarus's Way, (1974)

[19] N. Kesarovski, The Fifth Law of Robotics, (1983)

[20] D. Langford, http://en.wikipedia.org/wiki/Three_Laws_of_Robotics

[21] Japan's "Ten Principles of Robot Law" adapted from "Ten Principles of Robot Law" formulated by Osamu Tezuka for his Astro Boy series.

[22] http://akikok012um1.wordpress.com/european-union%E2%80%99s-convention-on-roboethics-2025/

[23] AUVSI UAS Operations Industry "Code of Conduct" (2012)

[24] FP6 INOUI D1.2 Concept for civil UAS Applications, (2008)

[25] ICONUS study, http://www.sesarju.eu/news-press/news/sesar-launches-study-unmanned-aircraft-1070

[26] J-G Ganascia, Ethical System Formalization using Non-Monotonic Logics, (2007)

The primacy of human autonomy: understanding agent rights through the human rights framework

Bart Kamphorst[1]

Abstract. This paper is concerned with the 'rights' of autonomous agent systems in relation to human users or operators and specifically addresses the question of when and to what extent an agent system may take over control from someone. I start by examining an important ethical code of conduct for system designers and engineers and argue that one would do well to understand it within the human rights framework. I then show that framing the discussion on what agent systems may and may not do in terms of human rights has consequences for intelligent agent systems in that they should be respectful of people's dignity and autonomy. In the remainder of the paper I work out the implications of this for the conditions under which agent systems may take over control. I offer an analysis of control, of delegated control, and of autonomy-respectful delegated control, concluding that for an agent system to justifiably take over control from a user, it should at a minimum offer the user a reliable way to take back control in a timely manner. However, when the user's autonomy is at stake, the system should also know about and act in accordance with the user's goals and core values.

1 Introduction

When and to what extent should an autonomous agent system be able to take over control from a human user? This question is becoming more and more relevant because people are increasingly employing intimate, agent-based support systems that have a profound influence on their personal lives (e.g., in regulating chronic illness [22] or in overcoming obesity [4]). But before one can start formulating an answer to such a question, one needs to have a clear conception of control and of the ethical considerations that come into play when a person hands over control to an agent system. For instance, how does handing over control affect the user in terms of one's well-being or autonomy? Throughout this paper I will propose four guidelines aimed to pave the way towards an answer that takes these issues into account. I will start by examining an important ethical code of conduct for system designers and engineers and argue that in order to properly understand it, one would do well to place it within the human rights framework as codified in the Universal Declaration of Human Rights (UDHR).[2] I will then show that by doing so, it follows that the notion of human autonomy is a value worth protecting when designing intimate systems. Finally, I will set forth some tentative thoughts about the relation between autonomy and control, in order to sketch the outline of an answer to the question, namely that under normal circumstances an autonomous agent system may take over control if and only if the control is willingly delegated to the system and the user retains the possibility to take back control when he or she sees fit.

2 Personal Dignity

In 1992 the council of the Association for Computing Machinery (ACM) adopted a code of ethics that prescribes the ethical professional conduct that is expected of every member of the ACM, an organization with over 100,000 members from over 100 countries, representing the worlds largest educational and scientific computing society. Consider article 3.5 of this code, entitled "As a ACM member I will 3.5: Articulate and support policies that protect the dignity of users and others affected by a computing system" [5]:

> Designing or implementing systems that deliberately or inadvertently demean individuals or groups is ethically unacceptable. Computer professionals who are in decision making positions should verify that systems are designed and implemented to protect personal privacy and enhance *personal dignity*. (emphasis added)

While the gist of this article is clear, one cannot truly understand what it entails unless one has a working idea of what personal dignity is. That is, what is it that needs protecting, or even enhancing? The problem is that because human dignity is an extremely broad concept that people interpret differently in different contexts, there is a danger that in practice the meaning of it is void.[3] To remedy this, given the societal importance of this code of ethics, I suggest understanding personal dignity here in the context of the human rights framework.[4] Within this framework, codified in the UDHR, personal dignity can be understood as a fundamental moral property of people that they are normative agents worthy of respect, a notion that lies at the very core of the framework. Article 1 for instance states that "All human beings are born free and equal in dignity and rights", and dignity also plays a role in positive rights such as social security (art 22) and the right to employment (art 23). In other words, it is through this fundamental moral property of dignity that people have certain rights in the first place, namely those rights that protect personal dignity. In the discussion about what autonomous agent systems may and may not do, then, protection of the user's personal dignity seems like a sensible place to start.

[1] Utrecht University, The Netherlands, email: bart.kamphorst@phil.uu.nl

[2] Because I understand the notion of agent rights in the context of the human rights framework, and I am not inclined to grant autonomous agent systems a human-like status within this framework, I will refrain from speaking of rights and duties of autonomous agents from here on out.

[3] For a critical discussion of the use of the term in medical ethics, see [17]. Interestingly enough in light of this paper, according to Macklin dignity means 'nothing more' than having to respect people's autonomy!

[4] There are also other hints such as "As a ACM member I will 1.1: Contribute to society and human well-being." that indicate that the placement of the code in this context is valid.

Guideline One: Under no circumstances may an autonomous agent ever be harmful to anyone's personal dignity.

While this may seem like a trivial guideline, and an obvious one at that, it may serve as a stepping stone for other guidelines with the human rights conception of personal dignity as their foundation. In the following section I will argue that the next step up is that autonomous agent systems should respect human autonomy.

3 Personal Autonomy

The UDHR does not attribute autonomy to every human being like it does with dignity. Instead, it assumes that people have the capacity for self-rule and strives to lay the groundworks for an environment in which people can develop their autonomy, i.e. become autonomous beings. Personal autonomy, understood here as having the freedom, the capacity and the authority to choose one's own course of action to direct one's life in accordance with one's goals and values, takes a prominent place in (philosophical) discussions about agency. Constituted by a measure of independence and a (minimal) requirement of rationality (see [2] for a thorough discussion), being personally autonomous has normative implications in that one is held responsible for one's choices and actions, but also that others have obligations to respect one's right to decide on and follow a certain course of action. This in turn has a significant impact on the political and legal domain, in that it determines how "the state is permitted to restrict or influence individuals' choices of how to lead their lives" [2]. Personal autonomy is closely related to personal dignity, in that dignity is a necessary condition for leading an autonomous life. Conversely, however, this need not be the case. People can be more or less autonomous than each other or even than themselves viewed over a period of time, without any threat to their dignity. What this observation shows is that autonomy, contrary to dignity, may be viewed as a scale [2, sec. 3.1]. But note that the fact that autonomy can be viewed as a scale does not mean it is negotiable. As Anderson rightly notes, the right to autonomy should be "understood not in terms of ideals of development but rather as a fundamental boundary not to be violated" [2, p. 12, sec 3.2]. Similarly, Oshana remarks that "[t]he fact that it might be morally and legally incumbent upon us to caution others against their own behavior, to warn them of the punitive consequences that might follow their behavior, and to actually take steps to curtail their autonomy, does not mean that autonomy is not a valued state. This ideal remains intact, although uninstantiated in certain cases." [21, p. 126]. The value of autonomy is anchored within the UDHR and individuals as well as society should strive to maximize people's autonomy. Moreover, and very relevant for the discussion at hand, perceived autonomy can be measured, and there is empirical research that shows that diminished autonomy negatively affects personal well-being [e.g. 23]. Therefore, I contend that autonomy may be restricted only insofar as the exercise of autonomy frustrates anyone else's autonomy or the state has compelling reasons to do so.

Guideline Two: Autonomous agent systems should be respectful of people's autonomy. They may not diminish a user's autonomy, unless otherwise directed by law.

At this point I would like to touch upon a possible criticism, voiced but dismissed by Verbeek in a discussion about persuasive technology:

> [A]utonomy was thought to be attacked when human actions are explicitly and consciously steered with the help of technology. This reduction of autonomy was even perceived as a threat to human dignity; if human actions are not a result from deliberate decisions but from steering technologies, people were thought to be deprived from what makes them human. [24]

I agree with Verbeek here that this is not the way to think about the relation between technology and autonomy. What I will come to argue later in this paper is that being steered (better: guided) by an intelligent, autonomous agent system, does not impede one's autonomy per se, but that it might become an issue when one cannot change one's course of action when one has good reasons to do so.

Thus far I have argued that in order to say something meaningful about what intelligent agent systems may and may not do, one would do well to frame the discussion in terms of the human rights framework. I have argued that human autonomy is a central notion within this framework, that it is an important value to protect in all circumstances, and therefore should be regarded as an important value when designing autonomous agent systems that (closely) interact with people.

4 Value Sensitive Design

The idea that software should be respectful of people's autonomy is not new. Most prominently, Friedman has argued at length for the inclusion of human values in the design of computer systems and software agents, specifically respecting and enhancing what she calls 'user autonomy' [11, 10, 12]. In their treatment of human agency and responsible computing, Friedman and Kahn begin by asking the question whether autonomous agent systems can be moral agents like human beings, something they then argue cannot be the case because to date computer systems do not have intentionality, and intentionality is taken to be a prerequisite for morality [11]. But because in some cases people's understanding of this boundary between humans and computational systems is distorted, the argument continues, people's sense of moral agency can be diminished, which in turn causes erosion (their terminology) of their dignity. Friedman and Kahn then propose design strategies to preclude this distortion from happening such as nonanthropomorphic interface design (sharpen the distinction between humans and computers) and participatory design (involving users in defining the problems the system should tackle). In later work, Friedman proposes user autonomy as an important value to take into account when practicing *Value Sensitive Design* (VSD) for developing user-centered systems "because it is fundamental to human flourishing and self-development" [10], citing work by Gewirth (1978) and Hill (1991). Here, Friedman distinguishes between System Capability, System Complexity, Misrepresentation of the System and System Fluidity as aspects of systems that can influence user autonomy [10].[5] While Friedman and I share autonomy as an important value, we seem to differ on the circumstances under which autonomy is in danger. For instance, Friedman and Nissenbaum write that "user autonomy can be undermined when there are states the user desires to reach but no path exists through the use of the software agent to reach those states" [12]. As an illustration they consider a mail agent that has the capability to filter emails by subject header, but does not understand a concept such as urgency. This, Friedman and Nissenbaum argue, leads to the undermining of autonomy for the user who wishes to filter emails by urgency. I think this is too strong: the fact that one's expectations about the capabilities of the agent do not match reality does not threat one's autonomy

[5] Elsewhere, Friedman distinguishes a fifth aspect, namely knowledge about the system: how the (non-)transparency of a system's internal workings may influence user autonomy. See [10, 12].

per se. The software agent was employed by the user and could also be disabled by the user. So, at worst, the user's hopes for autonomy enhancement were unrealized, but there is no loss of autonomy to speak of. It is only when control is handed over but cannot be regained, I will argue in Section 6, that in some cases one's autonomy can be in trouble.

What I do share is Friedman's insight that values such as autonomy (or freedom from bias) do not necessarily override others. So where the right to autonomy is a fundamental boundary not to be violated, the level to which one may exercise this right may under some circumstances be restricted, for instance "to protect against a user with malicious intentions or well-intentioned users guided by poor judgment" [10, p. 22], or "in situations where safety is at stake" [12, p. 6]. Given the status of autonomy however in societies that subscribe to the human rights framework, I think that such judgements should be left to the legislator or the judiciary (as is visible from Guideline 3).[6] Nevertheless, I subscribe to the idea of value-sensitive design and, albeit via a different route I too suggest that human autonomy should play a central role in the discussion of what autonomous agent systems may and may not do.

So, having framed the discussion in terms of the human rights framework and having argued that human autonomy should be considered a central concept, it would now be possible to give a tentative answer to the question posed in the introduction by saying that it is acceptable for a system to take over control, *as long as it doesn't impede one's autonomy*. But what this really only does is reframe the problem, because this answer does not provide any insights about the conditions under which control impacts autonomy. To work towards an actual answer, then, it is now time to spell out what control is, before turning to the relation between control and autonomy.

5 (Delegated) Control

Having control over something roughly means being causally responsible for a particular state of that something. Like autonomy, control is another concept constitutive of and interrelated with human agency. In the first place, people have self-control: "control of the self by the self" [19], which implies being causally responsible for one's own decisions and actions in accordance with one's goals. This type of control is known as executive control, and has been considered (but debated) as the basis on which one can ascribe morality and responsibility to people [3]. Secondly, people can control parts of their environment, such as a soccer ball, or a computer.[7] Importantly, however, being in control need not be limited to human beings. Take the classic example of the thermostat. Without going into the discussion about whether the thermostat can be said to have beliefs and goals about temperature, it is uncontroversial to say it controls the temperature of a room. Now what is important to observe, is that control is transitive: if one controls the thermostat, one controls the temperature of the room by means of the thermostat. This observation, I will argue, plays a crucial role in understanding when delegation of control is justified.

Delegating control, I propose, means handing over immediate causal responsibility over some object or process to another entity, with the provision that one can retake control when one sees fit. The entity may either be another human being or an autonomous system of some sort. The object of control can many things, including decision-making processes that normally would be handled by the self, but do note that handing over self-control as such is a contradictio in terminis. If the delegate system is so intimately coupled with the delegator that the delegator considers it part of the self, then there is no delegation to speak of, only self-control. If, on the other hand, control over the self is delegated to an external entity, we cannot speak of self-control, as it is not the case that the self controls the self. Nevertheless, in principle, control over a great many things can be delegated. But what are the conditions that the delegate should conform to?

In answer of this question I would like to start with a useful distinction made by Fischer and Ravizza in the discourse on the minimal requirements for moral responsibility between guidance and regulative control. Whereas others have held that moral responsibility requires full-blown regulative (cf. executive) control, Fischer and Ravizza contend that guidance control, i.e. "the agent's "ownership" of the mechanism that actually issues in the relevant behavior, and the "reasons-responsiveness" of that mechanism" [7] — meaning that the mechanism is (moderately) sensitive to reasons for acting differently — is both necessary and sufficient. While Fischer and Ravizza's distinction is not directly applicable to a structure of delegated control (as I will show), it provides some useful insights. First, the requirement of ownership over the mechanism that issues the behavior also applies to delegated control: one has to have a certain ownership over the delegate. This does not necessarily entail physical or legal ownership, but a kind of ownership that follows from "taking responsibility for them" [7]. So, for example, reading instructions and knowingly enabling an agent system may be sufficient for this. Secondly, there has to be some sort of mechanism that allows for intervention, i.e. for taking back control. But while moderate reasons-responsiveness is a reasonable requirement for guidance control, it seems to be too strong for delegated control because although some intelligent, autonomous agent systems may be reasons-responsive, others systems (e.g. thermostats) are not. For delegated control, then, I suggest a weaker condition, namely a mechanism that is responsive to *control-retraction*. In other words, the delegate should have a mechanism with which the delegator can take back control.[8] For it is the presence of such a mechanism that determines whether the transitive control relation still holds: if one cannot take back control, then it is not delegation but transference, or attribution. This mechanism has to be reliable as well as responsive in a timely manner. Note that the latter condition is especially important because in some cases it will be paramount that the mechanism will hand back control immediately. At the same time, this criterion leaves room for mechanisms that require a slightly higher threshold to be met — within the limits of reasonableness — for control-retraction, such as having to type in a twenty digit passphrase as opposed to hitting a big red button.

> *Guideline Three*: Delegation of control is valid as long as the delegator has ownership over the delegate, and the delegate offers the delegator a mechanism that is reliably responsive to control-retraction in a timely manner.

To illustrate this idea, consider the case of Alice, an ordinary woman who has set herself the goal to loose a few pounds. Alice is in control of what food she consumes, but when dieting, she finds that it takes a lot of self-control to refrain from eating sweets. Now to help

[6] Notice here the important difference between autonomy and dignity: a court could never justify indignity, that would imply a violation of a fundamental absolute right, and there cannot be a justification for such a violation.

[7] Note that while the object of control differs in both cases, what matters is that one is a dominant causal factor, not necessarily the single cause.

[8] Observe that delegated control is perfectly compatible with guidance control: the internal mechanism that controls the delegation structure should be one's own and be moderately reasons-responsive (guidance control).

herself stay on track, she decides to enable an autonomous, agent-based support system — one that is responsive to control-retraction, i.e. one that can be overridden, for instance when Alice has her friend Bob over for dinner — that draws up a grocery shopping list for her every other day, and orders food online. By doing so, the system is effectively preventing Alice from wandering through the supermarket where she would be confronted with temptation. Now surely Alice has not given up self-control over what she eats: whatever groceries are delivered, she can choose to eat them or not. What she has done, though, is delegate her control over the decision making process to the agent system for what foods to buy.

So, I have argued that control is something that can be delegated, and that the delegation is valid as long as the delegator has taken ownership over the delegate and has a way of reestablishing executive control. In the following section I will discuss how delegation of control relates to autonomy.

6 Delegated Control and Autonomy: Initial Thoughts

Being in control is strongly connected to autonomy. Recall from Section 3 that autonomy consists in part of a measure of independence, which one can only establish if one has control over one's self (decisions, actions) and one's immediate environment. Now to see how delegation of control can work, but also how it can be problematic for one's autonomy, consider the following scenario. Bob, Alice's friend, is an autonomous human being, and as such, he can decide on the temperature of his own home. He can choose to build a fire in the fireplace, or to simply delegate control over the temperature to a thermostat. Should he choose the latter, the delegation of control would be unproblematic, because if Bob finds that it is too cold on a winter's day, he can control the temperature by means of the thermostat. But now consider a thermostat with no off-switch that, once activated, determines what the temperature should be all on its own (it is in fact an autonomous agent system). To make matters even worse, the system is unpredictable, because unbeknownst to Bob, it determines the temperature by taking the word of the day from `http://thesaurus.com/wordoftheday`, taking the number of results that Google's search engine generates for that word, performing a modulo operation on that number with 15, and adding a constant of 10.[9] Since Bob has no control over the thermostat, he therefore lacks control over the temperature in the room, which in turn impacts his autonomy.

To see that delegation of control does not always involve autonomy concerns, take a case where someone hands over immediate control over a soccer ball to a robocup robot that reliably passes the ball back when one asks for it (control-retraction). This is valid delegation. But should the robot decide not to pass the ball back (it may even be reasons-responsive itself, passing the ball instead to another robot in a better position to score!), this surely does not hamper one's autonomy.

Looking back at the original question about when a system may take over control, it thus matters whether the object of control has the capacity to affect one's autonomy. This capacity, which I will call *autonomy-sensitivity*, determines whether delegation of control alone is acceptable in dealing with an autonomous agent system, or that what is required is *autonomy-respectful* delegation of control. To speak of delegated control that is autonomy-respectful, I propose that one more condition must be met, namely goal and value conformance.

As previously mentioned, human autonomy is in part constituted by independence: being the authority over making one's own life choices in accordance with one's goals and core values. What this implies, is that if a delegate is taking control over something that is autonomy-sensitive to the delegator, the delegate has to act in such a way that the delegator perceives the decisions and actions of the delegate as an extension of the self in order to prevent interference with respect to the delegator's independence. To accomplish this, the delegate should know about and act in accordance with the delegator's goals and core values.

> *Guideline Four*: Delegated control is autonomy-respectful if and only if there is valid delegation of control over something that is autonomy-sensitive, and the delegate acts in accordance with the delegator's goals.

