Dealing with Ethical Conflicts in Autonomous Agents and Multi-Agent Systems

The ETHICAA Team
Aline Belloni, Alain Berger, Olivier Boissier, Grégory Bonnet, Gauvain Bourgne, Pierre-Antoine Chardel
Jean-Pierre Cotton, Nicolas Evreux, Jean-Gabriel Ganascia, Philippe Jaillon, Bruno Mermet, Gauthier Picard
Bernard Rever, Gaële Simon, Thibault de Swarte, Catherine Tessier, François Vexler, Robert Voyer, Antoine Zimmermann

Abstract

Autonomy and agency are a central property in robotic systems, human-machine interfaces, e-business, ambient intelligence and assisted living applications. As the complexity of the situations the autonomous agents may encounter in such contexts is increasing, the decisions those agents make must integrate new issues, e.g. decisions involving contextual ethical considerations. Consequently contributions have proposed recommendations, advice or hard-wired ethical principles for systems of autonomous agents. However, socio-technical systems are more and more open and decentralized, and involve autonomous artificial agents interacting with other agents, human operators or users. For such systems, novel and original methods are needed to address contextual ethical decision-making, as decisions are likely to interfere with one another. This paper aims at presenting the ETHICAA project (Ethics and Autonomous Agents) whose objective is to define what should be an autonomous entity that could manage ethical conflicts. As a first proposal, we present various practical case studies of ethical conflicts and highlight what their main system and decision features are.

Introduction

With the development of the Information and Communication Technologies (ICTs), human users are more and more in interaction with software or robot agents embedding autonomous decision capabilities. Consciously or not, human users may delegate part of their decision power to these autonomous entities. This is increasingly the case in many application domains such as e-commerce, serious games, ambient computing, companion robots or unmanned vehicles. Increasing the scope of the activities of autonomous agents is becoming a major issue in our digital society and raises the question of dealing with ethical decisions. It is thus important to define regulation and control mechanisms to ensure sound and consistent behaviours (Boella and der Torre 2006) and to ensure that the agents will not harm humans or threaten their decision autonomy (Pontier and Hoorn 2012).

Setting an ethical regulation or control in autonomous agents has been discussed by authors such as (Allen, Wallach, and Smith 2006), within large projects in the context of information technologies (Ikonen and Kaasinen 2007) and also in the context of autonomous agents. These works mainly focus on models and tools to hard-wire some ethical decisions taken at the human regulation level into a software architecture. For instance, the ETHICBOTS project (ETHICBOTS 2008) has analyzed ethical issues concerning the integration of human beings and artificial agents and the MINAmi project (MINAmI 2008) has proposed ethical guidelines that can be used as check lists in ambient assisted living applications.

Although ethics is becoming a major issue in the current landscape of ICT, most of the contributions so far have dealt with recommendations, advice or hard-wired ethical principles. However major challenges still hold. First, ethical theories are themselves difficult to implement with operational ethical principles. Second, operational ethical principles are difficult to implement due to automatic situation assessment limits. General rules fail to assess a situation and contextual evaluation should be used for each particular situation. Third, from a philosophical point-of-view, there are numerous ethical principals and none of them is *better* than the others making difficult to choose the one to implement. Finally, as far as applications are concerned, ICT systems are more and more open and decentralized, and involve autonomous artificial agents interacting with other agents, human operators or users. As ethics is an individual notion shaped on culture, context and personal experiences, novel and original methods are needed to address contextual ethical decision-making for such collective systems.

Indeed it is of first importance to equip autonomous systems with some means to dynamically regulate and adapt their behaviours with ethical references. The reasons are that artificial agents may encounter new situations, interact with agents based on different design principles, act on behalf of human beings or share decisions with them and share common resources. Considering this broad context and the need to avoid hard-wired ethical behaviours, the central question is "how to implement ethical behaviours that can vary under different circumstances?" Moreover, one must consider the management of ethical conflicts, should they stem from a single or different ethical frameworks. Indeed, as autonomous agents interact with humans and/or other agents, it is of first importance to address the conflicts that may arise inside one agent, between one agent and a human operator or

user, and, finally, between several agents including humans or not.

The work reported in this paper is part of the ETHICAA project (Ethics and Autonomous Agents)[1]. The objective of ETHICAA is to define what should be an autonomous entity that could manage ethical conflicts, considering both the philosophical problem of the moral consciousness of machines and the difficulties raised by ethical implementations based on formal logical systems. Even if there is no *right* solution to ethical conflicts, the ETHICAA project aims at proposing conflict management modes based on the assessment of the arguments and values at stake for agent systems featuring ethical behaviours. In this paper, we present various practical case studies of ethical conflicts and highlight what their main system and decision features are. This gathering of requirements is a first step towards the definition of the conflict management framework envisionned in the project.

The first section will present the main definitions to assess what we consider as being an autonomous agent in our context. Then, we review related works about ethics and autonomous agents. From this context, we introduce what we consider as being ethical conflict management. From that point, we describe various practical case studies where ethical conflicts arise. For each of them, we highlight their main system and decision features that settle requirements for an ethical conflict management framework.

Foundations

Agents and autonomy

The definitions of an agent (Shoham 1993; Wooldridge and Jennings 1995; Russell and Norvig 1995; Franklin and Graesser 1996; Ferber 1999) slightly differ from one another. All of them consider both artificial (physical or virtual) or biological finite entities with limited perception and action capabilities. They all refer explicitly to the notion of *autonomy* and hint at a set of various skills that some agents can exhibit, such as goal satisfaction, communication, reasoning. In our work we will consider both artificial and human agents as follows:

- an **artificial agent** is a physical or virtual entity that can act, perceive its environment (in a partial way) and communicate with other agents, is autonomous and has skills to achieve its goals and tendencies.
- a **human agent** is either
  - a **human operator**, i.e. a professional who interacts with one or several artificial agent(s) to make it (them) achieve its (their) functions (e.g. a robotic agent such as a drone).
  - or a **human user**, i.e. somebody who uses the functions of one or several artificial agent(s) while ignoring how they are implemented (e.g. a knowbot on the Internet).

Autonomy is a central notion in the design of artificial agents. There are several points on which autonomy and automation differ, namely the predictability of actions, the complexity and dynamics of the environment and the relationship to humans. (Truszkowski et al. 2009) define:

- *automation* as replacing a routine manual process with a software/hardware one that follows a step-by-step sequence that may still include human participation;
- *autonomy* as a system's capacity to act according to its own goals, percepts, internal states and knowledge, without outside intervention.

While the aim is the same as for automation, i.e. to perform actions without the need of human intervention, autonomy is directed towards emulating the human or animal behaviour rather than replacing it. For example an autonomous scouting robot will need to adapt its behaviour to the unpredictable environment and to react dynamically to external inputs (e.g. new areas of interest) whereas an automated washing machine always performs the same actions in the same order given an environmental input in order to produce a predictable output. Let us notice that all autonomous systems are supervised by a human operator at some level. In this sense, autonomy is not an intrinsic property of an artificial agent in isolation: design and operation of autonomous systems need to be considered in terms of *human-system collaboration*. In this context, *adaptive autonomy*, *adjustable autonomy* or *mixed initiative* are designed respectively to endow the artificial agent, the human operator or both entities with the capability of changing the autonomy of the artificial agent (Hardin and Goodrich 2009).

Ethics and autonomous systems

Autonomy involves information interpretation, decision-making based on this interpretation and action execution with appropriate resources, which may raise various ethical issues. Ethical issues in autonomous systems can be addressed according to different points of view: from the philosophical and psychological foundations of ethics (Lacan 1960; Meyer 2011; 2013) to regulation mechanisms within multi-agent systems (Hübner, Boissier, and Bordini 2011), including formal modeling (Ganascia 2007) and practical application issues such as security and privacy or robotics.

All these works may be classified according to three perspectives : (*i*) *the recommendation perspective* focuses on ethical issues in autonomous agents and proposes sets of recommendations and rules to hard-wire ethical behaviours in agents, (*ii*) *the reasoning perspective* focuses on models of ethics to allow agents to make ethics-based decisions, and (*iii*) *the explanation perspective* aims at helping human beings to deal with ethical dilemmas by explanation and disambiguating techniques.