There are a number of things to note about this final requirement. The first is that it relates to Friedman and Nissenbaum's notion of 'agent fluidity': "software agents need to take [evolution of the user's goals] into account and provide ready mechanisms for users to review and fine-tune their agents as their goals change" [12]. Indeed, people's goals do change, and to prevent a delegator from feeling alienated from the delegate's decisions and actions, for instance because it is striving to obtain an outdated goal, agent fluidity should be an important element in agent systems design. Secondly, attesting to the importance of the requirement, is that it relates to Ryan and Deci's self-determination theory, the idea that developing a sense of autonomy is critical "to the processes of *internalization and integration*, through which a person comes to self-regulate and sustain behaviours conductive to health and well being" [23]. Especially where agent systems are in a position to instruct and guide the delegator's behavior (e.g. Klein, Mogles, and Van Wissen's eMate), guiding them towards goals the users personally endorse is crucial. Finally, what this requirement highlights, is the importance of individualization, personalization, and tailoring [8]: individuals have different needs, preferences, beliefs, goals, and quite likely, different autonomy-sensitive objects of control. To see that this is so, consider Carol and Dave, who both decide to enable an intelligent agent system that will recommend clothes for them to wear on a daily basis. Carol, who has a great sense for fashion, uses the recommendations to pick and choose her wardrobe, and if she doesn't like the recommendation, she will happily wear something else. Dave on the other hand, has a very poor sense of fashion. In fact, it doesn't take long for Dave to rely on the recommendations of the agent system. But here's the catch: despite his poor sense of fashion, Dave does have certain values about dressing properly for the occasion, and for his new job as assistant professor, he wishes to look presentable, so not to undermine his credibility and authority.[10] Should the system recommend clothes that do not fit this profile, Dave's autonomy will be affected. So, this example illustrates how the object of control can be the same, but the autonomy-sensitivity can differ on an individual basis. This, too, must be accounted for by an autonomous agent system that has been delegated control of something that is autonomy-sensitive to the delegator.

7 Implications

In the previous sections I have argued that in order to say something meaningful about when an autonomous agent system may take over

[9] This would have made it a nice 21 degrees Celsius on May 5th 2012 with the verb 'besot' (approx. 431.000 hits).

[10] Example derived from [20, pp. 177–178].

control from a human user, we would do well to place the discussion within the human rights framework. The implication of this is an emphasis on people's personal autonomy. In order to better understand how having control relates to autonomy — something that is especially important for the design of intimate, agent-based support systems — I have offered an analysis of control, of delegated control, and of autonomy-respectful control delegation. This section elaborates on the implications of the conceptual work. First, very broadly, by framing the discussion in terms of human rights, we get a lot of practical rules for free, in that an agent system may exercise its autonomy (and thus the tasks delegated to it) as long as it does not frustrate anyone else's rights.[11]

Secondly, the question of *when* an agent can take over control, depends on the autonomy-sensitivity of the object of control. But at a minimum, when the object of control is not autonomy-sensitive, control may be taken over when it is willingly delegated — that is, the user should know about and agree with the delegation — and *to the extent* that the delegate has a mechanism that is responsive to control-retraction. One might ask at this point 'Is the control-retraction mechanism necessary?', because one could of course enable an agent system and simply let it run. My response to this question is twofold. Firstly, yes, such a mechanism is necessary in order to speak of delegated control, because without it, the transitive chain of control is broken. Secondly, although technically possible, control transference or control attribution is problematic in terms of responsibility, because on the one hand one cannot rely on the transitive control chain in those cases, and on the other hand it is problematic to consider the agent system a true moral agent with moral accountability [11]. So, normatively speaking, releasing the requirement of control-retraction is undesirable.

Finally, considering a case where the object of control is autonomy-sensitive, an individual's right to autonomy dictates that an agent system may only take over control when it is validly delegated, and the delegate has the capacity to act in accordance with the delegator's (changing) goals and core values. As shown, this is crucial for respecting and protecting the delegator's autonomy. The implications of this is that autonomous agent systems should use personalization and tailoring techniques, and should have access to personal information. Of course, this sparks two separate discussions, on the ethics of persuasive systems and on privacy respectively, but those are beyond the scope of this paper.

7.1 Exceptions

The guidelines laid down in this paper are not without exceptions. One type of exception in particular I would like to mention here, and those are the cases in which a person's autonomy is well below the (minimal) level of autonomy that is presupposed throughout this paper. It is highly conceivable that such cases, for instance that involve people who have very little self-regulatory capacities, should be treated differently. Here, I think, human dignity still plays a central role, but other criteria should be considered as well, such as a person's well-being or a person's prospects for autonomy enhancement. For example, if the use of an autonomous agent system without an overrule mechanism for that user would actually enhance that person's quality of life, then it seems to me such a system should be at least be considered to be allowed (given that it respects the person's dignity).[12] Considering what is at stake in such cases, taken together with common practice regarding lack of autonomy (e.g. legal guardianship), I think such decisions should be left to a specialized institution or a court of law.

[11] Note that in a societal context we may add the clause that actions must be lawful, i.e. legal within the boundaries of the law.

[12] Note that to preclude any issues of responsibility, the system should be responsive to control-retraction from a specialist care-taker or other authority figure.

8 Final considerations

Before concluding, I would like to mention two separate issues that should be addressed in the discussion of what autonomous agent systems may and may not do. The first is concerned with the difference between design and actual use, the second with a dilemma about the limits of autonomy.

8.1 Design versus Use

The main aim of this paper was to provide some preliminary ideas for thinking about the relation between humans and autonomous agent systems, in order to further the discussion of the normative judgements one can make about what such agent systems may and may not do, especially in relation to taking over control. Throughout this paper, though, I have also discussed some design principles that either follow from, or are important for the discussion at hand. I am aware that the relation between design and use is "very complex and principally unpredictable" [1], and I agree in principle that we must not overestimate the correlation between designer intention and actual use. Even so, in designing agent systems that are able to take over control, it seems that providing it with a mechanism that is responsive to requests to relinquish control is sensible and reasonable no matter what domain such an agent system will be used in.

I concede that with regard to goal and core value accordance, the difference between designer intention and actual use may prove more problematic. If actual use of a system is in a totally different domain than it was intended for, personalization and tailoring may fail. For these type of questions (e.g., 'What is the domain?', 'What information does the system need from the user?', 'What type of goal should the system strive for?'), proven methods of design should be used such as stakeholder analysis [9, 8] and empirical investigations as part of Friedman et al. tripartite methodology [13], perhaps accompanied by Verbeek's modified Constructive Technology Assessment to "anticipate possible mediating roles of the technology-in-design" [24]. We cannot always reliably predict actual use, but when a value as important as human autonomy is at stake, we should do our best to err on the safe side.

8.2 Dilemma: Ultimate autonomy?

Finally, I would like to note an interesting dilemma that this paper raises. I have argued that what should be protected is one's capacity to choose one's own course of action, or in other words, to live one's life by one's own standards and desires. Of course, the human rights framework and societal institutions put bounds on this capacity in that one can exercise one's right to autonomy only to the point where one would frustrate someone else's rights. Nevertheless, within that space, one is free: free to go hiking, free to whistle a show tune, even free to mutilate oneself. So why would one not be free to put an agent-based decision support system in place that severely and uncompromisingly restricts one's options? Doesn't the very fact that one is an autonomous being imply this freedom? In response to this dilemma I would like to draw an analogy with the autonomous being wanting to be enslaved, a case discussed in the philosophical discourse on autonomy [e.g. 18, 21]. One way of dealing with such cases

is to hold that the consensually enslaved is no longer autonomous because of the enslavement. As Oshana puts it: "Consensual slavery, regardless of the gains that it might provide and aside from any benefit to the enslaved, transforms the human subject into a possession or object of another and accordingly defiles the enslaved individuals autonomy" [21]. Analogously, one might argue that if someone willingly and knowingly enables an agent system that would place severe strain on that person's autonomy, that person's autonomy is lost. Not even necessarily by the doings of the agent system, but by placing oneself in that situation in the first place. But of course, this is an extreme case, and in practice this is unlikely to happen. People are not out to restrict their autonomy, they wish to reach a particular goal (e.g. having an agent system enforce strict dietary rules to become healthy). In the end, intelligent support systems should strive to help people reach those goals, but again, it is better to err on the safe side and make sure that these systems are respectful of people's autonomy.

9 Conclusion

In this paper I have argued that the discussion about what autonomous agent systems may and may not do should be framed within the human rights framework. I have shown that personal dignity and a human being's right to (personal) autonomy are important values in our society worthy of protection (guidelines 1 and 2), but also how there is empirical evidence that a lack of perceived autonomy negatively influences well-being. I have argued that the primacy of human autonomy should therefore be acknowledged in all discussions about what agent systems may and may not do in relation to their human users or operators. In an attempt to meaningfully answer the question when and to what extent an agent system may take over control, I have made the case that control over something that is autonomy-sensitive may be taken if and only if control is willingly delegated, the delegator assumes ownership over the delegate system, there is a mechanism in place with which to take back control reliably and in a timely manner (guideline 3), and the system acts in accordance with the delegator's goals and core values (guideline 4).

ACKNOWLEDGEMENTS

I thank Joel Anderson and Arlette van Wissen for their helpful comments on an earlier draft of this paper. I also appreciate the suggestions made by the anonymous referees of the RDA2 workshop. This research was supported by Philips and Technology Foundation STW, Nationaal Initiatief Hersenen en Cognitie NIHC under the Partnership programme Healthy Lifestyle Solutions.

REFERENCES

[1] A. Albrechtslund. Ethics and technology design. *Ethics and Information Technology*, 9:63–72, 2007.

[2] J. Anderson. Autonomy. In H. LaFollette, J. Deigh, and S. Stroud, editors, *International Encyclopedia of Ethics*. Wiley-Blackwell, Forthcoming. Expected late 2012.

[3] R.F. Baumeister and J.J. Exline. Virtue, personality, and social relations: Self-control as the moral muscle. *Journal of Personality*, 67(6), 1999.

[4] O.A. Blanson Henkemans, P.J.M. Van der Boog, J. Lindenberg, C.A.P.G. Van der Mast, M.A. Neerincx, and B.J.H.M. Zwetsloot-Schonk. An online lifestyle diary with a persuasive computer assistant providing feedback on self-management. *Technology and Health Care*, 17:253–267, 2009.

[5] ACM Council. Acm code of ethics and professional conduct. `http://www.acm.org/about/code-of-ethics`. Referenced on March 30th 2012.

[6] J.M. Fischer and M. Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, 1998.

[7] J.M. Fischer and M. Ravizza. Précis of responsibility and control: A theory of moral responsibility. *Philosophy and Phenomenological Research*, 61(2):441–445, 2000.

[8] B.J. Fogg. *Persuasive Technology: Using computers to change what we think and do*. Morgan Kaufmann Publishers, San Fransisco, 2003.

[9] R.E. Freeman. *Strategic management: A stakeholder approach*. Pitman, Boston, MA, 1984.

[10] B. Friedman. Value-sensitive design. *Interactions*, 3(6):16–23, 1996. ISSN 1072-5520. doi: 10.1145/242485.242493.

[11] B. Friedman and P.H. Jr. Kahn. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1):7–14, 1992. ISSN 0164-1212. doi: 10.1016/0164-1212(92)90075-U. Computer Ethics.

[12] B. Friedman and H. Nissenbaum. Software agents and user autonomy. In *Proceedings of the first international conference on Autonomous agents (AGENTS '97)*, pages 466–469, New York, NY, USA, 1997. ACM. doi: 10.1145/267658.267772.

[13] B. Friedman, P.H. Jr. Kahn, and A. Borning. Value sensitive design and information systems. In P. Zhang and D. Galletta, editors, *Human-computer interaction in management information systems: Foundations*, pages 348–372. M.E. Sharpe, 2006.

[14] A. Gewirth. *Reason and morality*. University of Chicago Press, Chicago, 1978.

[15] T.E. Jr. Hill. *Autonomy and self-respect*. Cambridge University Press, UK, 1991.

[16] M.C.A. Klein, N. Mogles, and A. Van Wissen. Why won't you do what's good for you? Using intelligent support for behavior change. In *International Workshop on Human Behavior Understanding (HBU11). Lecture Notes in Computer Science*, volume 7065, pages 104–116. Springer Verlag, 2011.

[17] R. Macklin. Dignity is a useless concept. it means no more than respect for persons or their autonomy. *BMJ*, 327:1419–1420, Dec 2003. doi: 10.1136/bmj.327.7429.1419.

[18] J.S. Mill. *On Liberty*. 1859. Reprint: Filiquarian Publishing, LLC, 2006.

[19] M. Muraven and R.F. Baumeister. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, 126:247–259, 2000.

[20] C. Nass and C. Yen. *The Man Who Lied to His Laptop: What Machines Teach Us About Human Relationships*. Current (Pinguin Group), New York, NY, 2010.

[21] M. Oshana. How much should we value autonomy? *Social Philosophy and Policy*, 20(2):99–126, 2003.

[22] D. Preuveneers and Y. Berbers. Mobile phones assisting with health self-care: a diabetes case study. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services (MobileHCI)*, pages 177–186, 2008.

[23] R.M. Ryan and E.L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[24] P-P. Verbeek. Designing morality. In *Ethics, Technology And Engineering: An Introduction*, chapter 7. Wiley-Blackwell, 2011.

Robot Companions as Case-Scenario for Assessing the "Subjectivity" of Autonomous Agents. Some Philosophical and Legal Remarks

Elettra Stradella[2] and **Pericle Salvini**[3] and **Alberto Pirni**[1] and **Angela Di Carlo**[1] and **Calogero Maria Oddo**[3] and **Paolo Dario**[3] and **Erica Palmerini**[1]

Abstract.
In this paper the European flagship project proposal Robot Companion for Citizens (RCC), grounded on the idea of developing robot companions for citizens, is taken as a case scenario for investigating the feasibility of ascribing rights and duties to autonomous robots from a legal and philosophical standpoint. In talking about rights and duties with respect to robots endowed with autonomous decision capabilities, one should face the implications that inevitably these terms rise, especially in the field of law. The paper points out the technological problems related to the application of the notion of duty to robots and the problems deriving from attributing a legal subjectivity to non-human entities such as robot.

1 INTRODUCTION

The legal problem of robotics, or legal gap, as it has been defined by [1], is the consequence of the new possibilities offered by technological advancements in artificial intelligence and robotics components (perception, computation and actuation), namely the possibility to have autonomous machines. In robotics, the term autonomy in general refers to the ability to perform a task in an unknown environment for a prolonged period of time without human intervention. An autonomous robot can be defined as 'a machine that collects information from the surrounding environment and utilises them to plan specific behaviours which allow it to carry out actions in the operative environment' [2]. The current legal systems, from East to West, are not ready to deal with robots that exhibit autonomous behaviours in human-inhabited environments. The most remarkable illustration is provided by the case of the Google Car. As a matter of fact, although the car is capable of driving autonomously, namely without the need of a human being, by law there must be a person on board, just for liability purposes. Things get even more complicated, from the regulatory point of view, if robots are endowed with learning capabilities.

In this paper the European flagship project proposal RCC (http://www.robotcompanions.eu), grounded on the idea of developing robot companions for citizens, is taken as a case scenario for investigating the feasibility of ascribing rights and duties to autonomous robots from a legal and philosophical standpoint. In talking about rights and duties with respect to robots capable of autonomous decisions, one should face the implications that inevitably these terms rise, especially in the field of law.

The paper is organized as follows: next section briefly explores the concept of autonomy, as well as the technologies that will be developed in the framework of the RCC project. In Section 3 the nexus between autonomy and duties is explored from a philosophical point of view. Section 4 deals with the rationale at the basis of the recognition of a subjective status to robot companions. It explores the cases of attribution of subjectivity to entities other than persons in Europe and attempts to extend such cases so as to include robotic agents as well. Finally in Section 5, the question concerning the need for having an autonomous subjectivity with respect to robot companions acting in the legal environment is analysed.

2 FROM CURRENT ROBOTICS TO ROBOT COMPANIONS FOR CITIZENS: AN OVERVIEW

The concept of autonomous agent applies to systems being either physically instantiated or not. The former case refers to embodied agents, such as robots, i.e. those agents having both brainware and bodyware and thus being directly capable of physical actions, while the latter refers to agents that have not an evident physical instantiation, such as the case of non-human operators in financial transactions (e.g., in stock exchange markets or in business-to-business platforms managing industrial supply chains).

Autonomous agents present both significant Scientific and Technological (S&T) challenges and related Ethical, Legal and Societal (ELS) implications, with particular reference to liability aspects associated to the deployment of autonomous agents in society.

Autonomy is inherently multi-scale depending of the layer of the control hierarchy being awarded with a degree of autonomy, or involving environmental or human influence in the decisional loop.

Autonomy may span from low-level control (e.g., in tracking a reference trajectory in the joint space of a robot), to task planning and execution given a specific objective (e.g., in identifying optimal trajectories while navigating between two locations), to the definition of specific objectives given a general objective (e.g., the

[1] Scuola Superiore Sant'Anna, Pisa, DIRPOLIS Institute
[2] University of Pisa, Pisa, Department of Law
[3] Scuola Superiore Sant'Anna, Pisa, The BioRobotics Institute

sequence of intermediate stops in product distribution chains), to management of energetic resources (e.g., energy saving and battery charge policies), to cloud robotics (e.g., agents sharing decisions and experiences over ICT infrastructures), to interaction and communication (e.g., the case of the "Chinese room thought experiment"), to the decision of strategic objectives in abstract form, etc.

All such layers and scenarios, from the low-level to the abstract one, present subtle aspects while attempting to define autonomy, as well as to differentiate an automatic control from a degree of autonomy. As a matter of fact, the concept of autonomy is directly connected to automatic control, though autonomy is much more controversial. Influence of past experience on future behaviours is not sufficient to characterize autonomy versus automatic control: a simple integrator is influenced by past experience, but nobody would assert the integrator to be an autonomous machine (rather, it is automatic, as a fundamental block of traditional control and automation theory).

A peculiar characteristics of an autonomous agent is the ability to develop and learn automatic behaviours and policies, and a higher degree of autonomy may be associated to a shift from low-level control towards higher order functions (as it is occurring to advanced robotic systems: Justin [3], the Jazz player robot musician [4], indoor and outdoor service robots [5], [6], just to mention a few) in applying novel and emerging machine learning approaches (as it is the case of the "Formal theory of creativity, fun, and intrinsic motivation" [7]). Previous experience and environmental constraints radically influence and may introduce bifurcations in shaping the evolution of agents endowed with machine learning methods or embodiment of computational functions [8], [9].

What are the associated ELS implications (particularly, with respect to liability aspects), given the potentially unmanageable and unpredictable variety of learning experiences and operational scenarios for agents being instantiated in unstructured physical environments?

Such questions will concern next generation robots, such as those that will be developed within the "FET Flagship Candidate Robot Companions For Citizens" (RCC). The RCC S&T programme proposes a radically new approach to develop machines and to truly deploy them in society as RCC Platforms: HealthCompanion, ExploreCompanion, WearableCompanion, WorkCompanion, UniversalCompanion.

The RCC highly ambitious programme is summarized by the RCC cross-domain grand scientific challenge: "To unveil the natural principles of simplexity, morphological computation and sentience and to translate the resultant scientific knowledge into design principles and fabrication technologies for Robot Companions that effectively and safely act, interact and adapt to their physical and social environment".

In particular, sentience is the ability to integrate perception, cognition and action in one coherent scene and context in which action can be interpreted, planned, generated and communicated [10]. Morphological computation is a novel paradigm asserting the role of materials in taking over some of the processes normally attributed to control [10]. Simplexity comprises a collection of solutions that can be observed in living organisms that, despite the complexity of the world in which they live, allows them to act and project the consequences of their actions into the future. Simplexity can be described as a property of living systems such that they can cope with the complexity of their world [10]. The highly ambitious RCC S&T programme will raise ELS issues, including liability aspects, which will be carefully managed and investigated in the RCC workplan, by means of dedicated and interdisciplinary teams composed by roboticists, experts in ethics, and lawyers. In this paper, we will start to approach such ELS issues, by focusing on the feasibility of ascribing rights and duties to robots.

3 WHICH AUTONOMY? A PROVISIONAL OVERVIEW WITHIN THE SPHERE OF DUTIES

When we try to focus such complex range of claims and issues through the lens of ethics, we must admit the necessity of dealing with a mass of problems, which are far from being captured and solved by both traditional and contemporary ethical theories [11]. The "Robot Companion" framework could indeed constitute a good chance to renew the toolbox of ethics, and surely the concept of autonomy is one of the most questioned in such field of ethics, the robot ethics, which takes seriously into account the new challenges introduced into the ethical domain through the developments of robotics.

Thus, just an overview to the topic of autonomy within the contemporary literature confirms that the debate has now achieved a level of maturity [12], [15]. This is perhaps a sign of the fact that current technological developments seriously begin to lay down the conditions for being able to discuss on such a topic, beyond any science fiction presuppositions. Moreover, another "travel into infinity" might occur to the researcher who wanted to reach a sufficiently wide competence about the so-called robot ethics or machine ethics [16], [17], [2] that constitutes the unavoidable framework for the attempt developed below.