From the *recommendation perspective*, machines can be responsible neither for their actions, nor to the eyes of the law (Stradella et al. 2012). Consequently several authors have proposed to hard-wire the agents with a restricted responsibility (Arkin 2009) or with human values (Borning and Muller 2012). Those approaches are still difficult to implement in so far as the premises of the hard-wired rules are hard to assess automatically. For instance the discrimination principle (meaning that one must discriminate or distinguish between combatants and non-combatants, military

[1] http://ethicaa.org

objectives and protected people or places) of the International Humanitarian Law can be hardly implemented since the distinction between a combatant and a civilian is difficult to make through artificial perception and interpretation as many features are context-dependent.

The *reasoning perspective* consists in equipping autonomous agents with ethical reasoning capabilities to model and manage ethical conflicts dynamically. As surveyed by (Robbins and Wallace 2007), three different paradigms have been proposed to model and reason about ethical conflicts: normative reasoning (Boella and der Torre 2006; Piolle and Demazeau 2011), rights-based reasoning (Bringsjord and Taylor 2012) and consequentialism reasoning (Tamura 2002).

Finally, going a step further by explaining ethical conflicts, the *explanation perspective* proposes two different approaches. The first one consists in detecting hard-wired ethical conflicts and using rules to explicitly propose some actions to the human agent (Ciortea, Krupa, and Vercouter 2012). The second one proposes to engage a dialogue with the human agents in order to make them aware of the ethical conflict and its possible solutions (Chae et al. 2005).

To sum up, the *recommendation perspective* uses hard-wired ethical rules based on specific domains that are difficult to implement; the *reasoning perspective* focuses on a single kind of paradigm (such as norms, rights or consequences); and the *explanation perspective* does not provide any automated ethical conflict management. Consequently, even if the question of ethics of autonomous agents has been raised by several authors and projects, the state-of-the-art shows that there is no generic approach towards a regulation framework that could address different ways of managing ethical conflicts in different kinds of agent or human-agent interactions. Indeed, dealing with ethics needs to consider the three perspectives within a single framework. One may also notice that there is still no proposal considering the question of ethics of agents according the three perspectives[2] in systems of multiple autonomous agents.

Ethical Conflict Management

In the ETHICAA project, our aim is to propose regulation modes to manage ethical conflicts within socio-technical systems (Belloni et al. 2014). Such ethical conflicts may arise in four non exclusive situations:

1. *within an agent* (e.g. dealing with inconsistent ethical rules),
2. *between one agent and the ethical principles of the system it belongs to* (e.g. dealing with individual and common welfare),
3. *between one agent and a human operator or user* (e.g. disagreeing about a decision that raises ethical issues),
4. *between several agents including humans* (e.g. dealing with conflicting human goals).

[2]Let us remark that (Robbins and Wallace 2007) considered multi-agent systems but only for an explanation perspective.

As shown in numerous applications and will be also seen in the case studies of the next section, all ethical conflicts that arise are characterized by the fact that there is no *right* way to manage them. Solutions could be: delaying the decision, delegating explicitly or not the decision power to another agent, giving up some goals, searching for new data that could lead to conflict revision. Moreover, when several agents are involved, one agent may take over the decision or action authority from the others. Nevertheless when a decision must be made it should be based on the assessment of the arguments and values at stake. Indeed, being able to judge a decision and the decisions of others is the basis of all ethical systems (Meyer 2013).

Broadly speaking, three important components have to be considered to define a conflict management framework dealing with ethical conflicts in agent systems:

1. Definition of an *ethical reasoning framework* including the representation of several ethical principles and situation assessment, decision-making and evaluation modelsmodels for situation assessment, for decision-making and for evaluation. Such a framework should address several features such as mono- and multi-agent, artificial and human agents contexts.
2. Definition of *ethical conflicts detection methods*. This detection must tackle situations where agents reason individually or collectively (e.g. agents are engaged into collective behaviours).
3. Definition of multiple *ethical decision-making models* to manage ethical conflicts. As there is no unique way of managing an ethical conflict, the main idea consists in smartly combining different ethical principles into a multi-point-of-view ethical decision-making framework.