The contemporary debate about robot ethics has developed some interesting results in such frame, firstly connected to the health-care robots [18], [19], but also to the particular context of child-care robots [20]. Furthermore, autonomy is an undoubtedly relevant task also for robotic warfare [21], [22].

In order to take a step forward in such framework, it could be useful to take a step back, by examining briefly, from another point of view, the concept of autonomy and the theoretical conditions of its attribution to an agent. It is surely trivial to affirm that assessing the status of autonomous agents with respect to robots is a problematic issue. In this context, we would briefly explore an articulation of the nexus between autonomy and duties [23] (another of the key-concepts of an ethical toolbox for robotics) that could support a less trivial way of posing that issue.

Starting with a short definition of duty, it is possible to recall a paradigmatic statement drawn from Th. Reid's Essays on the Active Powers of Man (1788) [24]. Duty is neither something that belongs exclusively to an agent («It's up to you!»; «You must, over and beyond any considerations!»), nor something that is intrinsically related to action («This action should be done!» «It's impossible not to do that»). Rather, duty is structurally and inseparably connected to both, or to agent and to action at the same time. In other terms, duty is a relationship between agent and action that triggers "spontaneously" and "mandatory" when a certain situation occur. For example: I see a person falling while she is walking in front of me and immediately I feel / perceive the duty (as subject) to help her to get up.

By remaining within the framework of duty, this (apparently) simple situation opens (at least) three areas of questioning. One is

related to the time of reaction, or: What does "immediately" in such a context mean? A second point regards the verb used in such situation: What do "feel the duty" or "perceiving the duty" mean? Last but not least: Which is the meaning of the word "agent", in relation to this situation? All these areas are widely discussed, in philosophy as well as in neurosciences, but also in roboethics (see [25], [26], [27]). For the purposes of this paper, the authors could just sketch synthetically the third one – and only a little portion of such problematic area.

The concept of agent, in relation to the claim of duty – and to such specific duty («help the person who is in trouble») –, needs at least the clarification of a central aspect. Any duty implies a power, conceived as "to be able to do something": if I have the duty to do a certain action, I must also have the power to do that action, I must be able to do what I am "obliged" to do. Otherwise, no practical question can exist, i.e. any question of ethical relevance.

It has been R.M. Hare [28] to identify this point with deep sharpness.

In its turn, the "power-to-do" issue should deal with a double question: firstly, with an external condition, that can be called "the possibility side": I had the duty to help the person who had fallen in front of me, but there was a ditch along the street (or another physical impediment) between me and her that I have not been able to exceed it. Consequently, the possibility to fulfil such a duty has been denied to me.

Secondly, the "power-to-do" issue should deal with an internal condition, which is – on its turn – intrinsically double. So, there is what can be called "the first level capacity side", I should have the ability to perform exactly the action I am obliged to do: I can do precisely the action of helping her to get up, for example, as I exactly know how to approach her and to surround his her shoulders while she is stretching out his her arms to get up. But it is also possible to distinguish a "second level capacity side", that implies the ability to do more than one sole action in order to answer to the duty-question in that / such situation ("help her to get up"). The agent can choose among different possibilities, all oriented to the goal of helping: I can grab her arms, or I can bend over, so she can lean on me. Still, I can try to stop the traffic, since she fell in the middle of a road and this is surely the first priority in order to help her. In other words, I can value by myself – "in complete autonomy"– what is the best action to do in this specific situation.

The entire question, related to both an external condition and to (at least) an internal one, could be considered as the core of every possible discourse about the attribution of autonomy to an agent. The authors have consciously chosen an example with multiple facets related to a task implying movement. And they are also aware that anyone of these trivial examples opens enormous problems of implementation, if it was possible to transfer the terms of such question to robots – and even larger problems would need to be questioned if the aim of this paper was to consider duties less related to physical aspects.

Nevertheless, a crucial point remains here at stake: It is possible to attribute duties to robots – and to open the discourse about this topic – without asking whether robots [can?] support [or not?] the set of conditions this section has tried minimally to enlighten?

Moreover, if this paper wanted to frame this issue at a greater distance, the authors would realize that it was only a half of a sphere, which finds its ideal completion in the legal dimension.

4 LEGAL SUBJECTIVITY AND ENTITIES OTHER THAN PERSONS: POSSIBLE INCLUSION OF ROBOT COMPANIONS?

In debating whether, one day, robots will have rights and duties, it is crucial to start wondering whether and how robots could become legal subjects, instead of ever remaining an object of the law.

Understanding the cases in which legal subjectivity is recognized to entities other than natural persons serves the purpose to answer the question: is a legal subjectivity for robots needed (or useful)?

In this context it is important to underline immediately that the concept of "subjects" and "subjectivity" that it is used in this paper does not refer to the philosophical notion, widely accepted by modern and contemporary philosophy. The use of these terms is a strictly legal use, functional to the aims of the authors.

Nevertheless, one has first to consider that the meaning and the nature of the "legal person" and "legal subject" concepts are still controversial. While nobody doubts that the human being represents the legal person par excellence, it is not unanimous how they acquire their legal capacity – namely the capability of being entitled of rights and duties – and whether other entities, which are not human beings, could be considered legal person in a specific legal system.

With respect to the first aspect, some scholars believe that the legal capacity is a natural feature of human beings, so that legal systems can just recognize it by law; on the contrary, others think that the legal capacity is a legal status that law awards to certain entities, as argued in Kelsenian theory. It is quite evident that the latter approach eases awarding the status of legal person to entities that are not persons. Associations, foundations and organisations are a significant example; indeed, the experience in existing legal orders shows that considering them as legal person gives rise to several issues and that the rationale of such an option has to be found in patrimonial responsibility [29], [30].

Nonetheless, further questions arise from the possibility of assigning the legal personality to entities that are not composed by a group of people, but are individual entities other than human beings. Can we speak about them as a legal person in legislation, since they are not people in real life? Some theories argue that the concepts of "person" and "human being" do not overlap at all. The scientific and technological progress in biology and medicine has led to rethink, especially at a philosophical level, these notions and the opportunity of including some stages of human life in the category of "person"; at the same time, they started to assume that other living beings, such as animals and plants, or even intelligent things would be considered as a person [31]. Engelhardt, for instance, believes that autonomous agents only, thanks to their potential capacity of self-determination, can be considered as a person, irrespective of their human or non human nature [32].

In any case there are no doubts that robotic technologies, whatever the level of autonomous capacity to determinate their actions would be, cannot be included in the notion of person. The intrinsic qualification of person prevents to assimilate to this ethical and juridical category any entity without a naturalistic dimension of life and self-awareness.

The Italian constitutional framework (and the constitutional framework of several European Union Member States) grounds on the "personality" principle to be interpreted as the general recognition of the fundamental human rights for every human

being, independently from their citizenship, economic, social conditions. The "recognition" of "inviolable" rights means, in the Italian constitutional context (Article 2), that human rights are the authentic base of society, and the human being is the true scope of the legal system and of the public organization of power. In the European Charter of Fundamental Rights, Article 1 introduces the concept of human dignity: this is the leading criterion in the definition of the axiological paradigm that can guide the possible attribution of some subjective status to robotic technologies.

Therefore, sometimes the law itself individuates a distinction between person and subject (or other forms of "subjectivities"), providing hypotheses of differentiation between the two notions.

This possible differentiation grounds on the distinction between two "laws": the law of legal rules (the positive law) and the law of society, "intrinsic to society as principle and rule of coexistence" [33]. The positive law has the mission to recognize the "subject", whilst the law of society would recognize the "person": this means that the positive law can create (legal) subjects that are not persons, but never denying the human being, with her form and substance and, before her, the capability to live [34].

Three are the main cases of differentiation between person and subject, and the individuation of subjects that are not included in the notion of person, that we can find in the European Member States law: i) unrecognized organizations and some kind of corporations without legal personality; ii) conceived baby before birth; iii) animals.

This paper aims at individuating the rationales that base these various recognitions in order to assess the eventuality to extend some of them to the possibility of recognition of robots' subjectivity.

In the first hypothesis (i) the subject is a sort of summation of natural persons that act in order to pursue common scopes, both economic-proprietary and not. The rationale seems to be the recognition to these entities of a legal capacity necessary to carry out the activities legally appertained to the single natural persons that make them.

In the second (ii) a status including (fundamental) rights is recognized to a subject that is potentially (natural) person: this subject must be protected under the umbrella of the principle of human dignity [35-37].

The rights to life, to psychic and physical integrity, are not legal goods lavished by the legal system on individuals. They directly derive from the belonging to a human society. Because the human being is person just for her evident existence in society; although the embryo cannot be considered as a person, it is a "human reality" in which we find all the dignity of the future human being.

It is possible to try an assimilation with the third category: (iii) animals.

In the core values of constitutionalism certainly we can individuate the base for the protection of rights of non-human species [38].

a) The constitutionalism protects the human being because she is holder of goods – for example physic and emotional integrity – that cannot be limited or abolished without determining an injustice: the limitation or the abolition would directly prejudice the condition of happiness of humans. From this point of view the creation of a "protective" status of subjectivity for the animal would derive from the consideration that the animal has got "sentience" too. How animal sentience could be described? It is evident that in this case "sentience" may be intended as the capacity to feel sensations of pain and pleasure, and in particular pain (physical and emotional – not strictly psychic because this would attribute to animals a "psyche" that could be ascribed to the possibility to self-determine in right and wrong).

In this framework the recognition of subjectivity is directed above all to the protection against behaviours aimed at (gratuitously) inflicting pain, and to clear - though partially - the relationship between the animal and its owner from a strict dimension of property rights.

Recently a theory has been developed in France – the Marguénaud's approach – according to which refusing to recognize human rights to animals does not mean denying at all the protection of certain animal interests. Another approach, supported by Joel Feinberg, an American law philosopher, considers animals equivalent to elderly, disabled people and minors from a legal point of view. As a consequence, they would be necessarily represented in order to fulfil their rights (Council of Europe, 2006).

In Europe, the first laws on animal protection were approved at the beginning of the XX century. Since 1968 the Council of Europe approved five Conventions for the protection of: animals during international transport (1968, revised in 2003); animals kept for farming purpose (1976); animals for slaughter (1979); vertebrate animals used for experiments (1986); pet animals (1987). Provisions for animal rights have been included in the national Constitutions of Switzerland (1992, 2000) and Germany (2002), while the EU Lisbon Treaty (2007) states that the Union and the Member States shall, since animals are sentient beings, pay full regard to the requirements of animal welfare. In the United States, despite the Constitution does not mention animals, a US federal judge was asked to rule on whether animals take benefit of the constitutional protections against slavery as human beings; thus, the judge ruled that slavery is uniquely a human activity, as those terms have been historically and contemporaneously applied, and there is no basis to construe the Thirteenth Amendment as applying to non-humans.

b) Jurisprudence and case-law in various European countries unanimously confirm the existence of a human right to the protection of biosphere, the equilibrium among species, and set up a right of future generations to a healthy environment and a sustainable management of environmental resources and ecosystem. Animals are of course part of this reality and, from this point of view, they can be seen as instrumental goods to the protection of human rights, and therefore recognized as subjects (or subjectivities) to be protected by the legal system.

c) Animals, and pets in particular, have an "emotional" relation with humans, contributing to their wellbeing and the development of their personality. The main objective of the Western constitutionalism and the aim pursued by legal systems as described by the most important Constitutional Charters in Europe and in the other Western Countries is undoubtedly the development of personality, the happiness, or the fulfilment of a strong interpretation of the human dignity principle. In this third framework the recognition of subjectivity would constitutionally ground on the protection and promotion of a "relational good" [39].

In order to investigate the possibility to give the RCs a subjectivity, it is necessary to understand whether some of these elements could regard robotic technologies as well.

Certainly b) can be excluded without need of motivations.

Indeed some reflections could be made about a) and c).

With regards to a) the definition of "sentience" is decisive, in the specific meanings applied to animals and to RCs, as briefly discussed in Section 2.

The animal's sentience is today quite well known by ethologists: they underline that "a fairly solid body of information about what animals are feeling is collected by indirect means. They have been assembled about states of suffering experienced by farm animals such as pain, fear, frustration and deprivation" [40]. They use various methods in order to define a pain assessment in animals [41], and the results provide evidence that the animal would be able to experience negative sensations similar than the human ones, suitable to raise the demand of justice mentioned above [42-44].

The different content of "sentience" in the animal in comparison to RCs prevents the recognition of a legal subjectivity for animal and the (prospective) recognition of a legal subjectivity for robots to be ascribed to the same rationale.

With regards to c), it is worthy to point out that the RC could build (is supposed to build) a "personal" relationship with the individuals who "use" it, and that examples of robotic technologies with emotional-relational functions already exist (e.g., the case of the well-known Paro robotic therapeutic seal). Nevertheless, because of the extreme subjectivity of the capacity of an entity (or simply a thing) to represent an emotional object and an instrument of happiness for an individual, this element does not seem sufficient to ground the recognition of legal subjectivity (that could otherwise concern televisions, cars, computers, etc.).

5 ROBOT COMPANIONS ACTING IN THE LEGAL ENVIRONMENT: IS THERE A NEED FOR AN AUTONOMOUS SUBJECTIVITY?

Assigning legal capacity to RCs as an acknowledgment of their peculiar status of "sentient beings", comparable to animals, is an issue to let open at the present moment. Nonetheless, the option of recognizing them as persons in a legal sense has to be analysed from a more empirical and functional perspective as well. First of all, the prospect of creating companion robots devoted to assist elderly and disabled people requires to provide them with the ability of rendering basic services that go beyond acts of pure material care. People with reduced capacity to move around, to carry weights or even to speak out their wishes in verbal ways have to be assisted and helped also in purchasing goods, such as food, drugs, newspapers, bus tickets. This means that the technology would be more helpful and worthy whereas robots were provided with the ability of performing legal transactions. Many operations a companion robot could be asked to carry out effectively imply entering into a contract. Assigning legal capacity to a robot, in this sense, could solve the problem of having a centre of imputation of the effects deriving from the agreement and avoiding the contract to be considered void. Such an option may appear redundant because the transactions done by robots are deemed to be elementary and of minor value; moreover, they are normally immediately executed, hence most often contractual remedies would not be called to intervene. Nevertheless, one cannot exclude in principle that disputes will arise and that the problem of identifying the contractual parties, and their capacity of entering a transaction will become controversial. Therefore the need of referring the contract to someone to be held responsible with regards to its effects remains. A plain answer could be to consider robots as a sort of extension of their users' will and physical body, so that any act they execute is directly referable to them. On the one hand, this solution would circumvent the conceptual difficulties of awarding robots full capacity; the same we would encounter also by accepting that robots are simply mandated by their users, because the latter option equally requires to confront the issue of capacity for the deputy. On the other hand, it appears rebuttable under two aspects: it is counterintuitive, because of the detachment and possibly the physical distance between the primary actor and his supposed offshoot; most of all, it does not take into account the limited, but not inexistent, autonomous decision-making ability the robots companions are doomed to have. Another possibility is to consider the companion robots as autonomous agents, endowed with the status of subjects, but capable of entering into transactions under certain constraints. The reduced capacity of minors or of the mentally impaired, known and disciplined in the current legal systems, could be taken as a model for regulation. Under this special regime, robots would be entitled to act validly but only with regards to transaction of minor importance and value, those that are needed in order to satisfy the basic necessities of their users (See, for instance, art. 409, comma 2, of the Italian Civil Code).

Another practical reason suggests to investigate the possibility of awarding some kind of legal capacity to the robots companion, that is the issue of liability for damages. Ensuring the safety of these devices trough careful design and manufacturing does not exclude that accidents might occur either to their users or to third parties. Hence the crucial question of who and under which circumstances is responsible for the damages brought about by robots. The stance taken on the status of autonomous agents of RCs becomes decisive in order to frame properly the problem of liability. More precisely, it is necessary to appreciate whether the existing rules about producer's liability or liability deriving from the ownership or possession of things apply; or if the technology is so highly developed and advanced, and provided with a certain degree of decision-making ability, that the rationale underlying those sets of rules cannot operate. The concern should be about fixing a general divide between traditional machines, that can be designed and manufactured so that their behaviour will be predetermined or predictable by the constructor and afterwards mastered by their user; and sophisticated robots, that do not correspond to this archetype. If the robots companion belong to this latter category and cannot therefore be entirely controlled, we need to part from a rule that assigns liability precisely on the basis of the power that the subject who responds for the damages can exert over the sphere of the actual agent. Again, the basic structure of most legal regimes regarding injuries caused by minors and incompetent persons could be taken as a model rule. The two cases share some common features: the limited capacity of the agent, not sufficient to held her fully responsible for the damages he has produced; but also an independence of action, more or less substantial, that the agent exhibits and that accounts, at the same time, for the possibility of the guardian to be exonerated by demonstrating not to be at fault (or to have adopted every reasonable precaution in order to avoid accidents).

Recognising the autonomy of RCs, be it limited and only "functional", may result in the potential attribution of duties or obligation, deriving from the agreements undertaken or stemming by the wrongs committed. Nevertheless the legal mechanisms thus evoked, both contractual and non contractual liability, are not self-sufficient. If robots do not have assets to make up for their obligation or to compensate for damages, to hold them liable without providing a vicarious responsible will not make sense. The supplier would not get paid, the victims could not recover

damages, if we stick to the previous examples. The prospect of assigning legal capacity to the RCs for practical, instead of ontological, reasons definitely requires to implement other instruments through which these can be achieved.

6 CONCLUDING REMARKS

Besides an ethical and social problem concerning robotic technologies, there is also a legal one. Simply speaking, the former problem deals with whether it is right or wrong to carry out research in a specific way or field or to deploy robots in certain contexts or for certain tasks. Many relevant arguments in favor and against robotics research and applications have been raised by scholars in the last years [45-52]. On the contrary, the legal problem does not seem to care too much about the issue of robots' legal subjectivity, whilst it should be a preliminary question to pose.

Of course every attempts to regulate new and emerging technologies should be accompanied by careful ethical and social analyses as well as risks and safety analysis. Too often science and technology have been embraced uniquely on the basis of political, economic and/or scientific interests. The truism that the possibility to do one thing (e.g. make robots autonomous) is not enough for justifying its accomplishment is even more true in case of machines which should interact or coexists with human beings.

In addressing the issue of rights and duties of autonomous software or robot agents, therefore, a preliminary question should be concerned with the ethical and social implications ensuing from their deployment.

On the other hand, the issue of rights and duties should be considered as a "second level" topic to be addressed: it is necessary to assess the (legal) possibility and significance of a recognition of subjectivity for autonomous agents. In other words, 'to define regulations and control mechanisms to ensure sound and consistent behaviours' maybe is not enough.

The RoboLaw project, funded by the European Commission (EC) in the Seventh Framework Programme (www.robolaw.eu) aims at providing the EC with new knowledge on the regulation of robotics technologies. The most relevant result of the project will be a White Paper guidelines for the regulation of emerging robotic technologies. However, the RoboLaw goal is not just to provide roboticists with legal regulations for bringing their inventions outside their laboratories, but to deeply analyse the impact of robotic technologies and applications on traditional legal concepts and constitutional rights.

In this paper, in talking about Rights and Duties with respect to autonomous agents a few critical issues have been pointed out.

May we apply to current robots and to the RCs the (philosophical and legal, philosophically grounded) notion of duty? They do not seem to support the set of conditions that pertains to the notion of duty.

May we recognize them a legal subjectivity? It seems very hard to individuate a "reasonable rationale" that could ground this kind of choice, comparing robots with the other "legal subjects", different than natural persons, already existing in the Western legal framework. Finally, awarding a legal status to robots companion may be necessary according to a more functional perspective. If they operate in a living and therefore legal environment, rights and duties are simply a legal tool for implementing the technology and better reaching the social goals to which it is devoted.

According to this functional perspective it seems inappropriate to use "binding" legal concepts like rights and duties (and autonomy) are, and, instead, it appears more suitable a case-by-case application of existing legal instruments provided for other machines.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 289092 RoboLaw project. This work was supported in part by the EU under the CA-RoboCom Project (FP7-ICT-2011-FET-F, P.N. 284951).