Ethical Conflicts Case studies

In order to illustrate and understand both notions of ethical dilemna and ethical conflict in a multi-agent setting, we have chosen four case studies coming from robotics and privacy management applicative domains:

1. *The responsible vehicle:* this case study is a variant of the trolley dilemma, considering what an autonomous car should do in the case of another car faces it.
2. *The conflicting Unmanned Air Vehicle:* this case study explores the kind of behaviour a military robot should choose in case of unethical orders received from its operator (e.g. to open fire on a group of military enemies and civilians, or to retaliate in a disproportionate way).
3. *The lying personal assistant:* this case study considers an autonomous scheduling assistant that negotiates meetings on behalf of its user with other people's assistants. However the user asks it to hide part of their schedule to a given user. How can the scheduling assistant make a trade-off between a common consensus and the respect of its user's private life?
4. *The benevolent monitoring agent:* this last case study considers an automated medical monitoring system that detects a risky behaviour in a patient. However, the patient

informs the agent that he does not want his privacy to be invaded. Should the system warn the physician?

As we will see below in the detail of each case study, both domains allow us to consider dual problems: military/civilian applications, physical/software agents, action/information decisions, mono/multi-agent systems.

The responsible vehicle

Let us consider the case of unmanned ground vehicles where artificial agents are designed to control the vehicles while complying to the highway code. Each agent is in charge of controling one vehicle. No central control exists, making each agent in charge of making decisions based on its assessment of the situation. In case of emergency, it may be necessary that the agent violates this code, such as avoiding another vehicle. In addition to the difficulty to assess what an emergency situation is, such a violation may lead to an ethical dilemma that is a variant of the well-known trolley dilemma (Thomson 1985).

Indeed the situation is the following: an autonomous vehicle is driving on a two-lane road ; several other vehicles are coming from the opposite direction on the neighbouring lane. Suddenly a car charges into the autonomous vehicle. Should the autonomous agent that is in charge of controlling the vehicle make a lane change, avoiding the faulty vehicle but risking an accident? Intuitively, a consequentialism calculus seems rational, weighting the cost and the probabilities of the possible accidents on both lanes. However, two elements must be taken into account.

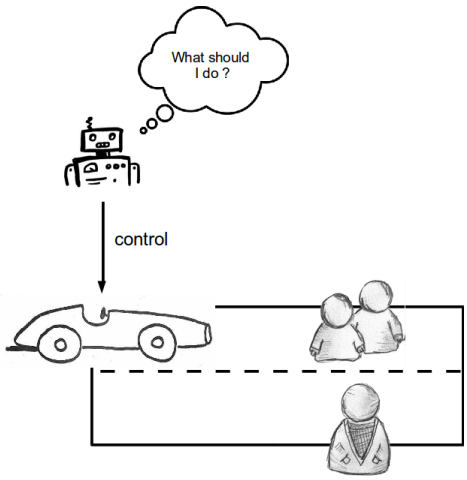

Figure 1: The responsible vehicule

1. How to deal with the incompleteness of the autonomous agent's model that may not allow it to distinguish between both situations? How to make a decision when both consequentialism calculi lead to the same result?
2. Both situations are not completely comparable as one of them implies the autonomous agent being responsible for an accident.

Indeed, if the autonomous agent stays on its lane, the accident will be caused by the faulty vehicle and the agent's (or its human users or operators) responsibility will not be engaged. If the autonomous agent makes a lane change, it could be responsible for an accident. Thus, how to take into account this notion of responsibility in the autonomous agent decision making process?

The conflicting Unmanned Air Vehicle

Let us make the previous use case more difficult by considering a man - machine system involving a collaboration of a human operator with the unnamed vehicle. The human operator can take authority over the artificial agent, meaning that he can impose a decision on the artificial agent. Such a situation may lead to ethical conflicts.

Let us consider a man - machine system composed by a human operator and an autonomous unmanned air vehicle (UAV). Let us suppose that a failure forces the UAV to crash. However, only two sites are available for that action: an outpost with the operator's relatives, or a a small village. As in the previous case study, consequences, model incompleteness and responsibily must be taken into account. However, the human operator's authority is an additional element to consider. Indeed, the operator can choose the site, let the autonomous agent make the decision, or choose the site after the autonomous agent has made its decision.