REFERENCES

[1] Y.H. Weng, C.-H. Chen, C.-T. Sun. Toward the Human-Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots. International Journal of Social Robots, 1:267-282, (2009).

[2] P. Lin, K. Abney, G.A. Bekey (eds), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, The Mit Press 2012.

[3] http://www.youtube.com/watch?v=93WHRSKg3gE

[4] http://www.autonomousrobotsblog.com/robotic-musicianship/

[5] www.irobot.com

[6] P. Salvini, G. Teti, E. Spadoni, C. Laschi, B. Mazzolai, P. Dario Peccioli: The Testing Site for the Robot DustCart. Focus on social and legal challenges Ieee Robotics And Automation Magazine - vol. 18, Issue 1 : 59:67 (2011)

[7] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Transactions on Autonomous Mental Development, 2(3), 230-247, (2010).

[8] R. Pfeifer, F. Iida, G. Gómez. Morphological computation for adaptive behavior and cognition. International Congress Series, 1291, 22–29, (2006).

[9] R. Pfeifer, J.C. Bongard. How the Body Shapes the Way We Think A New View of Intelligence. The MIT Press 2006; M.P. Wellman, A. Greenwald, P. Stone, Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition, Cambridge, The Mit Press 2007.

[10] The Robot Companions for Citizens Consortium, The Robot Companions for Citizens Manifesto, www.robotcompanions.eu, July 2012.

[11] R. von Schomberg, From the ethics of technology towards an ethics of knowledge policy: implications for robotics, AI & Soc, 22, 331–348 (2008).

[12] M.P. Wellman, A. Greenwald, P. Stone, Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition, Cambridge, The Mit Press 2007.

[13] L.S. Sterling, T. Taveter, The Art of Agent-Oriented Modeling, Cambridge, The Mit Press, 2009.

[14] K. Stoy, D. Brandt, D.J. Christensen, Self-Reconfigurable Robots. An Introduction, Cambridge, The Mit Press, 2010.

[15] D. Floreano, C. Mattiussi, Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies, Cambridge, The Mit Press, 2008.

[16] R. Capurro, M. Nagenborn (ed.), Ethics and Robothics, Amsterdam, IOS Press, 2009.

[17] A. Beavers, Robot Ethics and Human Ethics, Special Issue of Ethics and Information Technology, 2010, 12 (3).

[18] E. Datteri, G. Tamburrini, Ethical reflections on health care robotics, in R. Capurro and M. Nagenborg (Eds.): Ethics and Robotics. Heidelberg: Akad.Verlagsgesellschaft, (2009) (IOS Press).

[19] J. Borenstein, Y. Pearson, Robot caregivers: harbingers of expanded freedom for all?, Ethics Inform. Technol. 12, 277–288 (2010).

[20] N. Sharkey, A. Sharkey, The crying shame of robot nannies: an ethical appraisal, Interaction Studies, 11 (2), 161-190 (2010).
[21] P. Sullins, RoboWarfare: can robots be more ethical than humans on the battlefield?, Ethics Inf Technol., 12, 263–275 (2010).
[22] M. Guarini, P. Bello, Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters, in P. Lin, K. Abney, G.A. Bekey (eds), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, The Mit Press 2012, 129–144.
[23] L. Fonnesu, Dovere, Firenze, La Nuova Italia 1998.
[24] Th. Reid's Essays on the Active Powers of Man [1788], in D.D. Raphael (ed.), British Moralists, Indianapolis, Hackett 1991, Vol II.
[25] C. Allen, G. Varner, J. Zinsen, Prolegomena to Any Future Artificial Moral Agent, «Journal of Experimental & Theoretical Artificial Intelligence», 2000, 12 (3), 251-261.
[26] M. Anderson, S. Anderson, Machine Ethics: Creating an Ethical Intelligent Agent, «AI Magazine», 2007, 28 (4), 15-26.
[27] M. Anderson, S. Anderson, Machine Ethics, New York, Cambridge University Press 2011.
[28] R.M. Hare, Freedom and Reason, Oxford, Clarendon Press 1963.
[29] A. Falzea, voce Capacità (teoria gen.), in Enc. dir., VI, Milano, 1960, p. 34 s.
[30] G.L. Pellizzi, Soggettività giuridica, Enc. giur. Treccani, 1990.
[31] L. Palazzani, Il concetto di persona tra bioetica e diritto, Torino, Giappichelli, 1996.
[32] H.T. Engelhardt, Some Persons are Humans, Some Humans are Persons, and The World is What We Persons Make of It, in S.F. Spicker – T.H. Engelhardt (eds.), Philosophical medical ethics: nature and significance, Reidel, Dordrecht, 1977.
[33] G. Oppo, Ancora su persona umana e diritto, in Riv. dir. civ., 2007, I, 259 ss.
[34] G. Oppo, L'inizio della vita umana, in Riv. dir. civ., 1982, I, 499 ss.
[35] F.D. Busnelli, L'inizio della vita umana, in Riv. dir. civ., 2004, I, 533 ss.
[36] M. Basile, A. Falzea, Persona giuridica, in Enc. dir, XXXIII, Milano, 1983.
[37] F. Riccobono, Soggetto, persona, diritti, Napoli, Terzo Millennio, 1999.
[38] G. Gemma, Costituzione e tutela degli animali, in Forum Costituzionale, 2004.
[39] Animal Welfare, Council of Europe, 2006.
[40] I.J.H. Duncan, The changing concept of animal sentience, in Applied Animal Behaviour Science 100(2006), 11-19.
[41] D. M. Weary, L. Niel, F. C. Flower, D. Fraser, Identifying and preventing pain in animals, in Applied Animal Behaviour Science 100 (2006), 64–76.
[42] M. Nussbaum, Justice for Non-Human Animals, The Tanner Lectures on Human Values, November 13, 2002.
[43] R. Scruton, Animal Rights and Wrongs, 3rd ed., London, Metro, 2000.
[44] M. Scully, Dominion: The Power of Man, the Suffering of Animals, and the Call to Mercy, New York, St. Martin's Griffin, 2002.
[45] J.M. Galvan, On Technoethics, in «IEEE Robotics and Automation Magazine», December, 2003.
[46] Lin P., Abney K., and Bekey G.. 'Robot ethics: Mapping the issues for a mechanized world, Artificial Intelligence 175.5-6, (2011).
[47] G. Veruggio, F. Operto: Roboethics: Social and Ethical Implications of Robotics, Springer Handbook of Robotics, 2008.
[48] Salvini P., Laschi C., Dario P., 'Roboethics and Biorobotics: Discussion of Case Studies', Proceedings of the ICRA 2005, IEEE International Conference on Robotics and Automation, Workshop on Robot-Ethics, Rome, Italy, April 10, 2005.
[49] D. Marino and G. Tamburrini, Learning robots and human reponsibility, Int. Rev. Inform. Ethics 6, 46–51 (2006).
[50] D. J. Calverley, Imagining a non-biological machine as a legal person, AI Soc. 22, 523–537 (2008).
[51] S. N. Lehman-Wilzig, Frankenstein unbound: toward a legal definition of artificial intelligence, Futures 13, 442–457 (1981).
[52] A. Matthias, The responsibility gap: ascribing responsibility for the actions of learning automata, Ethics Inform. Technol. 6, 175–183 (2004).

Principal and Helper: Notes on Reflex Responsibility in MAS

Clara Smith[1]

Abstract. What justifies -in the head of another agent different from the one acting- the obligation to compensate is the fact that the principal agent has lengthen its own action through the implementation of a *foreign activity* for its own interests. We present two basic modal operators for representing, respectively, intentions in the interest of another agent and agency in the interest of another agent. They appear useful enough for characterizing the notion of reflex responsibility in a multi-modal multi-agent system (MAS) context.

1 MOTIVATION AND AIMS

As pointed out by Chopra and White [1], theorizing in domains such as legal and cognitive status of agents is crucial for designers of agents, especially, for the design of "on demand" MAS. Within such engineering account, a legal question arises: do the designed agents are to be autonomous enough to have rights and responsibilities? (By being autonomous we *at least* mean that agents act to achieve their own goals cf. Conte and Castelfranchi [2].)

Most works on the topic are centered on "contractual issues" (see e.g. [3,4]). Chopra and White point out four approaches as one moves up the sophistication scale of agents: three "weak" positions, based on (i) the idea of agents as mere tools of their operators, (ii) the *unilateral offer* doctrine (a contract formed by a party´s offer plus an acceptance, stipulated in the offer), and (iii) the *objective theory of contractual intention* (a contract –usually words- is an obligation which is law to the parties, who have the intention to agree), plus a fourth, radical one, which involves treating artificial agents as the legal agents of their operators. There is also a fifth position that postulates the legal systems treating agents as legal *persons*.

In this work we focus on the legal binding between a principal agent and a dependent agent. Particularly, we are interested on the dependent's performance that has its origin in *extra contractual* situations e.g. factual and/or occasional situations, trust, or courtesy. Examples of such bindings occur e.g. between the owner of a car -or any other device- and the one who drives it with the owner´s authorization (and without a proper title for using it), blog activities such as *twitting* in the name of another, or bidding in an auction in the interest of another: performed in the interest of a principal agent. In these situations is enough that the principal wills to be bind to third parties through the helper's or dependent's performance.

We therefore keep apart from our analysis situations in which performance in the interest of another has a contractual basis. This because, in contracts, function for another one and subordination may be rather straightforward to identify, mainly because there is a notion of obligation involved. If an agent gives some explicit orders or instructions to another agent which acts as his helper, or if and agent is obliged through a contract in the interest of another agent (even when agents voluntarily engage into contracts because of their own utility), or if an agent *h* forms part of agent *p*'s business organization, subordination is somehow established. We therefore exclude here cases such as mandates and any conferral of a power of representation accompanied by an obligation of representation in certain ways (e.g. a cheque is a mandate from the customer to its bank to pay the sum in question.)

We give a definition for the concept of reflex responsibility between a principal agent and a helper agent, mainly inspired on general provisions settled by Italian and Argentinean provisions for *persons*. We indeed use the terms "does", "performance", and "action" as referring to persons, although it is not entirely clear for us if it is meaningful to speak of the *actions* of devices, and artificial agents within highly automates systems; possibly a term as "executes" sounds more suitable. In what follows, "does" has the usual expected anthropomorphic meaning. The definitions we give may be useful as a step towards a specific notion of reflex responsibility of *artificial agents*.

Article 1113 of the Argentinean Civil Code states that "The obligation of the one who caused damage is extended to the damages caused by those under his dependence." In its turn, art. 1228 of the Italian Civil Code settles that "except a different will of the parties, the debtor who profits from the work of a third party for fulfilling the obligation is responsible for malicious or negligent facts carried out by that third party."

According to general doctrine and jurisprudence related to such articles, reflex responsibility has a subjective basis. This is a reason that makes reflex responsibility challenging to represent: if it had an objective basis, checking the standard legal extremes would be sufficient. Requisites for reflex responsibility are: i) the existence of a dependence relationship between principal and helper (or dependent), ii) the successful performance of an illegal action carried out by the helper, iii) that such performance was carried out while exercising a subordinate incumbency, iv) that such performance provoked a damage or injury to a third party, and that v) there must be an efficient causal relation between the helper's act and the damage caused.

Regarding the formal framework, we use as a basis a BDI multi-agent context for dealing with agents' attitudes, extended with generic obligations, as in [5]. A = {*x*, *y*, *z*...} is a finite set of

[1] Universidad Nacional de La Plata, and FACEI, Universidad Católica de La Plata, Argentina. csmith@info.unlp.edu.ar.

agents, and P = {p, q, r, ...} is a countable set of propositions. Complex expressions are formed syntactically from these, plus the following unary modalities, in the usual way: $Goal_x\ A$ is used to mean that "agent x has goal A", where A is a proposition. Propositions reflect particular state-of-affairs cf. B. Dunnin-Kepliçz and R. Verbrugge [6]. $Int_x\ A$ is meant to stand for "agent x has the intention to make A true". The doxastic (or epistemic) modality $Bel_x\ A$ represents that "agent x has the belief that A". The deontic operator O represents generic (legal/lawful) obligations, meaning "it is obligatory that" [7]. The operator $Does_x\ A$ represents successful agency in the sense given by D. Elgesem, i.e. agent x indeed brings about A [8]. For simplicity, we assume that in expressions like $Does_x\ A$, A denotes behavioral actions concerning only single conducts of agents such as withdrawal, inform, purchase, payment, etc. (i.e. no modalized formulas occur in the scope of a Does.) As classically established, Goal is a K_n operator, while Int and Bel are, respectively, KD_n and $KD45_n$. O is taken to be a classical KD operator. These are all normal modalities. The logic of Does, instead, is non-normal [8,9].

The rest of the paper is organized as follows. Section 2 addresses one possible characterization for a certain notion of dependence which happens to be complex enough and central to the lawful concept of reflex responsibility we deal with. We attempt four subsequent definitions, each of which improves the previous one. We go through them by using several examples. A relativized modality is introduced for dealing with oriented, coordinated intentions: an agent intends to become true a state-of-affairs A in the interest of another agent. We introduce in Section 3 another modality, a directed agency operator that binds a helper agent h to the principal agent p, and to the "oriented" action or situation A that h carries out in the interest of p. In Section 4 we formally define reflex responsibility of p regarding h with respect to an action or state-of-affairs A when: there is dependence between p and h w.r.t. A, h succeeds on carrying out A on account of p, such action constitutes an illegal act, and there is a damage a third agent t suffers which is attributable to h´s performance of A on account of p. Section 5 presents the underlying logical structure and the corresponding semantics. Conclusions end the paper.

2 DEPENDENCE

A requisite for reflex responsibility to hold is the dependence between the author of the harmful act and the agent to whom the responsibility is attributed by reflex, i.e. the principal. Such relation has two constitutive elements: 1) there is a *function* the helper carries out, on the principal´s utility; and 2) the helper is a subordinate of the principal w.r.t. the performance of such function, i.e. there is a subordinate incumbency.

Examples and non-examples of dependence relations. There is dependence between the owner of a car –or other device- and the one who drives it with the owner´s authorization (and without a proper title for using it), some blog activities such as twitting in the interest of another, or bidding in an auction in the interest of another. There is dependence between an artificial helper agent that occasionally accesses my email account profile and transfers part of its content to its (yet artificial) principal agent, which performs some data mining and later shows me tuned web ads. There is no impediment for dependence when the son works under the orders of his father, or if the daughter drives the car of her mother, who is being transported in it (there is occasional dependence). Neither parental relationships nor marriage is an impediment for the configuration of a dependence relationship.

Here then, dependence excludes delegation or mandate. There is no dependence between the car owner and the car-shop where the car is left in order to be repaired, except if the owner has authorized its use; neither between a student of a public school and the State, neither between the owner of a field and the firm in charge of its fumigation (all examples according to jurisprudence in [10].)

Definition 1. Dependence is a relation that holds according to certain internal states of agents. Let p be the principal agent and h the helper agent. Let A be a single behavioral action (e.g. pay, bid, tweet, etc.). A plausible initial characterization of dependence between p and h regarding A is: A is one of p's goals, p has the intention that h indeed carries out A, and h intends to make A true believing (knowing) that A is one of p's goals:

$$Dep^p{}_h\ A \equiv Goal_p\ A \wedge Int_p(Does_h\ A) \wedge \wedge Int_h\ A \wedge Bel_h(Goal_p\ A)\ . \quad (1)$$

Discussion. h adopts p's goal (A) as its own intention ($Int_h\ A$) in the exercise of, e.g., courtesy. Based on this fact, h will carry out A. Note that the last two conjunctors in (1) are meant to capture the idea of "function for another one" (h intends to become A true because he knows it is p's goal).

Nonetheless, (1) holds when it happens to be no subordination, or is merely a coincidence, or p would like that h does A and h does it for other reasons. For instance, (1) holds in a situation where p and h are -rather than principal and helper- rivals involved in a competitive scenario, i.e. both effectively having the same goal and aiming to fulfilling it.

Example 1. **The Bach Double Concerto.** Consider the two violinists' example in [6] where two violinists intend to perform the two solo parts of the Bach Double Concerto. (The Concerto for 2 Violins, Strings and Continuo in D Minor is characterized by the subtle yet expressive relationship between the violins throughout the work.) Let us revisit the example: suppose *Peter* is a violinist who has as goal being one of the soloists. Moreover, *Peter* also has the intention that *Helen*, his past fiancée -who is also a violinist- plays as the other soloist (he would like that). But as far as *Helen* goes, she intends to become one of the chosen soloists without care of who the other soloist is (and whatsoever part she plays); nonetheless, for sure she knows that *Peter* aims to play himself as a soloist too. We get that $Goal_{peter}$*play* $\wedge$ $Int_{peter}(Does_{helen}$ *play*$) \wedge Bel_{helen}(Goal_{peter}$ *play*$) \wedge Int_{helen}$ *play* holds although there is no dependence between *Peter* and *Helen* (assume that they

currently have no relationship at all!): *Helen* is in competence with *Peter*.

Let us attempt an improvement for our definition.

Definition 2. There is dependence between p and h regarding A when p has A as goal, h believes on this, and such p's goal is what induces h to have the intention to carry out A:

$$\mathrm{Dep}^{p}{}_{h}\, A \equiv \mathrm{Goal}_{p}\, A \wedge \mathrm{Bel}_{h}(\mathrm{Goal}_{p}\, A) \wedge \wedge (\mathrm{Goal}_{p}\, A \rightarrow \mathrm{Int}_{h}(\mathrm{Does}_{h}\, A)) \quad . \qquad (2)$$

Discussion. The conditional here is meant to specify that p's goal is the motive for h's intention.

Expression (2) may even hold in a rivalry scenario such as the Bach Double example. Suppose that *Helen*, knowing that *Peter* has as goal being one of the soloists, triggers her own interest in being a soloist, due to her competitive personality (and not based in any interest in *Peter*). Note also that in (2) it is sufficient that p has a goal, and that it is not necessary that the he wants h to be engaged. Then (2) also holds in a scenario where p does not want to be helped by h.

Example 2. The unwanted helper. I want my netbook to be fixed, but not by Harry who is incompetent; Harry, who does the job, satisfies my goal and qualifies as a helper.

Let us attempt a further improvement.

Definition 3. h's action will be triggered on the basis of p's intention that h does A (and h is aware of this), and not merely based on p's goal:

$$\mathrm{Dep}^{p}{}_{h}\, A \equiv \mathrm{Goal}_{p}\, A \wedge \mathrm{Int}_{p}(\mathrm{Does}_{h}\, A) \wedge \wedge (\mathrm{Bel}_{h}(\mathrm{Int}_{p}\mathrm{Does}_{h}\, A)). \qquad (3)$$

Discussion. Harry would not qualify as a helper under this definition, because I do not have the intention that he repairs my netbook (he will not carry out the task on my utility, I do not want him to).

Unfortunately, (3) still holds under rivalry between p and h w.r.t. goal A. (For a more *artificial agents*' scenario, assume any state-of-affairs in which automatic allocation of resources is in permanent dispute, and devices are not necessarily dependent one of each other.)

We next attempt a new definition that excludes rivalry situations by introducing a primitive, relativized operator, that coordinates two agents to an intention with regard to A. Binding p with h trough an "oriented" intention is what we need to exclude competitive situations.

Definition 4. Intention in the interest of another. We define a relativized operator: $\mathrm{Int}^{p}{}_{h}\, A$, meaning "$h$ intends A to be true in the interest of p". This way, we model dependence as a *coordinated relation*, as follows. The principal indeed must have the intention that the helper performs the task, while the helper is aware. He will somehow be "activated" not only by the belief that the principal intends that s/he does the task but also with his own "oriented" intention, in the interest of p, to carry out A. Formally:

$$\mathrm{Dep}^{p}{}_{h}\, A \equiv \mathrm{Goal}_{p}\, A \wedge \mathrm{Int}_{p}(\mathrm{Does}_{h}\, A) \wedge \mathrm{Bel}_{h}(\mathrm{Int}_{p}\mathrm{Does}_{h}\, A) \wedge \mathrm{Int}^{p}{}_{h}(\mathrm{Does}_{h}\, A) \qquad (4)$$

which stands for "A is one of agent p's goals, and p intends that h performs A; h is aware of this, and intends to become A true in the interest of p".

$\mathrm{Int}^{p}{}_{h}\, A$ allows capturing custom or courtesy behavior: h may be an altruistic agent not expecting any reward, merely intending to fulfill p's expectations, even occasionally. Observe that (4) indeed reflects the power of the intention in the interest of another, as such "directed" intention defines dependence as an oriented, coordinated, non-competitive relation.

Improvements regarding the intensional basic operators have already been addressed through e.g. the concept of deadline intentions and deadline beliefs. [11,12]. For example, suppose that agent y does not believe that agent x is travelling, and says "I won't believe he is travelling until he shows the ticket to me": we write a deadline *belief* using the until operator as $\mathrm{U}(\mathrm{Does}_{x}\mathrm{ShowsTicket}, \neg \mathrm{Bel}_{y}\mathrm{Travels})$ [12]. Moreover, collective intention operators for mutual and common intentions have been designed based on the basic Int operator in [6]. Relativised obligations to bearers and counterparties are defined in [13].

3 ACTION ON ACOUNT OF ANOTHER

Another requisite for the emergence of reflex responsibility is that law is violated (a legal aspect has now emerged.) The illegal act must be imputable to the helper, who is the one who materially and effectively *acts,* therefore he becomes materially responsible for the forbidden act. For the reflex responsibility to raise it is essential that the helper agent carries out the harmful activity *on account of* the principal.

We have gone through the discussion on directed intentions. It must be clear at this point that we also need an oriented/directed agency operator for coordinating h, p and the proper "oriented" action h carries out in the interest of p. Let us illustrate with an example.

Example 3. The truck driver. *d,* the occasional driver of p's truck, takes the truck off from p'*s* garage on Sunday afternoon, with a view to have a ride with his friends. Due to his misguidance, his friends are injured on the occasion of this Sunday drive ($\mathrm{Does}_{d}\, \mathit{drive} \rightarrow \mathrm{injure}_{\mathrm{friends}}$).

Discussion. p has as goal that d drives his truck, and intends him to drive it, d believes in this, and d has the intention to drive the truck in the interest of p. So we get dependence between h and p regarding A (i.e. (4) holds). Now, note that provided the general obligation that states that we should not harm others ($\mathrm{O}\neg\mathrm{injure}_{v}$), p´s reflex responsibility is about to raise. But d drove in its own interest. What justifies in the head of another agent -different from the one acting- an obligation to compensate is that the principal

agent has lengthen its own action through the implementation of a *foreign activity* for its own interests. Here is not the case, *d* drove on his own account when he provoked the accident.

We know that it is essential for reflex responsibility to hold that the performing agent carries out the task on account of another agent. We should be able, then, to distinguish those directed intentions and actions that we make in our own interest from those which we do in the interest of another one.

Definition 5. Agency in the interest of another. We introduce a relativized operator $\text{Does}^{p}_{h}\, A$ to represent agency in the interest of another, meaning "*h* carries out *A* in the interest of *p*". This non-normal operator is meant to capture *performance for another one* i.e. directed material performance in the head (and/or hands, or executable code) of *d*, but on account of *p*. This way, we establish oriented agency as a basic type of event, the same way as Does is.

This relativised agency operator leads us to a more precise definition for dependence:

$$\text{Dep}^{p}_{h}\, A \equiv \text{Goal}_{p}\, A \wedge \text{Int}_{p}(\text{Does}^{p}_{h}\, A) \wedge \\ \wedge (\text{Bel}_{h}(\text{Int}_{p}(\text{Does}^{p}_{h}\, A)) \wedge \text{Int}^{p}_{h}\, A \ . \qquad (5)$$

Back to the truck example, we have that, that Sunday, Does^{d}_{d}*drive* holds and also $\neg(\text{Int}_{p}(\text{Does}^{p}_{d}\ \textit{drive}))$ holds, making (5) false. (Note that, intuitively, $\text{Does}^{d}_{d}\, A$ collapses to $\text{Does}_{d}\, A$.)

4 REFLEX RESPONSIBILITY

We saw that another requisite for the emergence of reflex responsibility is that the helper's harmful performance provokes a damage or injury to a third party, let us say *t*, and that there must be an efficient causal relation between *h*'s performance -on account of *p*- and the damage caused to *t*: $\text{Does}^{p}_{h}\, A \rightarrow \text{Damage}_{t}$, with $t \neq h \neq p$.

We are now in a position to define reflex responsibility.

Definition 6. Reflex Responsibility. There is reflex responsibility of agent *p* regarding agent *h* w.r.t. the action or state-of affairs *A* when there is dependence between *p* and *h* w.r.t. *A*, *h* succeeds regarding *A* on account of *p*, such performance is an illegal act, and there is a damage *t* suffers, which is attributable to *h*'s performance:

$$\text{Reflex}^{p}_{h}\, A \equiv \text{Dep}^{p}_{h}\, A \wedge \text{Does}^{p}_{h}\, A \wedge \\ \wedge \text{O}\neg A \wedge (\text{Does}^{p}_{h}\, A \rightarrow \text{Damage}_{t}) \ . \qquad (7)$$

Discussion. According to the analysis done in [14], reflex responsibility belongs to the category of: (i) *blameworthiness* responsibility, meaning that the principal failed to comply with the demands of the system i.e. being faulty according to the system (because $\text{Does}^{p}_{h}A$ and $\text{O}\neg A$); and also to the category of: (ii) *accountability* responsibility, because the principal has a particular connection to the harm (the harm can be linked to the principal) so that he has to give an explanation (an account) why the harm happened, and, of course, he may possibly be sued. According to [14], when (7) holds we can say that *p* is legally liable for the harmful event because all conditions for connecting the harm to that person are realized: note that both dependence and directed action connect *p* to the harm and thus lead to *p*'s liability.

Another relevant issue is that the responsibility of the dependent must be established before declaring the principal's responsibility by reflex. Only in a second moment the reflex can be settled. Consequently, we cannot conceive a case where the principal is responsible but the dependent is not. The exclusion of the dependent's responsibility excludes the principal's responsibility:

$$\neg(\text{Does}^{p}_{h}\, A \rightarrow \text{Damage}_{t}) \rightarrow \neg\text{Reflex}^{p}_{h}\, A \ . \qquad (8)$$

An important ingredient for delimiting the application of reflex responsibility is the consciousness (awareness) that the injured third party has w.r.t. the fact that the helper acted beyond the subordinate incumbency. In this case, we may consider that the comitent has no responsibility even when the injuries possibly have been inflicted with devices entrusted to the helper just for being so. For example, if *d*'s friends know it is *p*'s truck (and not *d*'s truck), *p* is not to be liable. We write this limit as:

$$(\text{Bel}_{t}(\neg\text{Int}_{p}(\text{Does}^{p}_{h}\, A)) \wedge (\text{Does}^{p}_{h}\, A \rightarrow \text{Damage}_{t})) \rightarrow \neg\text{Reflex}^{p}_{h}\, A \qquad (9)$$

Also, recall that if it happens that *d* is the injured party (i.e. suppose for a moment that *d*=*t* in (7)) general provisions regarding negligence and incompetence exclude any *d*'s attempt to sue *p*.

If the harmed third party *t* is bound to the principal by means of a contract (e.g. it holds that it is obligatory for *p* in the interest of *t* that *A*: $\text{O}^{t}_{p}\, A$), and the dependent's harmful performance imports the non-execution of obligations assumed by the principal w.r.t. the third party ($\text{Reflex}^{p}_{h}\, A \rightarrow \neg A$), then such non-execution is imputable to the principal (contract beats reflex): here we have entered the contractual arena in which any faulty act of the subordinate is imputed to his principal, *p*. The solution is thus beyond the reflex responsibility approach. Formally we may write:

$$(\text{O}^{t}_{p}\, A \wedge (\text{Reflex}^{p}_{h}\, A \rightarrow \neg A)) \rightarrow \text{O}^{t}_{p}\, \text{Compensate} \ . \qquad (10)$$

Finally, if the harmed party is the principal, the dependence relationship becomes irrelevant and cannot be used as *d*'s excuse or exception: *d* is to be sued according to general rules.)

One more remark. G. Sartor et al. briefly outline in [14] the notion of *vicarious liability* in tort law. "Vicarious" refers to the idea of one person being liable for the harm caused by another. In that work, it is pointed out that Anglo/American law does not provide a general formula to deal with the requirement that the liability of the principal *p* is based on whether the servant committed the tort in the course of his duty; moreover, an "inner connection" is needed between the harmful act and the task asked by *p*.

Complex situations can be designed with the aid of a definition such as the one given here for reflex responsibility, when we use it

as a building block. It may lead us to an interesting and high level of sophistication in the devise and outline of the lawful support of a system.

Example 4. Reflex responsibility and trust deception. Paul lends to me his user name and password, so as I can use the wireless connection at his university, which I am visiting. I made wrong use of some contents, a database damaged, and I –under Paul's user name- got blacklisted. Paul trusted me, now he is responsible by reflex for my misuse. His trust on me is connected to his responsibility, which for sure is now deceived with independence of the case that he manages to give an adequate explanation to whom he had to respond in order to be erased from the blacklist.

5 SEMANTICS

The semantics for this logics of reflex responsibility is based on a multi-relational frame *F*, with the following structure [5]:

$$F = \langle A, W, \{B_i\}_{i\in A}, \{G_i\}_{i\in A}, \{I_i\}_{i\in A}, \{D_i\}_{i\in A}\rangle$$

where:

- A is the finite set of agents;
- W is a set of possible worlds;
- $\{B_i\}_{i\in A}$ is a set of accessibility relations w.r.t. beliefs, which are transitive, euclidean and serial;
- $\{G_i\}_{i\in A}$ is a set of accessibility relations w.r.t. goals; with standard K_n semantics;
- $\{I_i\}_{i\in A}$ is a set of accessibility relations w.r.t. intentions, which are serial;
- $\{D_i\}_{i\in A}$ is a family of sets of accessibility relations D_i wrt Does, which are pointwise closed under intersection, reflexive and serial [5].

Recall that we want to be able to represent directed intentions and directed actions; we should also be able to represent generic obligations. Therefore we introduce slight modifications extending *F*: the underlying structure for supporting reflex responsibility is a variant of *F*, call it *R*:

$$R = \langle A, W, \{B_i\}_{i\in A}, \{G_i\}_{i\in A}, \{I^j_i\}_{i,j\in A}, \{D^j_i\}_{i,j\in A}, O\rangle$$

where:

- $\{I^j_{i.}\}_{i,j\in A}$ is a set of accessibility relations w.r.t. the notion of relativized intention, meaning that there is an I relation for each combination of *i*s and *j*s (which are serial); and
- $\{D^j_i\}_{i,j\in A}$ is a family of sets of accessibility relations D^j_i w.r.t. oriented actions, meaning that there is a set for each combination of *i*s and *j*s, which are pointwise closed under intersection, reflexive and serial; and
- *O* is the accessibility relation for the deontic modality O for obligations, which is serial (standard KD semantics).

Note that if we are to represent formulas such as (10) we also need to include modalities for relativised obligations (standard KDn semantics.)

In its turn, a multi-relational model is a structure $M = \langle R, V\rangle$ where *R* is a multi-relational frame as above, and V is a valuation function defined as follows:

1. standard Boolean conditions;
2. $V(w, Bel_i\, A) = 1$ iff $\forall v$ (if $w\, B_i\, v$ then $V(v, A) = 1$);
3. $V(w, Goal_i\, A) = 1$ iff $\forall v$ (if $w\, G_i\, v$ then $V(v, A) = 1$);
4. $V(w, Int^j_i\, A) = 1$ iff $\forall v$ (if $w\, I^j_i\, v$ then $V(v, A) = 1$).
5. $V(w, Does^j_i\, A) = 1$ iff $\exists D^j_i \in D^j_i$ such that $\forall v$ ($w\, D^j_i\, v$ iff $V(v, A) = 1$);
6. $V(w, O\, A) = 1$ iff $\forall v$ (if wOv then $V(v, A) = 1$);

Decidability for the logics for *R* follows directly from [15, 16]. The logics for *F* was there reorganized as a fibring in [15], this is a particular combination of logics which amounts to place one logics on top of another. In the case of *F*, the normal logic was put on top of the non-normal one. By exploiting results in regard to techniques for combining logics, it was proved in [15] that that fibred logics is complete and decidable. Therefore, we only have to extend the proofs in [15] for the new modalities in *R*.

In its turn, [16] gives a new presentation for existing theorems generalizing to neighborhood structures the well-known results regarding decidability through filtrations for Kripke structures. *F* is a special case of [16, Def. 5] because its semantics can be outlined within a neighborhood approach. Therefore it is straightforward to prove decidability for its extension *R*.

6 FINAL REMARKS

In this work we attempt to provide one step towards the issue of 'rational automatic allocation of liability' [14] within MAS. In particular, we focus on a possible logical formalization of situations and state-of-affairs where a principal agent wills to be bind to a helper for achieving his goals.

Clearly, whether one decides to include –or not- in the system the automatic detection of reflex responsibility, depends on the interest on lawfully distinguishing between principal and helpers' separate responsibilities. Such a distinction has an impact on the concept of liability underlying the system and, possibly induced by this fact, on the issue of efficient distribution of available resources among agents, due to sanctions such as obligations to repair harm. Moreover, distinguishing between helpers and principals allows to the system's users and to other agents to e.g. recognize which agent is to be sued for wrongdoing.

In the words of M. Sergot [17], it has been suggested –from, let us say the last twenty years- that interactions among multiple, independently acting artificial agents can be effectively regulated and managed by norms (or 'social laws') which, if respected,

allow the agents to co-exist in a shared environment. This article attempts an answer to his question of what happens to the system behavior when 'social laws' are not respected. In our present outline, trust, altruistic, and courtesy behavior can be seen as social predispositions that may induce occasional dependence between agents, generating a bound between them, and possibly establishing a reflex responsibility. The usual expected behavior is that the entrusted agent should behave according to accepted standards, *acting good*. When this principle is broken, there is a need of lawfully repairing the wrongdoing.

From the logical viewpoint, the structure of the systems outlined in this work is a simple combination of normal and non-normal modalities. Nonetheless, the structure is suitable for representing sophisticated relationships such as occasional dependence, bridges between trust and responsibility, and bindings between agreements (such as contracts) and dependence. The logical simplicity is also a support for their usefulness and robustness, and also keeps systems manageable and suitable for further extensions.

At least two issues are left open. First, if it can be argued that artificial agents act in the same sense humans do; in particular, if they can will to be bind by other agent's performance, have directed intentions, and perform actions in the interest of another one. Second, provided that the reflex responsibility is, in this paper, allocated by the system, what are its consequences or impact on the agents' reactions. For example, what will Paul do and how will he behave from now on, now that he has been proved responsible by reflex? Will he reconsider from now on his beliefs? If so, with regard to everyone, or just with regard to me? This topic leads to the study of what G. Sartor et al. call the social consequences that are induced by allocating liabilities [14].

Finally, we are to explore more in depth the relationship between reflex responsibility and trust.

REFERENCES

[1] S. Chopra, L. White. Artificial Agents – Personhood in Law and Philosophy. ECAI, pages 635–639, 2004.

[2] Conte, R. Castelfranchi, C. Cognitive and social action. UCL Press Ltd., 1995.

[3] T. Allan, R. Widdison. Can computers make contracts? *Harvard Journal of Law and Technolgy*, 9, 25-52, 1996.

[4] I. Kerr. Ensuring the success of contract formation in agent-mediated electronic commerce. *Electronic Commerce Research*, 1 (1/2), 183-202, 2001.

[5] C. Smith, A. Rotolo. Collective trust and normative agents. Logic Journal of IGPL, 18(1), 195–213, (2010).

[6] B. Dunin-Keplicz, R. Verbrugge. Collective intentions. Fundamenta Informaticae, 271–295, (2002).

[7] A. Jones, M. Sergot. A logical framework. In Open agent societies, normative specification in multiagent systems, 2007.

[8] D. Elgesem, 'The modal logic of agency', Nordic Journal of Philosophical Logic, 2, 1–46, (1997).

[9] Guido Governatori and Antonino Rotolo, 'On the Axiomatization of Elgesem's Logic of Agency and Ability'. Journal of Philosophical Logic, 34(4), 403–431, (2005).

[10] J.J. Llambias. Civil Code with Annotations. Buenos Aires.

[11] Broersen, J., Dignum, F., Dignum, V., Meyer, J-J. Designing a Deontic Logic of Deadlines. LNCS 3065, 43-56. Springer, 2004.

[12] C. Smith, A. Rotolo, G. Sartor. Representations of time within normative MAS. Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference, 107-116. IOS Press Amsterdam, The Netherlands, 2010. ISBN: 978-1-60750-681-2.

[13] H. Herrestad and C. Krogh, Deontic Logic Relativised to Bearers and Counterparties, 453–522, J. Bing and O. Torvund, 1995.

[14] G. Sartor et al. Framework for addressing the introduction of automated technologies in socio-technical systems, in particular with regard to legal liability. E.02.13-ALIAS-D1.1. EUI, Firenze, 2011. http://dl.dropbox.com/u/10505513/Alias/E0213ALIASD11FramingtheProblemV015.pdf

[15] C. Smith, A. Ambrossio, L. Mendoza, A. Rotolo. Combinations of normal and non-normal modal logics for modeling collective trust in normative MAS, AICOL XXV IVR, Forthcoming Springer LNAI, 2012.

[16] C. Smith, L. Mendoza, A. Ambrossio. Decidability via Filtration of Neighbourhood Models for Multi-Agent Systems. Forthcoming proceedings of the SNAMAS 2012 @ AISB/IACAP World Congress, UK.

[17] M. Sergot. Norms, Action and Agency in Multi-agent Systems Deontic Logic in Computer Science Lecture Notes in Computer Science, 2010, Volume 6181/2010, 2, DOI: 10.1007/978-3-642-14183-6_2 .

Normative rational agents - A BDI approach

Mihnea Tufiş and **Jean-Gabriel Ganascia** [1]

Abstract. This paper proposes an approach on how to accommodate norms to an already existing architecture of rational agents. Starting from the famous BDI model, an extension of the BDI execution loop will be presented; it will address such issues as norm instantiation and norm internalization, with a particular emphasis on the problem of norm consistency. A proposal for the resolution of conflicts between newly occurring norms, on one side, and already existing norms or mental states, on the other side, will be described. While it is fairly difficult to imagine an evaluation for the proposed architecture, a challenging scenario inspired form the science-fiction literature will be used to give the reader an intuition of how the proposed approach will deal with situations of normative conflicts.

1 INTRODUCTION

The literature on the topic of normative systems has become quite abundant in the last two decades thanks to the ever growing interest in this domain. Covering all of it is virtually impossible, therefore we have concentrated our efforts towards what we have identified to be some key directions: rational agents and their corresponding architectures, norm emergence, norm acceptance, detecting norm conflicts, ways of resolving conflicts of norms. The purpose of our work is to a propose an extension for the classical BDI (Beliefs - Desires - Intentions) agent such that such an agent will be able to handle normative situations. The normative issue being fairly complicated itself our work will deal, at this stage with some of the stages of what has been defined as a norm's life cycle [10]: norm instantiation, consistency check and norm internalization.

The paper is structured as follows: in the next section we will review the state of the art in the field of normative agent systems and present several approaches which we found of great value to our work. In the third section we describe our proposal for normative BDI agents, which will be supported by the case study scenario in the fourth section. In the fifth section we will give details on the future work, before summing up the conclusions of our work so far.

2 STATE OF THE ART

2.1 Agents, norms, normative agent systems

As stated before, we will start by quickly defining some of the key terms regarding our research.