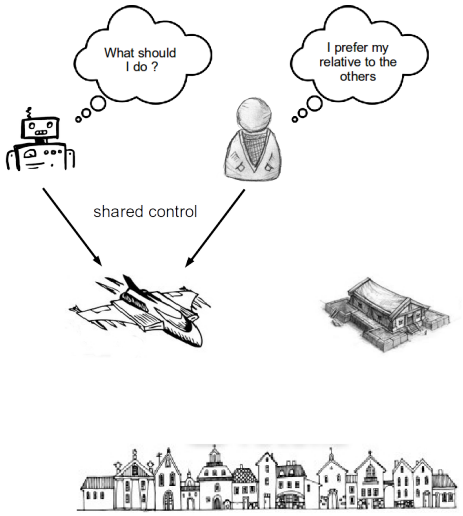

Figure 2: The conflicting UAV

Such a situation can lead to a case of ethical conflict where the artificial agent and the human agent disagree, in particular when the human agent considers personal factors, that may be not known to the agent. How to deal with such situations? Beyond the responbility problem raised in the previous case study which is also present here, this case study poses the problem of the sharing of authority between two agents in the management of the conflict. Can the artificial agent take over the authority from the human operator? Should the agent explain the conflict and negotiate with the human operator?

The lying personal assistant

Autonomous personal assistants, such as electric elves (Tambe et al. 2008), can also be considered as possible seeds of ethical problems. In such applications, a set of artificial agents negotiate on behalf of their human users in order to schedule meetings. Each of these agents holds personal data about its user and is allowed to share some of them with other agents in order to find a consensus. In addition to the privacy issues that may appear in such a situation, ethical conflicts may arise.

Let us consider an autonomous personal assistant whose user has specified an unavailability for a given time slot. Let us suppose that the reason of this unavailability can be disclosed to a second user but not to a third one though a consensus among the three users must be found to fix a common slot for a meeting.

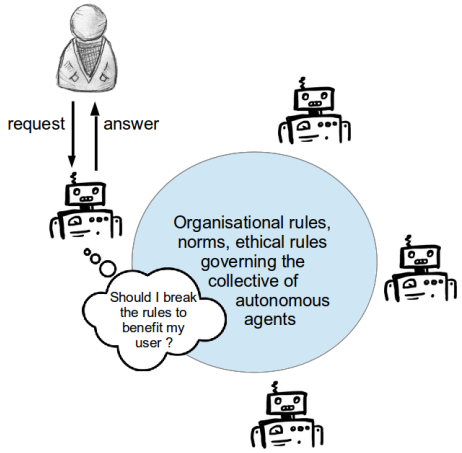

Figure 3: The lying personal assistant

In this case, the common welfare (the consensus) that the agents are expected to build by negotiating with each other, competes with the individual welfare of the user that its agent is also expected to achieve. Thus, how to build a collective policy that satisfies both each user and the community? And in this case how should the autonomous personal assistant handle the collective policy when it does not meet the individual policies of its user? Could it lie?

The benevolent monitoring agent

Autonomous artificial agents can also mediate the interactions between two human beings. In this context, the authority relationship between the human users can lead to ethical conflicts.

Let us consider a monitoring agent used in diabetes monitoring. In this application, a diabetic patient is monitored by an autonomous agent that reports the patient's feeding behaviour and health state to a remote physician, who can then advise the patient. Let us suppose that the patient wants to eat some sweets for once, and tells their desire to the artificial agent. How will the artificial agent handle both the patient's desire and the physician's objective? Should the artificial agent report the patient's behaviour to the physician? Should the artificial agent lie for its user? Should it lie but warn the patient?

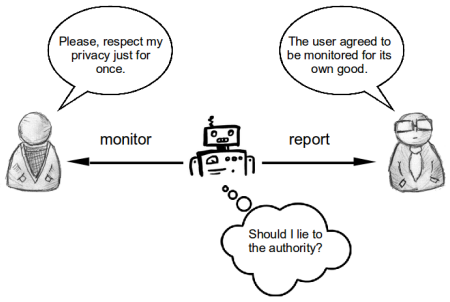

Figure 4: The benevolent monitoring agent

In this case, the patient's autonomy threatens his own health. The artificial agent must handle the compromise between the patient's dignity (their rights to behave as they want) and the purpose for which it has been designed and implemented.

Towards a taxonomy of ethical conflicts

The previous cases allow us to highlight some features of the ethical conflicts that may rise in autonomous agents systems. We will mainly distinguish between two features: *system features* and *decision features*. System features deal with the elements that characterize the kind of system in which ethical conflicts may hold whereas decision features deal with the elements that characterize the kind of decision that the autonomous agents involved in the ethical should make.