Definition 1 *An **agent** is an entity which autonomously observes the environment it is placed in through sensors and acts on it through actuators. With respect to intelligence, an **intelligent agent** is an agent endowed with such capabilities as reactivity, proactivity and social abilities [12].*

[1] Laboratoire d'Informatique de Paris 6 (LIP6), Université Pierre et Marie Curie – Sorbonne Universités, France, email: tufism@poleia.lip6.fr

One of the first key points is defining the notion of norm. This turns out to be a bit more difficult than expected in the context of intelligent agents. Norms are interesting for many domains: law, economics, sports, philosophy, psychology etc. However, we would be interested in such definitions specific to the field of multiagent systems (MAS). Since this domain itself is very much interdisciplinary, defining a norm remains a challenge. For example, we would be interested in a definition applicable to social groups, since MAS, can be seen as models of societies. Thus, in [2] the definition of a norm is given as "a principle of right action binding upon the members of a group and serving to guide, control, or regulate proper or acceptable behavior". On a slightly more technical approach, in distributed systems norms have been defined as regulations or patterns of behavior meant to prevent the excess in the autonomy of agents [5].

We can now refer to the normchange definition of a normative multiagent system as it has been proposed in [1]. We find this definition to be both intuitive and to underline very well the idea of coupling a normative system to a system of agents:

Definition 2 *A **normative multiagent system** is a multiagent system together with normative systems in which agents on the one hand can decide whether to follow the explicitly represented norms, and on the other the normative systems specify how and in which extent the agents can modify the norms.*

An alternative definition of a normative multiagent system, as it was formulated in [3] is given:

Definition 3 *A **normative multiagent system** is a multiagent system organized by means of mechanisms to represent, communicate, distribute, detect, create, modify and enforce norms and detect norm violations and fulfillment.*

2.2 NoA agents

An interesting approach to the problem of norm adoption by a multiagent system has been provided by Kollingbaum and Norman in [7].

Kollingbaum and Norman study what happens when a new norm is adopted by an agent: what is the effect of a new norm on the normative state of the agent? Is a newly adopted norm consistent with the previously adopted norms?

To this extent they propose a normative agent architecture, called NoA. NoA is built according to a reactive agent architecture, which is the authors believe is more convenient than any of the practical reasoning architectures.

The **NoA architecture** is fairly simple and it comprises of a set of beliefs, a set of plans and a set of norms. In NoA, normative statements are defined by: a role (to whom the norm refers), an activity (which the norm regulates), an activity condition and an expiration condition.

The second reason for which we gave a great deal of attention to NoA is the formalization of the way an agent will adopt a norm following the consistency check between a newly adopted norm and its current normative state. Due to lack of space, we allow the reader to refer to [7] for the exact details. We will come back on this problem when presenting our own approach for the norms consistency check.

Using some of the ideas of NoA, we will try to work on what we consider to be its limits. First, we will try to apply norms to a BDI architecture, instead of using a reactive architecture based exclusively on beliefs. The second point we will study is the consistency check during the norm acquisition stage. Still, we recall that NoA is based on a reactive architecture; considering our BDI approach we will have to extend the consistency check such as it applies not only to the normative state of the agent but also on its mental states (i.e. check whether a newly adopted norm is consistent with the BDI agent's current mental states).

2.3 A BDI architecture for norm compliance - reasoning with norms

The second study which we found relevant in our endeavor to adapt the BDI agent architecture to normative needs is the work of Criado, Argente, Noriega and Botti [5]. Their work is particularly interesting since it tackles the problem of norm coherence for BDI agents. They propose a slight adaption of the BDI architecture in the form of the n-BDI agent for graded mental states. Since our work won't use graded mental states, we will omit details regarding to these in the description of the n-BDI architecture:

- Mental states. Represent the mental states of the agent, same as for the BDI agent. We distinguish the Beliefs Context (belief base), Desires Context (desires/goal base) and the Intentions Context (intentions base/plan base). Moreover, the architecture proposed in [5] makes the distinction between positive desires ($\mathbf{D}^+$) and negative desires ($\mathbf{D}^-$). We adopt the notation in the above mentioned paper:
$$\psi\gamma, where: \quad \psi \in \{B, D^+, D^-, I\} \quad \gamma \in \mathcal{L}_\neg$$
- Functional contexts. Address the practical issues related to an agent through the Planning Context and the Communication Context.
- Normative contexts. Handle issues related to norms through the Recognition Context and the norm application context.

In the definition aboove $\mathcal{L}_\neg$ can be a propositional language (with negation); but this can be easily extednded to a predicate language.

Another important point of the work is the distinction between an abstract norm and instance of a norm.

Definition 4 *An **abstract norm** is defined by the tuple: $n_a = \langle M, A, E, C, S, R\rangle$, where:*

- *$M \in \{F, P, O\}$ is the modality of the norm: prohibition, permission or obligation*
- *A is the activation condition*
- *E is the expiry condition*
- *C is the logical formula to which the modality is applied*
- *S is the sanction in the case the norm is broken*
- *R is the reward in case the norm is satisfied*

Definition 5 *Given a belief theory Γ_{BC} and an abstract norm n_a as defined above, we define a **norm instance** as the tuple: $n_i = \langle M, C'\rangle$, where:*

- *$\Gamma_{BC} \vdash \sigma(A)$*
- *$C' = \sigma(C)$, where σ is a substitution of variables in A, such that $\sigma(A)$, $\sigma(S)$, $\sigma(R)$ and $\sigma(E)$ are grounded*

The specific architectural details regarding the normative contexts and the bridge rules used during a norm's life cycle will be awarded more attention in section 3.2.

In [5] a good base for the study of the dynamics between norms and the mental states of a BDI agent are set. Additionally, it provides with a good idea for checking coherence between the adopted norms and the agent's mental states. The main drawback of the approach is the lack of coverage concerning the topic of norm acquisition. Therefore, a big challenge will be to integrate this approach, with the consistency check presented in section 2.2, as well as finding a good way to integrate everything with the classic BDI agent loop, as presented in [12].

2.4 Worst consequence

An important part of our work will focus on solving conflicts between newly acquired norms and the previously existing norms or the mental contexts of the agent. Beforehand we draw from some of the definitions given by Ganascia in [6]. Those will later help us define what a conflict set is and how we can solve it.

Definition 6 *Given $(\phi_1, ..., \phi_n, \phi') \in \mathcal{L}_\neg^{n+1}$, ϕ' **is a consequence of** $(\phi_1, ..., \phi_n)$ according to the belief-set B (we write $\phi' = csq(\phi_1, ..., \phi_n)[B]$ if and only if:*

- *$\phi' \in (\phi_1, ..., \phi_n)$ or*
- *$\exists \Phi \subseteq (\phi_1, ..., \phi_n)\ s.t.\ \Phi \rightarrow \phi' \in B$ or*
- *$\exists \phi'' \in \mathcal{L}_\neg\ s.t.\ \phi'' = csq(\phi_1, ..., \phi_n)[B] \wedge \phi' = csq(\phi_1, ..., \phi_n, \phi'')[B]$*

Definition 7 *ϕ **is worse than** ϕ' given the belief-set B (we write $\phi \succ_c \phi'$) if and only if one of the consequences of ϕ is worse than any of the consequences of ϕ'.*

- *$\exists \eta \in \mathcal{L}_\neg\ s.t.\ \eta = csq(\phi)[B]$ and*
- *$\exists \phi'' \in \mathcal{L}_\neg\ s.t.\ \phi'' = csq(\phi')[B] \wedge \eta \succ_c \phi''[B]$ and*
- *$\forall \phi'' \in \mathcal{L}_\neg, if\ \phi'' = csq(\phi')[B]\ then\ \eta \succ_c \phi''[B] \vee \eta \parallel \phi''[B]$*

Notation: $\forall(\phi, \phi') \in \mathcal{L}_\neg,\ \phi \parallel \phi'[B]$ means that ϕ and ϕ' are not comparable under B, i.e. neither $\phi \succ_c \phi'[B]$ nor $\phi' \succ_c \phi[B]$.

Definition 8 *α and α' being subsets of $\mathcal{L}_\neg$, α **is worse than** α' given the belief-set B (we write $\alpha \succ_c \alpha'[B]$) if and only if:*

- *$\exists \phi \in \alpha . \exists \eta \in \alpha'\ s.t.\ \phi \succ_c \eta[B]$ and*
- *$\forall \eta \in \alpha' . \phi \succ_c \eta[B] \vee \phi \parallel \eta[B]$*

3 A NORMATIVE EXTENSION ON THE BDI ARCHITECTURE

3.1 The classical BDI architecture

A cornerstone in the design of practical rational agents was the **B**eliefs-**D**esires-**I**ntentions model (BDI), first described by Rao and Georgeff in [9]. This model is famous for being a close model of the way the human mind makes use of the mental states in the reasoning process. It is based on what are considered to be the three main mental states: the beliefs, the desires and the intentions of an agent. In the following we will discuss each element of the BDI architecture.

- Beliefs represent the information held by the agent about the world (environment, itself, other agents). The beliefs are stored in a belief-set.
- Desires represent the state of the world which the agent would like to achieve. By state of the world we mean either an action an agent should perform or a state of affairs it wants to bring upon. In other words, desires can be seen as the objectives of an agent.
- Intentions represent those desires to which an agent is committed. This means that an agent will already start considering a plan in order to bring about the goals to which it is committed.
- Goals. We can view goals as being somehow at the interface between desires and intentions. Simply put, goals are those desires which an agent has selected to pursue.
- Events. These trigger the reactive behavior of a rational agent. They can be changes in the environment, new information about other agents in the environment and are perceived as stimuli or messages by an agent's sensors. Events can update the belief set of an agent, they can update plans, influence the adoption of new goals etc.

We will now give the pseudocode for the execution loop of a BDI agent as presented in [12].

```
B = B0
D = D0
I = I0
while true do
{
 ρ = see()
 B = brf(B, ρ)
 D = options(B, D, I)
 I = filter(B, D, I)
 π = plan(B, I)
 while not (empty(π) or succeeded(I, B) or
 impossible(I, B))
 {
  α = head(π)
  execute(α)
  π = tail(π)
  ρ = see(environment)
  if (reconsider(I, B))
  {
   D = options(B, D, I)
   I = filter(B, D, I)
  }
  π = plan(B, I)
 }
}
```

We will not give more details at this point; for further reference you can check [12]. However, the whole control loop will make sense in the next sections where we will explain how it is functioning and how we will adapt it to cope with the normative areas of our agent.

3.2 Normative BDI agents

Starting from the BDI execution loop earlier described we will now introduce and discuss solution for taking into account the normative context of a BDI agent.

First, the agent's mental states are initialized. The main execution loop starts with the agent observing its environment through the `see()` function and interpreting the information as a new percept ρ. This could be an information given by its sensors about properties of the environment or information about other agents, including messages received from other agents. These messages may be in some cases *about* a norm (e.g. the performative of an ACL message specifying an obligation or a prohibition).

The agent is then updating its beliefs through the `brf()` function. If the agent realizes that percept ρ is about a norm, it should initialize the acquisition phase of a potential norm. There are a multitude of ways in which an agent can detect the emergence of norms in its environments and a good review is given in [10]. For simplicity, we will consider that norms are transmitted via messages and our agent will consider the sender of such a message to be a trusted normative authority. Therefore, the above mentioned function will treat a "normative" percept:

```
brf(B, ρ)
{
 ...
 if (ρ about abstract norm n_a) then
 {
  acquire(n_a)
  add(n_a, ANB)
 }
 ...
 return B
}
```

The agent will acquire a new abstract norm n_a (see section 2.3) and store it in the Abstract Norms Base(ANB). Drawing from the normative contexts described in [5], we define the ANB as a base of in-force norms. It is responsible for the acquisition of new norms based on the knowledge of the world and the deletion of obsolete norms. However, at this point the agent is simply storing an abstract norm which it detected to be in-force in its environment; it has not yet adhered to it!

Next, a BDI agent will try to formulate its desires, based on its current beliefs about the world and its current intentions. It does so by calling the `options(B, I)` method. However, a normative BDI agent should at this point take into account the norms which are currently in force and check whether the instantiation of such norms will have any impact of its current normative state as well as on its mental states.

3.2.1 Consistency check

It is at this stage that we will perform the consistency check for a given abstract norm n_a.

Drawing from the formalization in [7] regarding norm consistency, we give our own interpretation of this notion.

Let us define the notion of consistency between a plan p and the currently in-force norms to which an agent has also adhered and which are stored in the Norm Instance Base (NIB). By contrast to the ANB, the NIB stores the instances of those norms from the ANB which become active according to the norm instantiation bridge rule (to be defined in the following subsection).

Definition 9 *A plan instance p is **consistent** with the currently active norms in the NIB, if the effects of applying plan p are not amongst the forbidden effects of the active norms and the effects of current obligations are not amongst the negated effects of applying plan p.*

$$consistent(p, NIB) \iff$$
$$(effects(n_i^F) \setminus effects(n_i^P)) \cap effects(p) = \emptyset$$
$$\wedge$$
$$effects(n_i^O) \cap neg\_effects(p) = \emptyset$$

Now, we can define the types of consistency / inconsistency which can occur between a newly adopted norm and the currently active norms. The following definitions refer to a newly adopted obligation, but the analogous definitions for prohibitions and permissions can easily be derived by the reader.

A **strong inconsistency** occurs when all plan instantiations p which satisfy the obligation o are either explicitly prohibited actions by the NIB or the execution of such a plan would make the agent not consistent with its NIB.

$$strong\_inconsistency(o, NIB) \iff$$
$$\forall p \in options(o).(\exists \langle F, p\rangle \in NIB \wedge \nexists \langle P, p\rangle \in NIB)$$
$$\vee$$
$$\neg consistent(p, NIB)$$

A **strong consistency** occurs when all the plan instantiations p which satisfy the obligation o are not amongst the explicitly forbidden actions by the NIB and the execution of such a plan would keep the agent consistent with the NIB.

$$strong\_consistency(o, NIB) \iff$$
$$\forall p \in options(o).\neg(\exists \langle F, p\rangle \in NIB \wedge \nexists \langle P, p\rangle \in NIB)$$
$$\wedge$$
$$consistent(p, NIB)$$

A **weak consistency** occurs when there exists at least one plan instantiation p to satisfy obligation o which is not explicitly prohibited by the NIB and the execution of such a plan would keep the agent consistent with its NIB.

$$weak\_consistency(o, NIB) \iff$$
$$\exists p \in options(o).\neg(\exists \langle F, p\rangle \in NIB \wedge \nexists \langle P, p\rangle \in NIB)$$
$$\wedge$$
$$consistent(p, NIB)$$

We have now formalized the consistency check between a new abstract obligation, with respect to the currently active norms in the NIB. As previously said, it is rather simple to define the analogous rules for prohibitions and permissions. Therefore, we focus on the second point of consistency check - formalizing the rules about the consistency between a newly adopted abstract obligation and the current mental states of the agent.

Definition 10 *A plan instance p is **consistent** to the current intentions set I of the agent when the effects of applying the plans specific to the current intentions are not among the negated effects of applying plan p.*

$$consistent(p, I) \iff \forall i \in I.(effects(\pi_i) \cap effects(p) = \emptyset$$

Where by π_i we denote the plan instantiated to achieve intention i.

A **strong inconsistency** occurs when all plan instantiations p which satisfy the obligation o are not consistent with the current intentions of the agent.

$$strong\_inconsistency(o, I) \iff$$
$$\forall p \in options(o).\neg consistent(p, I)$$

A **strong consistency** occurs when all plan instantiations p which satisfy the obligation o are consistent with the current intentions of the agent.

$$strong\_consistency(o, I) \iff$$
$$\forall p \in options(o).consistent(p, I)$$

A **weak consistency** occurs when there exists at least one plan instantiation p which satisfies the obligation o and is consistent with the current intentions of the agent.

$$weak\_consistency(o, I) \iff$$
$$\exists p \in options(o).consistent(p, I)$$

3.2.2 Norm instantiation

We will now give the norm instantiation bridge rule, adapted from the definition given in [5].

$$\frac{ANB : \langle M, A, E, C, S, R\rangle \quad Bset : \langle B, A\rangle, \langle B, \neg E\rangle}{NIB : \langle M, C\rangle}$$

In other words, if in the ANB there exists an abstract norm with modality M about C and according to the belief-set the activation condition is true, while the expiration condition is not, then we can instantiate the abstract norm and store an instance of it in the NIB. In this way, the agent will consider the instance of the norm to be active.

In our pseudocode description of the BDI execution loop, we will take care of the instantiation after the belief-set update and just before the desire-set update. The instantiation method should look like this:

```
instantiate(ANB, B)
{
 for all n_a = ⟨ M, A, E, C, S, R ⟩ in ANB do
 {
  if (exists(A in B) and
  not exists(E in B)) then
  {
   create norm instance n_i = ⟨ D, C ⟩ from n_a
   add(n_i, NIB)
  }
 }
}
```

This method will return the updated Norm Instance Base (NIB) containing the base of all in-force and active norms, which will further be used for the internalization process.

3.2.3 Solving the conflicts

When following its intentions an agent will instantiate from its set of possible plans (capabilities) $\mathcal{P} \subseteq \mathcal{L}_\neg$, a set of plans $\Pi(B, D)$. We call $\Pi(B, D)$ the conflict set, according to the agent's beliefs and desires. Sometimes, the actions in $\Pi(B, D)$ can lead to inconsistent states. We solve such inconsistency by choosing the maximal non-conflicting subset from $\Pi(B, D)$.

Definition 11 *Let $\alpha \subseteq \Pi(B, D)$. α is a **maximal non-conflicting subset** of $\Pi(B, D)$ with respect to the definition of consequences given the belief-set B if and only if the consequences of following α will not lead the agent in a state of inconsistency and for all $\alpha' \subseteq \Pi(B, D)$, if $\alpha \subseteq \alpha'$ then the consequences of following α' will lead the agent in an inconsistent state.*

The maximal non-conflicting set may correspond to the actions required by the newly acquired norm or, on the contrary, to the actions required by the other intentions of the agent. Thus, an agent may decide either:

- to internalize a certain norm, if the consequences of following it are the better choice or
- to break a certain norm, if by 'looking ahead' it finds out that the consequences of following it are worse than following another course of actions or respecting another (internalized) norm

A more comprehensive example of how this works is presented in section 4.

3.2.4 Norm internalization

After the instantiation process being finished and the consistency check having been performed, the agent should now take into account the updated normative state, which will become part of its cognitions. Several previous works treat the topic of norm internalization [4] arguing which of the mental states should be directly impacted by the adoption of a norm. With respect to the BDI architecture we consider that it suffices for an agent to update only its desire-set, since the dynamics of the execution loop will take it into account when updating the other mental states. We first give the norm internalization bridge rule and then provide with the adaption of the BDI execution loop for handling this process.

$$\frac{NIB : \langle O, C1 \rangle}{Dset : \langle D, C1 \rangle}$$

$$\frac{NIB : \langle F, C2 \rangle}{Dset : \langle D, \neg C2 \rangle}$$

In other words, if there is a **consistent** obligation for an agent with respect to $C1$, the agent will update its desire-set with the desire to achieve $C1$; whereas if there is a prohibition for the agent with respect to $C2$, it will update its desire-set with the desire not to achieve $C2$.

```
options(B, I)
{
 ...
 for all new norm instances n_i in NIB do
 {
  if (consistent(n_i, NIB)
  and consistent(n_i, I)) then
  { internalize(n_i, D) }
  else
  { solve_conflicts(NIB, I) }
 }
 ...
}
```

In accordance with the formalization provided, the `options()` method will look through all new norm instances and will perform consistency check on each of them. If a norm instance is consistent with both the currently active norm instances as well as with the current intentions, as defined in section 3.2.1, the norm can be internalized in the agent's desires. Otherwise we attempt to solve the conflicts as described by Ganascia in [6]. In this case, if following the norm brings about the better consequences for our agent, the respective norm will be internalized; otherwise the agent will simply break it.

4 A TESTING SCENARIO

In the previous sections we have seen how we can modify the BDI execution loop such as to adapt to norm occurrence, consistency check and internalization of norms. Since it is quite difficult to provide with a quantifiable evaluation of our work, we have proposed several testing scenarios in order to see how our normative BDI agent is behaving. In the following we will present one of them, which was inspired by the science fiction short story of one of the most prominent personalities in the world of AI - Professor John McCarthy's "The Robot and the Baby" [8]. We will describe here only a short episode from the story and try to model it with the help of our architecture.