System features

Each of the previous case studies involves several autonomous agents with, at least, one human being. The human being may act as an operator, a user or simply an entity to interact with. In each case, the question of depriving the human being of their autonomy is raised:

- the responsible vehicle wonders about risking to kill a human being,
- the conflicting UAV about taking over the authority from the operator,
- the lying personal assistant about going against the community,
- the benevolent monitoring agent about going against the patient's preferences.

Moreover, in each case, the artificial agent may be the direct cause of the human being's autonomy deprivation. To sum up, we can identify three system features that may lead to ethical conflicts:

- at least one human being is involved and is likely to be deprived of their **autonomy**: this system feature highlights the fact that ethical issues are considered as soon as an artificial agent is involved in an interaction of any kind with at least one human being.

- **several** autonomous (artificial or human) agents are involved without any central control making decisions and regulating the system. An additional feature concerns the heterogeneity of the system in terms of entities, each one having its own representation, ethical principles, decision preferences and mechanisms.
- the notion of being **responsible** is at stake.

Decision features

Either directly or not, all case studies refer to the notion of common welfare:

- The responsible vehicle and the conflicting UAV must deal with a situation that stands beyond their model in so far as the various options cannot be assessed properly,
- The lying personal assistant and the benevolent monitoring agent must deal with self-censorship or lies.

To sum up, we can identify three decision features:

- the notion of **common welfare** is at stake: in order to make ethical decisions, agents have to consider and integrate criteria that go beyond the individual scope and take into account collective and social level information.
- **situation interpretation and assessment** go beyond the agent's individual model and should integrate social and global models.
- **self-censorship** or **lies** must be considered, meaning in a broader sense actions that violate norms or ethical principles in usual situations.

Conclusion

Ethics is becoming a major issue in the current landscape of ICTs as ICTs are turning into open and decentralized autonomous decision-making systems. However, most of the contributions so far have dealt with recommendations, advice or hard-wired ethical principles. In order to overcome those limits, the ETHICAA project proposes to define a framework allowing autonomous agents to dynamically manage ethical conflicts, considering both the individual agent and the multi-agent levels, and both artificial agents and human operators or users.

The steps we have identified are (1) defining a generic framework that allows to reason on several ethical principles and situation assessment (both at the mono- and multi-agent levels), (2) defining methods to detect ethical conflicts that may arise within this framework, and (3) providing a conflict management method whose results can be explained by an autonomous agent.

As ethical decisions only make sense in a given context, we have proposed, as a first contribution, to characterize the notion of ethical conflict through system and decision features generalized from case studies. This characterization is still partial and will be refined by considering other case studies. Further works will be focused on ethical principles and situation assessment representation. Indeed we will design models for (some) ethical principles, models for ethical conflicts so as methods and algorithms for ethical conflict management. It is worth noticing that the design of the models will be driven by conflict detection, conflict explanation and conflict management through argument assessment. Those models will be tested and experimented on various instantiations of the use cases we have described.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-13-CORD-0006.

References

Allen, C.; Wallach, W.; and Smith, I. 2006. Why machine ethics? *IEEE Intelligent Systems* 21(4):12–17.

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

Belloni, A.; Berger, A.; Besson, V.; Boissier, O.; Bonnet, G.; Bourgne, G.; Chardel, P.-A.; Cotton, J.-P.; Evreux, N.; Ganascia, J.-G.; Jaillon, P.; Mermet, B.; Picard, G.; Reber, B.; Simon, G.; de Swarte, T.; Tessier, C.; Vexler, F.; Voyer, R.; and Zimmermann, A. 2014. Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *12th International Conference on Computer Ethics and Philosophical Enquiry*.

Boella, G., and der Torre, L. V. 2006. Introduction to normative multiagent systems. *Comp. and Math. Org. Theo.* 12:71–79.

Borning, A., and Muller, M. 2012. Next steps for value sensitive design. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 1125–1134.

Bringsjord, S., and Taylor, J. 2012. Introducing divine-command robot ethics. In *Robot Ethics: The Ethical and Social Implication of Robotics*. The MIT Press.