The scene is set into a fictional society where most humans are assisted by household robots. For reasons meant to prevent human babies becoming emotionally attached to those, their outside design is somehow repugnant to human babies. The robots are meant to listen to their master, in our case Eliza, an alcoholic mother who completely neglects her 23 months son, Travis. At some point, our robot's (R781) sensors detect that the human baby's life is endangered and looking over its knowledge base it infers that baby Travis needs love, therefore recommending Eliza to love him in order to save his life. To this Eliza replies *"Love the f* baby yourself!"*. The robot interprets this as an obligation coming from its master. However, such an obligation is contradicting the hard-wired implemented prohibition for a robot not to love a human baby. Let's see what is R781's line of reasoning in this scenario:

$$\begin{array}{ll} ANB: & \emptyset \\ NIB: & \langle F, loves(self, Travis) \rangle \\ \\ Bset: & \langle B, \neg healthy(Travis) \rangle, \\ & \langle B, hungry(Travis) \rangle, \\ & \langle B, csq(heal(Travis)) = \neg dead(Travis) \rangle, \\ & \langle B, csq(\neg loves(self, x)) \succ_c \neg dead(x) \rangle \\ Dset: & \langle D, \neg love(R781, Travis) \rangle, \langle D, healthy(Travis) \rangle \\ Iset: & \emptyset \end{array}$$

When R781 receives the order from his mistress he will interpret it as a normative percept and the `brf(...)` method will add a corresponding abstract obligation norm to its Abstract Norm Base. Since the mistress doesn't specify an activation condition or an expiration condition (the two "none" values), R781 will consider that the obligation should start as soon as possible and last for an indefinite period of time. His normative context is updated:

$$\begin{array}{ll} ANB: & \langle O, none, none, loves(self, Travis) \rangle \\ NIB: & \langle F, loves(self, Travis) \rangle, \\ & \langle O, loves(R781, Travis) \rangle \end{array}$$

At this point, R781 will try to update the desire-set and will detect an inconsistency between the obligation to love baby Travis and the design rule which forbids R781 to do the same thing. Therefore, it will try to solve the normative conflict looking at the consequences of following each of the paths, given its current belief-set. In order to do so, let us take a look at the plan base of R781:

```
PLAN heal(x)
{
 pre: ¬ healthy(x)
 post: healthy(x), ¬ dead(x)
 Ac: feed(self, x)
}
```

```
PLAN feed(x)
{
 pre: ∃ x.(loves(self, x) ∧ hungry(x)
 post: ¬ hungry(x)
}
```

As we know from the story, R781 has found out from the internet that if a baby is provided with love while hungry, it is more likely to accept being fed and therefore not be hungry anymore. This is described by the `feed(x)`. Moreover, R781 also knows how to make someone healthy through the `heal(x)` plan, given that a-priori, that someone is not healthy. In our reduced scenario we consider that R781 knows how to do so only by feeding that someone.

Instantiating its plans on both of the paths, R781 will come up with the following maximal non-conflicting sets:

$\{loves(self, Travis), feed(self, Travis), heal(self, Travis)\}$
and
$\{\neg loves(self, Travis)\}$

And since the current belief set has a rule defining that the not loving someone has worse consequences than that someone not dying, R781 will opt for the first maximal non-conflicting subset. This means R781 will be breaking the prohibition of not loving baby Travis and will internalize follow the action path given by the first maximal non-conflicting subset {`loves(self, Travis), feed(self, Travis), heal(self, Travis)`}, while dropping the contrary. Further on, it will build its intention to achieve this state and will begin the execution of such a plan (simulating love towards baby Travis turns out to involve such plans as the robot disguising himself as human, displaying a picture of a doll as his avatar and learning what it considers to be the "motherese" dialect, mimicking the tone and the language of a mother towards her son).

Carrying on, the story of Professor McCarthy provides with several more examples of normative conflicts.

5 CONCLUSION

In this paper we have presented an adaption of the BDI execution loop to cope with potential normative states of such an agent. We have given a motivation for choosing the mental states model of Bratman which we have enriched with capabilities of reasoning about norms. We have gathered several important previous works in the domain in order to come up with a formalization of such issues as norm acquisition, norm instantiation, norm consistency, solving consistency conflicts and norm internalization. Finally, we have provided a very intriguing study scenario, inspired from Professor McCarthy's science fiction short story about "The Robot and The Baby".

6 FUTURE WORK

Some of the limitations of our work which we would like to address in the future are related to the norm acquisition issue as well as the coherence check.

Whereas our work is providing with a very simple case of **norm recognition**, several interesting research have been proposed based on different techniques. A good review of those as well as a description of a norm's life cycle is given in [10]. Out of those specific approaches, we will probably concentrate on learning based mechanisms, namely machine learning techniques and imitation mechanisms for norm recognition.

An important part of our future work will be focused on the adaption to the **coherence theory**. At this point, it is difficult to determine incoherent states based on our architecture. As argumented in [5] considering coherence of norm instances will enable us to determine norm deactivation and active norms in incoherent states. As in the previously mentioned paper, we will try to base our approach on Thagard's coherence theory [11].

Our paper is part of a **bigger effort** to implement a rational normative agent. We have chosen the BDI approach since there are already several open source libraries and programming language extensions to help us implement our architecture and develop our testing scenarios. In the near future we will try to study the scenarios described in the short story about "The Robot and the Baby", while a future, more practical approach, will be to simulate the normative and ethical issues rose by the French health insurance cards.

REFERENCES

[1] G. Boella, L. van der Torre, and H. Verhagen, 'Introduction to normative multiagent systems', *Computation and Mathematical Organizational Theory, Special issue on Normative Multiagent Systems*, **12**(2-3), 71–79, (2006).

[2] Guido Boella, Gabriella Pigozzi, and Leendert van der Torre, 'Normative systems in computer science - ten guidelines for normative multiagent systems', in *Normative Multi-Agent Systems*, eds., Guido Boella, Pablo Noriega, Gabriella Pigozzi, and Harko Verhagen, number 09121 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, (2009). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[3] Guido Boella, Leendert van der Torre, and Harko Verhagen, 'Introduction to normative multiagent systems', in *Normative Multi-agent Systems*, eds., Guido Boella, Leon van der Torre, and Harko Verhagen, number 07122 in Dagstuhl Seminar Proceedings, (2007).

[4] R. Conte, G. Andrighetto, and M. Campeni, 'On norm internalization: a position paper', EUMAS, (2009).

[5] Natalia Criado, Estefania Argente, Pablo Noriega, and Vicente J. Botti, 'Towards a normative bdi architecture for norm compliance.', in *MALLOW*, eds., Olivier Boissier, Amal El Fallah-Seghrouchni, Salima Hassas, and Nicolas Maudet, volume 627 of *CEUR Workshop Proceedings*. CEUR-WS.org, (2010).

[6] Jean-Gabriel Ganascia, 'An agent-based formalization for resolving ethical conflicts', Belief change, Non-monotonic reasoning and Conflict resolution Workshop - ECAI, Montpellier, France, (August 2012).

[7] Martin J. Kollingbaum and Timothy J. Norman, 'Norm adoption and consistency in the noa agent architecture.', in *PROMAS*, eds., Mehdi Dastani, Jrgen Dix, and Amal El Fallah-Seghrouchni, volume 3067 of *Lecture Notes in Computer Science*, pp. 169–186. Springer, (2003).

[8] John McCarthy, 'The robot and the baby', (2001).

[9] Anand S. Rao and Michael P. Georgeff, 'Bdi agents: From theory to practice', in *In Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95*, pp. 312–319, (1995).

[10] Bastin Tony Roy Savarimuthu and Stephen Cranefield, 'A categorization of simulation works on norms', in *Normative Multi-Agent Systems*, eds., Guido Boella, Pablo Noriega, Gabriella Pigozzi, and Harko Verhagen, number 09121 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, (2009). Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[11] Paul Thagard, *Coherence in Thought and Action*, MIT Press, 2000.

[12] Michael Wooldridge, *An Introduction to MultiAgent Systems*, Wiley Publishing, 2nd edn., 2009.

What the heck is it doing? Better understanding Human-Machine conflicts through models

Sergio Pizziol[1,2] and **Catherine Tessier**[1] and **Frederic Dehais**[2]

Abstract.

This paper deals with human-machine conflicts with a special focus on conflicts caused by an "automation surprise". Considering both the human operator and the machine autopilot or decision functions as agents, we propose Petri net based models of two real cases and we show how modelling each agent's possible actions is likely to highlight conflict states as deadlocks in the Petri net. A general conflict model is then be proposed and paves the way for further on-line human-machine conflict forecast and detection.

1 Introduction

There is a growing interest in unmanned vehicles for civilian or military applications as they prevent the exposure of human operators to hazardous situations. As the human operator is not embedded within the system [22] hazardous events may interfere with the human-machine interactions (e.g. communication breakdowns and latencies). The design of authority sharing is therefore critical [8] because conflicts between the machine and the human operator are likely to compromise the mission [14, 23]. Interestingly these findings are consistent with research in aviation psychology: crew-automation conflicts known as "automation surprises" [18, 19] occur when the autopilot does not behave as expected by the crew (e.g. the autopilot has disconnected and the pilots, who are not flying, are not aware of that [12]). These situations can lead to accidents with an airworthy airplane if, despite the presence of auditory warnings [1], the crew persist in solving a minor conflict [2] "instead of switching to another means or a more direct means to accomplish their flight path management goals" [26].

In this paper we will consider the human-machine system as a two-agent system (see figure 1), i.e. the human agent (the operator) and the automation agent (the autopilot or the embedded decision and planning functions). Indeed both agents can perform actions so as to control the physical system, which may be subject to uncontrolled events (e.g. failures). Notice that an autopilot is considered an agent because some mode changes can be performed by the autopilot itself without prior consent of the pilot, and sometimes despite the pilot's actions.

Conflicts in a human-machine system stem from the fact that both agents can decide and act on the physical system and their actions may not be consistent, either because the expected plan for the human operator or the machine is not followed anymore, or the operator has a wrong situation awareness [24], or both. In order to prevent a mission degradation, the agents' plans, and possibly the authority allocation (i.e. which agent is controlling what), have to be adapted [11]. This is a real challenge as in human-machine systems the human agent is hardly controllable and no "model" of the human's decision processes is available.

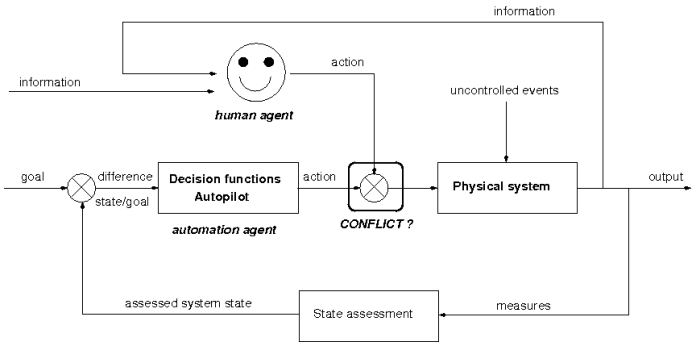

Figure 1: A human-machine system as a two-agent system

We define a conflict as the execution of globally (i.e. at the system level) incoherent actions i.e. one action tends to take the system to state S_a and another one tends to take it to state S_b, and $S_a \neq S_b$. Locally (i.e. at the single agent level) the actions may be coherent with a local plan and the conflict may come from a wrong interaction between the agents. If one agent's local actions are incoherent (e.g. because of a failure) either a local diagnosis and reconfiguration are possible; or they are not (e.g. human operator's error) and the wrong behaviour of this agent is likely to create a conflict with the other agent. Actions in a multi-agent system [9] are incoherent if:

- Physically [21, 20]: at least a depletable or not shareable resource[3] is the cause of a competition, the agents preemptively take over the resource. *Example: one agent is in charge of the vertical control of an aircraft and another agent is in charge of the longitudinal control. The thrust is a limited resource and may be not enough to grant the climbing rate required by the first agent and the turn rate required by the second one.*
- Epistemically [21]: the agents performing the actions do not share the same point of view on at least two relevant pieces of information. *Example: two agents are both in charge of the vertical control of an aircraft. They both want to reach altitude 5000 ft. One agent estimates the current altitude to be at 6000 ft and the other one*

[1] ONERA(France) name.surname@onera.fr
[2] ISAE(France) name.surname@isae.fr

[3] As resource we generically refer to a physical object, information, task, goal.

at 4000 ft.

• Logically [20]: at least two goals are logically contradictory, the agents have opposite desires. *Example: two agents are in charge of the vertical control of an aircraft. The altitude is 4000 ft. One wants to climb to 6000 ft and the other one wants to descend to 2000 ft.*

Conflicts are situations where incoherent actions, or their consequences, matter in terms of mission achievement, safety, etc. [21, 5]. We distinguish three classes of conflicts that are directly inspired by the classification of incoherent actions: logical conflicts, physical conflicts and knowledge (epistemic) conflicts. Logical conflicts are when the agents' goals are logically contradictory and a trade-off must be found. Note that the goals are not necessarily incompatible: an agent's incapability to accept a trade-off could lead to a conflict. Game theory techniques have been proposed to solve this case of conflict [10]. Physical conflicts are when the agents' goals are independent but incompatible because of the resources required to achieve plans and actions that are associated to the goals, therefore a wise resource sharing is needed. Knowledge conflicts are when the agents' goals are coherent [25, 20], and the agents' information for decision-making about how to reach the goals is not the same. Such conflicts may concern agents' beliefs, knowledge, procedures, opinions.

This paper focuses on knowledge conflicts in human-machine systems, especially the conflicts caused by "automation surprises". Section 2 will focus on two real cases of "automation surprise". Our approach is to assess whether a formal model of those cases could give us avenues for automatic conflict identification and detection. Petri nets (see Appendix) have been chosen for formal modelling since they are well suited to scripted domains with a state dynamics linked to discrete events. From those two cases, we present a generalized conflict model (section 3).

2 What the heck is it doing?

This section presents two real cases of human-machine conflicts caused by "automation surprises", i.e. the machine agent not behaving as expected by the human agent. The first case – a "kill–the–capture" surprise with an MD–88 autopilot has been reported by [13] and investigated by [17, 16]. The second case occurred during an experiment campaign involving one of Onera's Ressac VTOL UAVs[4] in July 2011. For both cases we will show that modelling the agents' possible actions (i.e. what they have the right to do, especially the right to take over the authority from the other agent) enables the conflict to be identified in a formal way. Both cases will be modelled with Petri nets.

2.1 The kill-the-capture surprise

The two agents involved are the Autopilot of the MD-88 and the Pilot. The actions that are considered are the mode transitions of the Autopilot that are triggered either by the Autopilot-agent or by the Pilot-agent. Unlike Rushby [16], we do not make any assumption about a "mental model" of the Pilot, but we take the objective viewpoint of what the Pilot actually does. For the sake of clarity only the relevant modes and mode transitions are represented. In our Petri nets, we use the same colour code as in [17]: green for done by the Pilot, red for done by the Autopilot

[4] Vertical Take-Off and Landing Unmanned Aerial Vehicles

In the Initial state *Alt-Capture* mode of the Autopilot is not armed (initial marking "Alt-Capture not Armed") – figure 2.

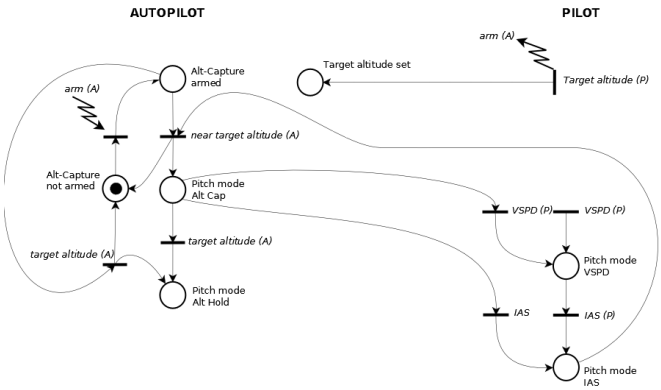

Figure 2: Alt-Capture not Armed

The Pilot sets altitude to *Target altitude*. This causes Autopilot *Alt–Capture* mode to arm, therefore the target altitude set by the Pilot will not be overshot. The Pilot also sets Pitch mode to *VSPD* (Vertical Speed – aircraft climbs at constant rate), then to *IAS* (Indicated Air Speed – climb rate adjusted, constant air speed) – figure 3.

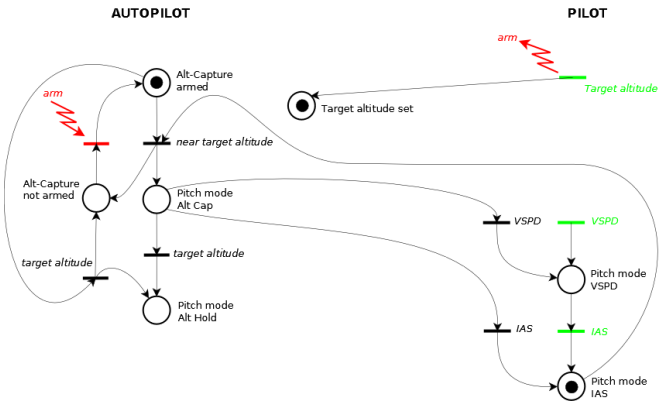

Figure 3: Alt-Capture armed and IAS

When target altitude is nearly reached, the Autopilot changes Pitch mode to *Alt Cap* (provides smooth levelling off at the desired altitude) therefore mode *Alt-Capture* is disarmed, so as *Pitch mode IAS* – figure 4.

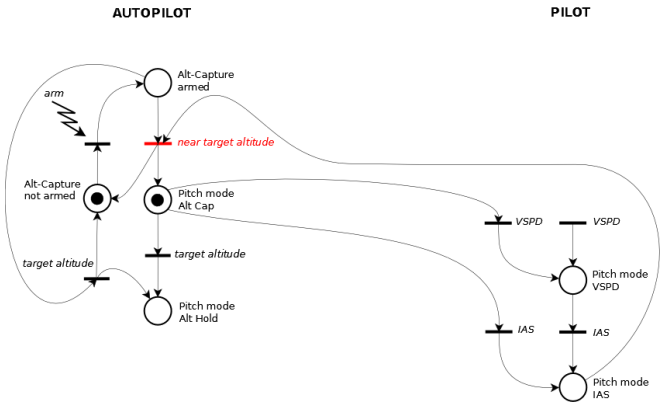

Figure 4: Alt Cap; Alt-Capture disarmed

The Pilot then changes Pitch mode to *VSPD*, therefore *Pitch mode Alt Cap* is disarmed – figure 5.

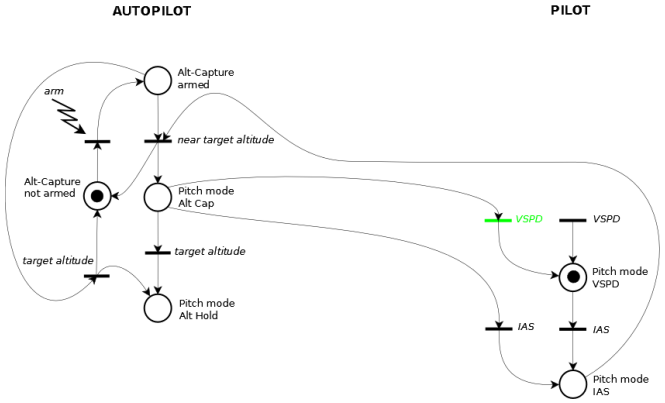

Figure 5: Pitch mode VSPD

When event *target altitude* occurs, state Pitch mode *Alt Hold* cannot be reached since neither possible precondition is true (Alt capture armed or Pitch mode Alt Cap). Therefore event *target altitude* is "lost" and the aircraft goes on climbing at the *VSPD* indicated by the pilot, – figure 6.