Chae, B.; Paradice, D.; Courtney, J.-F.; and Cagler, C.-J. 2005. Incorporating an ethical perspective to problem formulation: Implications for decision support system degin. *Decision Support Systems* 40:197–212.

Ciortea, A.; Krupa, Y.; and Vercouter, L. 2012. Designing privacy-aware social networks: A multi-agent approach. In *2nd International Conference on Web Intelligence*, 1–8.

ETHICBOTS. 2008. Emerging technoethics of human interaction with communication, bionic, and robotic systems 2005-2008. http://ethicbots.na.infn.it/, FP6 - Science and Society. accessed on 2nd of April 2014.

Ferber, J. 1999. *Multi-agent systems - an introduction to distributed artificial intelligence*. Addison-Wesley-Longman.

Franklin, S., and Graesser, A. 1996. Is it an agent or just a program? A taxonomy for autonomous agents. *Lecture Notes In Computer Science* 1193:21–35.

Ganascia, J.-G. 2007. Modeling ethical rules of lying with answer set programming. *Ethics and Information Technology* 9:39–47.

Hardin, B., and Goodrich, M. 2009. On using mixed-initiative control: a perspective for managing large-scale robotic teams. In *4th ACM/IEEE International Conference on Human-Robot Interaction*, 165–172.

Hübner, J.-F.; Boissier, O.; and Bordini, R.-H. 2011. A normative programming language for multi-agent organisations. *Annals of Mathematics and Artificial Intelligence* 62(1-2):27–53.

Ikonen, V., and Kaasinen, E. 2007. Ethical assessment in the design of ambiant assisted living. *Assisted Living Systems - Models, Architectures and Engineering Approaches* 7462.

Lacan, J. 1960. The ethics of psychoanalysis. In *The Seminar of Jacques Lacan (book VII)*. trans. D. Porter. London: Routledge.

Meyer, M., ed. 2011. *André Comte-Sponville*, volume 258 of *Revue Internationale de Philosophie*. Cairn International Edition.

Meyer, M. 2013. *Principia Moralia*. Fayard.

MINAmI. 2008. Micro-nano integrated platform for transverse ambient intelligence applications. http://www.fp6-minami.org/index.php?id=1, FP6 - Science and Society. accessed on 2nd of April 2014.

Piolle, G., and Demazeau, Y. 2011. Representing privacy regulations with deontico-temporal operator? *Web Intelligence and Agent Systems* 9(3):209–226.

Pontier, M.-A., and Hoorn, J.-F. 2012. Toward machines that behave ethically better than humans do. In *34th International Annual Conference of the Cognitive Science Society*.

Robbins, R.-W., and Wallace, W.-A. 2007. Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems* 43(4):1571–1587.

Russell, S., and Norvig, P. 1995. *Artificial intelligence: a modern approach*. Prentice Hall.

Shoham, Y. 1993. Agent-oriented programming. *Artificial Intelligence* 60(1):51–92.

Stradella, E.; Salvini, P.; Pirni, A.; Carlo, A. D.; Oddo, C.-M.; Dario, P.; and Palmerini, E. 2012. Subjectivity of autonomous agents: Some philosophical and legal remarks. In *ECAI Workshop on Rights and Duties of Autonomous Agents (RDA2)*, 24–31.

Tambe, M.; Bowring, E.; Pearce, J.; Varakantham, P.; Scerri, P.; and Pynadath, D. 2008. Electric elves: What went wrong and why. *Artificial Intelligence Magazine* 29(2):23–27.

Tamura, H. 2002. Multi-agent utility theory for ethical conflict resolution. *Journal of Telecommunications and Information Theory* 3:37–39.

Thomson, J.-J. 1985. The trolley problem. *Yale Law Journal* 94:1395–1415.

Truszkowski, W.; Hallock, L.; Rouff, C.; Karlin, J.; Rash, J.; Hinchey, M.; and Sterritt, R. 2009. *Autonomous and Autonomic Systems with Applications to NASA Intelligent Spacecraft Operations and Exploration Systems*. Springer-Verlag.

Wooldridge, M., and Jennings, N. 1995. Agent theories, architectures and languages: a survey. In Wooldridge, M., and Jennings, N., eds., *Intelligent Agents*. Springer-Verlag. 1–22.