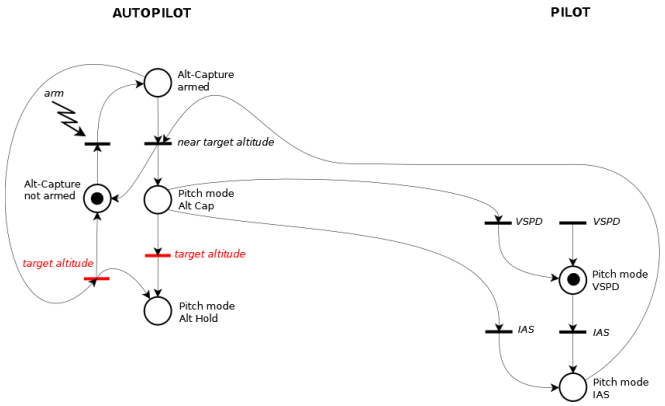

Figure 6: Event target altitude lost – "Oops, it didn't arm" [13].

The "Oops, it didn't arm" uttered by the pilot reveals that he does not understand why the aircraft goes on climbing. In fact, his actions on the Autopilot modes have destroyed the Autopilot sequence. Formally the Petri net is blocked on the Autopilot side (i.e. no transition can be fired anymore). This is a *knowledge conflict* [21] as the consequences of the agents' actions were neither assessed properly nor explained to one another.

2.2 Rain and automation

The second case of "automation surprise" occurred by chance during an experiment involving an Onera Ressac VTOL UAV in July 2011. Indeed the experiment was meant to test some properties of the Ressac planner and was not an ad-hoc scenario to bring about "automation surprise". The UAV mission requires two nominal pilots: the ground control station pilot (Gp) and the field pilot (Fp). For regulatory issues a third operator, the security pilot (Sp), can take over the manual piloting (as long as he wants) to deal with any unexpected event. About a dozen of other members of the Ressac team were checking the mission plan execution and performing other tasks.

There are five piloting modes (cf Table 1), one is totally automated (Nominal autopiloting- Autonav), three are partially automated modes and have been developed by Onera (Nominal autopiloting- Operator flight plan, Nominal manual- high level, Nominal manual- assisted), and the last one is a direct piloting mode (Emergency manual) using the on-the-shelf equipment of the vehicle (Yamaha RMax). This last mode can be engaged only by the Safety pilot who has always pre-emption rights, activating an exclusion switch cutting off the automatism. Notice that the Ressac software architecture has no visibility on the state of the switch. Flight phase transitions are allowed only in Nominal autopiloting mode.

	Automation	Gp	Fp	Sp	Phase achievement
Nominal autopiloting- Autonav	*				*
Nominal autopiloting- Operator flight plan	*	*	*		*
Nominal Manual- high level	*		*		
Nominal Manual- assisted	*		*		
Emergency Manual				*	

Table 1: Piloting modes, agents' involvement and phase achievement

So two nominal modes are possible i.e. Nominal autopiloting and Nominal manual piloting. When Nominal autopiloting is engaged, Ressac flies autonomously according to its plan, i.e. for this particular experiment:

- Phase 1: heading from the initial position to waypoint alpha
- Phase 2: heading from waypoint alpha to waypoint beta
- Phase 3: heading from waypoint beta to waypoint gamma

The following Petri nets represent the actions (transitions) and states (places) of the Ressac software agent (right) and of the human operator agent, i.e. what happens on the Gp's interface and the possible actions of the Sp (left). The procedure to follow (see figure 7 left) matches the plan (see figure 7 right) except the fact that it includes the case of the Sp taking control of Ressac to deal with an emergency: in that case the procedure is stopped. Initial state is human agent and software agent both in the state Phase 1.

In the Nominal autopiloting configuration the occurrence of Event A (waypoint alpha reached by Ressac) fires transition Phase 1/Phase 2 for the software agent. This transition emits Event B (information waypoint alpha reached displayed on the Gp interface) which updates the procedure: human agent state is Phase 2, so as software agent state.

Phase 2/ Phase 3 operates the same way with Event C (waypoint beta) and D (information displayed on the Gp interface and procedure updated).

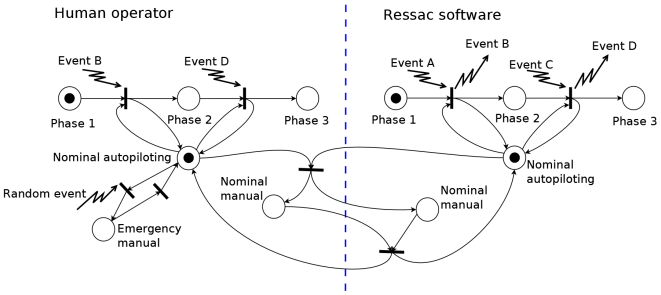

Figure 7: Initial state

What happened in July 2011 is the following sequence: Ressac was flying Phase 1 heading for waypoint alpha, when it began to rain. This random event made the Safety pilot Sp take over the control on Ressac. On the Petri net of figure 8 transition Random event is fired by the human agent and Emergency manual place is marked.

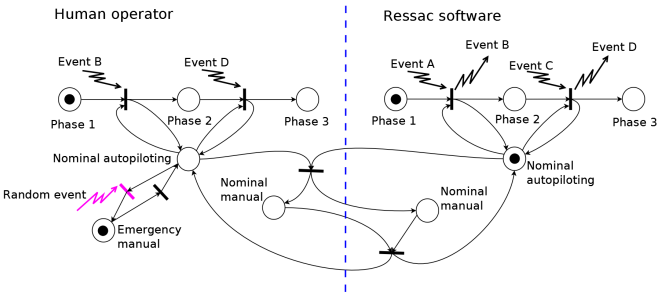

Figure 8: Rain and emergency manual mode

While operating Ressac manually in order to make it land, the Sp unintentionally flew it over waypoint alpha. Therefore Event A is generated, and the software agent engages Phase 2 (figure 9).

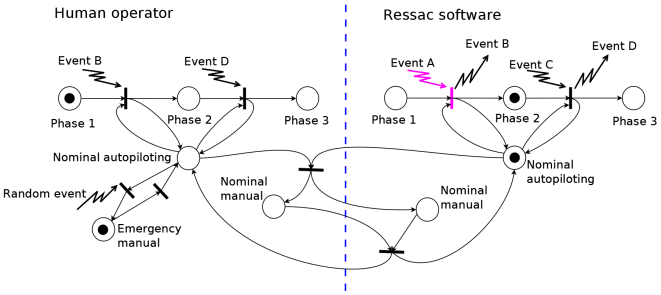

Figure 9: Software state update

Event B is emitted but lost on the human agent side, since one precondition (Nominal autopiloting) is no longer verified (figure 10).

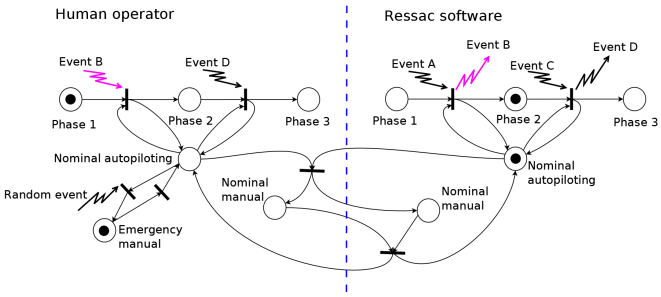

Figure 10: Lost of the event for the procedure update

The rain stopped and the Sp decided that the nominal plan could be resumed. Transition Emergency manual to Nominal autopiloting is fired (figure 11). The nominal plan was resumed (Phase 2) and Ressac headed waypoint beta. The human operators, who were expecting Phase 1 to be resumed, did not understand what Ressac was doing and began to panic. This is again a *knowledge conflict* [21] in which the human operators considered the behaviour of the machine as a failure. Indeed none of the test team members properly interpreted the behaviour of Ressac.

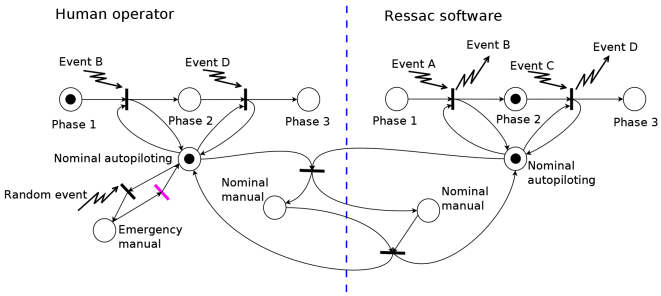

Figure 11: What the heck is it doing?

Notice that the marking of the Petri net (figure 11) is such that: (i) place Phase 2 is marked on the software agent side whereas place Phase 1 is marked on the human agent side ; (ii) one place Nominal piloting is marked (software agent side) whereas the other one is not marked (human agent side). Nevertheless it is a matter of semantic inconsistencies and not of formal inconsistencies within the Petri net model. Indeed for case (ii), the two places Nominal piloting do not represent the same state, otherwise a unique place would have been used: one is the software agent state and the other one is the human agent state.

Identifying conflicts through semantic inconsistencies would involve an explicit enumeration of all possible inconsistencies, which is hardly possible. Therefore what is relevant here from a formal point of view is not the semantic inconsistencies but the fact that the human agent part of the Petri net model is blocked (Event B will never occur again and Phase 2 will never be marked).

The next section will focus on a generalization of agent conflict representation, detection and solving.

3 Towards a model of human-automation conflict

3.1 Conflict model

In a multi-agent system different agents are often interested in the knowledge of the same state variables. Those variables can seman-

tically describe the physical environment state or the agent internal state. The values of those state variables can be affected by the agents' actions.

Let us consider two agents A1 and A2 that both have the right to act on a common device to change its state. The state of the device must be successively S1 then S2 and the agents must always have the same knowledge about the device state. The initial state is S1 In figure 12, both agents' knowledge is the same, i.e. the device state is S1 (left). The result of the firing of T1 is that both agents' knowledge is that system state is S2 (right). Note that transition T1 represent a synchronization of both agents about their shared decision.

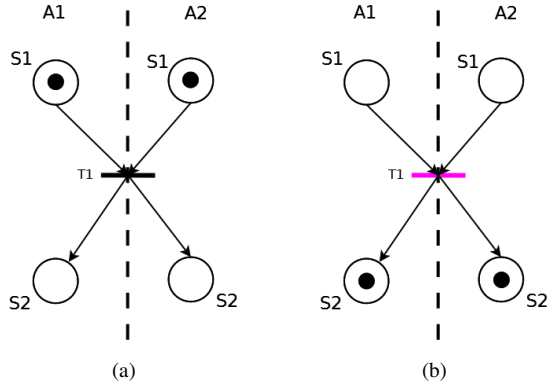

Figure 12: Two-agent system, correct design

As far as figure 13 is concerned, A2 need A1 to fire transition T2, i.e. both agents' knowledge must be S1 to make the device evolve to S2. On the contrary the firing of transition T1 only makes A1's knowledge state evolve to S2 (transition T1 is "hidden" from A2)(left). If T1 is fired, the result is that A1's knowledge is S2 whereas A2's is S1 and transition T2 is dead (right). This is a conflict.

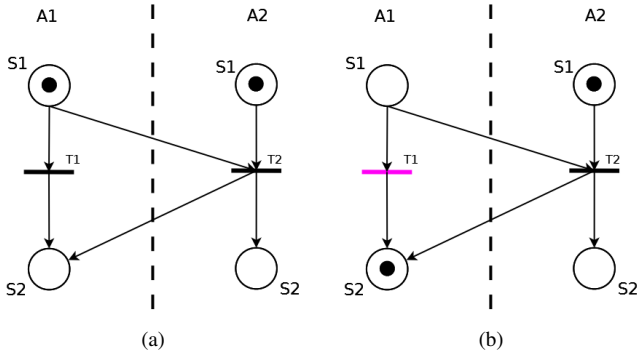

Figure 13: Two-agent system, incorrect design

3.2 Conflict solving

In figure 13 T1 is a 'hidden transition" so far as agent A2 cannot see it neither the consequences of its firing. That is the case for the "Rain and automation" example, figure 10.

Two solutions are then possible. The first one is to remove T1, i.e. agent A1 has no right to fire T1. In this case we get the ideal case in figure 12, we allow only *shared decisions* represented by transition T1. The second solution is to *inform* A2 of the firing of T1, see figure 14.

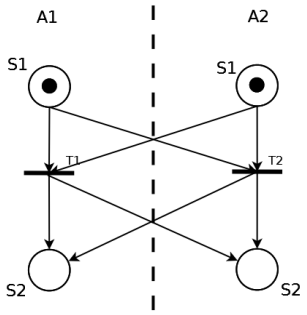

Figure 14: Two-agent system, another correct design

If A2 is a human operator the effect of a transition on his knowledge is not sure: the feedback he receives from the other agent can be lost or misinterpreted. A pseudo-firing [3] for T1 can model this kind of uncertainty, see figure 15 (left). The firing of T1 leads to the uncertain marking for the agent A2 state represented in figure 15 (right) by empty markers.

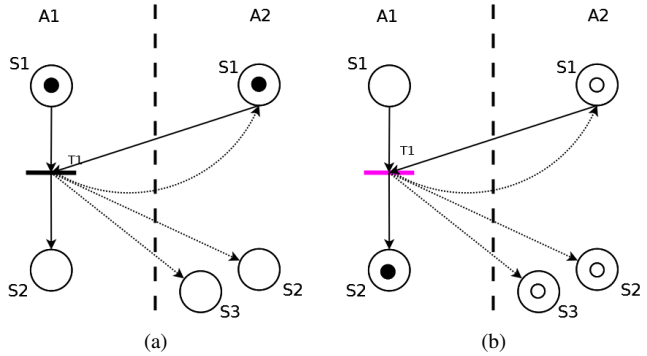

Figure 15: Two-agent system, A2 is human. Pseudo firing on correct design

For that reason the second solution proposed (inform the other agent) has an uncertain effect if A2 is human. This kind of transition is considered as a vulnerability by some researchers [7]. In other works the not nominal effect of a transition can be restored informing the human operator again or differently [6].

4 Conclusion and further work

Starting from two real cases of "automation surprises", we have shown that a formal model allows us to characterize a Human-Machine conflict: for both cases the Petri net model features a deadlock (i.e. at least one transition cannot be fired). We have then proposed a general Petri net based conflict model that paves the way for automatic conflict detection through "hidden" transitions identification and liveliness properties checking. We have also given two possible design solutions to prevent conflicts: *share the decision* or *inform the other agent*.

Nevertheless if the agent being informed is human the problem of the correct reception and interpretation of the information has to be considered. Therefore uncertainty has to be modelled so as to feed an estimator of the human agent's knowledge state: such an estimator,

which is further work, can be based on the human agent's actions and "internal state" [15].

Current work focuses on further aircraft autopilot-pilot interaction modelling – especially some cases that led to accidents – so as to put to the test the generic conflict model we have proposed. The next steps will be on-line conflict forecast and detection and experiments in our flight simulator.

5 Appendix: Petri Nets

A Petri net $< P, T, F, B >$ is a bipartite graph with two types of nodes: P is a finite set of places; T is a finite set of transitions [4]. Arcs are directed and represent the forward incidence function $F : P \times T \rightarrow \mathbb{N}$ and the backward incidence function $B : P \times T \rightarrow \mathbb{N}$ respectively. An *interpreted Petri net* is such that conditions and events are associated with places and transitions. When the conditions corresponding to some places are satisfied, tokens are assigned to those places and the net is said to be marked. The evolution of tokens within the net follows transition firing rules. Petri nets allow sequencing, parallelism and synchronization to be easily represented.

References

[1] D.B. Beringer and H.C. Harris Jr, 'Automation in general aviation: Two studies of pilot responses to autopilot malfunctions', *The International Journal of Aviation Psychology*, **9**(2), 155–174, (1999).

[2] E. Billings, *Aviation automation : the search for a human-centered approach*, Lawrence Erlbaum associates, Inc., Mahwah, NJ, USA, 1996.

[3] J. Cardoso, R. Valette, and D. Dubois, 'Possibilistic Petri nets', *Systems, Man, and Cybernetics*, **29**(5), 573 – 582, (1999).

[4] R. David and H. Alla, 'Discrete, continuous, and hybrid petri nets.', (2005).

[5] F. Dehais and P. Pasquier, 'Approche générique du conflit.', *ERGO-IHM*, (2000).

[6] F. Dehais, C. Tessier, L. Christophe, and F. Reuzeau, 'The perseveration syndrome in the pilot's activity: guidelines and cognitive countermeasures', *7th International Working Conference on Human Error, Safety, and System Development HESSD, year = 2009, volume = , number = , pages = .*

[7] Michael Feary., 'A toolset for supporting iterative human automation interaction in design', *NASA Ames Research Center, Tech. Rep. 20100012861*, (2010).

[8] T. Inagaki, 'Automation and the cost of authority', *International Journal of Industrial Ergonomics*, **31**(3), 169–174, (2003).

[9] Nicholas R. Jennings, 'Controlling cooperative problem solving in industrial multi-agent systems using joint intentions', *Artif. Intell.*, **75**(2), 195–240, (1995).

[10] S. Kraus, 'Negotiation and cooperation in multi-agent environments', *Artif. Intell.*, **94**(1-2), 79–97, (1997).

[11] S. Mercier, C. Tessier, and F. Dehais, 'Authority management in human-robot systems', *11th IFAC/IFIP/IFORS/IE Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, (2010).

[12] R. Mumaw, N. Sarter, and C. Wickens, 'Analysis of pilots' monitoring and performance on an automated flight deck', in *Proceedings of the 11th International Symposium for Aviation Psychology*, Columbus, OH, USA, (2001).

[13] E. Palmer, '"Oops, it didn't arm." A case study of two automation surprises', in *Proceedings of the Eighth International Symposium on Aviation Psychology*, eds., R. S. Jensen and L. A. Rakovan, pp. 227–232, Columbus, OH, USA, (1995).

[14] R. Parasuraman and C. Wickens, 'Humans : Still vital after all these years of automation', *Human factors*, **50**(3), 511–520, (2008).

[15] S. Pizziol, Fr. Dehais, and C. Tessier, 'Towards human operator "state" assessment', in *ATACCS'2011 - 1st International Conference on Application and Theory of Automation in Command and Control Systems*, Barcelona, Spain, (2011).

[16] J. Rushby, 'Using model checking to help discover mode confusions and other automation surprise', *Reliability Engineering and System Safety*, **75**(2), 167–177, (2002).

[17] J. Rushby, J. Crow, and E. Palmer, 'An automated method to detect potential mode confusions', in *18th AIAA/IEEE Digital Avionics Systems Conference*, St Louis, MO, (1999). Presentation slides.

[18] N.B. Sarter and D.D. Woods, 'How in the world did we ever get into that mode? Mode error and awareness in supervisory control', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **37**(1), 5–19, (1995).

[19] N.B. Sarter, D.D. Woods, and C.E. Billings, 'Automation surprises', in *Handbook of Human Factors and Ergonomics (2nd edition)*, ed., G. Salvendy, 1926–1943, New York, NY: Wiley, (1997).

[20] N. Su and J. Mylopoulos, 'Conceptualizing the co-evolution of organizations and information systems', in *Conceptual Modeling - ER 2006*, Tucson, AZ, USA, (2006).

[21] C. Tessier, H.-J. Müller, H. Fiorino, and L. Chaudron, 'Agents' conflicts: new issues', in *Conflicting agents - Conflict management in multi-agent systems*, eds., C. Tessier, L. Chaudron, and H.-J. Müller, Kluwer Academic Publishers, (2000).

[22] A.P. Tvaryanas, 'Visual scan patterns during simulated control of an Uninhabited Aerial Vehicle (UAV)', *Aviation, space, and environmental medicine*, **75**(6), 531–538, (2004).

[23] H. Van Ginkel, M. de Vries, J. Koeners, and E. Theunissen, 'Flexible authority allocation in unmanned aerial vehicles', in *Conference Proceedings of the Human Factors and Ergonomics Society*, San Francisco, CA, USA, (2006).

[24] C.D. Wickens, 'Situation awareness: review of Mica Endsley's 1995 articles on situation awareness theory and measurement', *Human factors*, **50**(3), 397–403, (2008).

[25] R. Wilensky, 'Planning and understanding: A computational approach to human reasoning', in *Reading, MA: Addison-Wesley.*, (1983).

[26] D. Woods and N. Sarter, 'Learning from automation surprises and going sour accidents', in *Cognitive engineering in the aviation domain*, eds., N. Sarter and R Amalberti, 327–353, Lawrence Erlbaum, New York, (2000).