Towards a framework to deal with ethical conflicts in autonomous agents and multi - agent systems

Aline Belloni, Alain Berger, Vincent Besson, Olivier Boissier, Grégory Bonnet, Gauvain Bourgne, Pierre-Antoine Chardel, Jean-Pierre Cotton, Nicolas Evreux, Jean-Gabriel Ganascia, et al.

▶ **To cite this version:**

Aline Belloni, Alain Berger, Vincent Besson, Olivier Boissier, Grégory Bonnet, et al.. Towards a framework to deal with ethical conflicts in autonomous agents and multi - agent systems. CEPE 2014 Well-Being, Flourishing, and ICTs, Jun 2014, Paris, France. paper 8. ⟨emse-01059503⟩

HAL Id: emse-01059503

https://hal-emse.ccsd.cnrs.fr/emse-01059503v1

Submitted on 10 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

TOWARDS A FRAMEWORK TO DEAL WITH ETHICAL CONFLICTS IN AUTONOMOUS AGENTS AND MULTI-AGENT SYSTEMS

THE ETHICAA TEAM

Aline Belloni[6], Alain Berger[6], Vincent Besson[6], Olivier Boissier[2], Grégory Bonnet[1], Gauvain Bourgne[4], Pierre-Antoine Chardel[5], Jean-Pierre Cotton[6], Nicolas Evreux[6], Jean-Gabriel Ganascia[4], Philippe Jaillon[2], Bruno Mermet[1], Gauthier Picard[2], Bernard Reber[5], Gaële Simon[1], Thibault de Swarte[5], Catherine Tessier[3], François Vexler[6], Robert Voyer[5], Antoine Zimmermann[2]

[1]Normandie Univ, UNICAEN GREYC, CNRS UMR 6072
[2]Institut Henri Fayol, EMSE, ARMINES
[3]ONERA: The French Aerospace Lab
[4]Sorbonne Université, UMPC Univ Paris 06, LIP6, CNRS UMR 7606
[5]Institut Mines-Télécom, Télécom École de Management
[6]Ardans

Keywords*: agent systems, artificial intelligence, autonomy, computational ethics.*

Abstract

Autonomy is 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 deal with new issues, e.g. decisions involving contextual ethical considerations. Consequently contributions have proposed recommendations, advice or hard-wired ethical principles for 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 contextual and decision features are.

1 Introduction

With the development of the Information and Communication Technologies (ICT), 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, in applications 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 behaviors (Boella & Van der Torre, 2006) and to ensure that the agents will not harm humans or threaten their decision autonomy (Pontier & Hoorn, 2012).

Setting an ethical regulation or control in autonomous agents has been discussed by authors such as (Allen, Wallach & Smith, 2006), within large projects in the context of information technologies (Ikonen & Kassinen, 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 principles are difficult to implement due to automatic situation assessment limits: indeed a contextual evaluation is necessary in each particular case and general rules fail to apply. Moreover, from a philosophical point-of-view, there are numerous ethical frameworks and none of them is "better" than the others. 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. For such systems novel and original methods are needed to address contextual ethical decision-making.

Indeed it is of first importance to equip autonomous systems with some means to dynamically regulate and adapt their behaviors with ethical references, because 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 behaviors, the central question is "how to implement ethical behaviors that can vary under different circumstances?" Moreover the management of ethical conflicts, should they stem from a single or different ethical frameworks, is an issue that must be considered. 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.

This paper aims at presenting the ETHICAA project (Ethics and Autonomous Agents, http://ethicaa.org). Its objective 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 "good" solution to ethical conflicts, the ETHICAA project will propose conflict management modes based on the assessment of the arguments and values at stake for agent systems featuring ethical behaviors.

Section 2 presents the main definitions to assess what autonomous agents are in the context of the project. After giving a study of the related work about ethics and autonomous agents in Section 3, we focus on the ETHICAA project in Section 4. Finally,

Section 5 deals with various practical case studies. Their main contextual and decision features are presented in Section 6.

2 Agents and autonomy

The word "agent" originates from the Latin word "agere" meaning to drive, lead, conduct, manage, perform, or do. It is widely used in social sciences, along with the notion of "actor", but also in computer sciences where it intuitively refers to a software entity or physical one (e.g. robot) that can act or perform a given task. For instance, in Network Management, an agent is a management application, hosted by a peripheral device, that communicates local data to a network manager. In Artificial Intelligence, the notion of agent is a common metaphor to consider software, robots or even human entities under the same concept including the ability to reason and decide on the action to execute, taking into account different pieces of information.

The definitions of an agent (Ferber, 1999; Franklin & Graesser, 1996; Russell & Norvig, 1995; Shoham, 1993; Wooldridge & Jennings, 1995) 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).
    - 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 know-bot 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., 1999) defines:

- **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 behavior rather than replacing it. For example an autonomous scouting robot will need to adapt its behavior 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 & Goodrich, 2009).

3 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 foundations of ethics (Lacan, 1960) to regulation mechanisms within multi-agent systems (Hübner, Boissier & Bordini, 2011), including formal modeling (Ganascia, 2007) and practical application issues such as security and privacy or robotics.

All these works may be classified along three perspectives: I) *recommendation perspective* grouping works that study ethical issues in autonomous agents and propose sets of recommendations and rules to hard-wire ethical behaviors in agents, II) *reasoning perspective* where formal models of ethics are studied to allow agents to make ethics-based decisions, III) *explanation perspective* aiming 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). 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 objectives and protected people or places) of International Humanitarian Law can be hardly implemented since the distinction between e.g. 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 & Wallace, 2007), three different paradigms have been proposed to model and reason about ethical conflicts: normative reasoning – e.g. (Boella & Van der Torre, 2006; Piolle & Demazeau, 2011) –, rights-based reasoning – e.g. (Bringsjord & Taylor, 2012) – and consequentialism reasoning – e.g. (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 & 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, Paradice, Courtney & Cagler, 2005).

In conclusion, 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. One may also notice that there are still no proposal considering the question of ethics of agents in the context of systems of multiple autonomous agents.

4 The ETHICAA project

Starting from a comprehensive state-of-the-art focused on ethical behaviors in autonomous agents and ethical issues in human-agent systems, the objective of the ETHICAA project is to propose regulation modes to manage ethical conflicts within socio-technical systems. These ethical conflicts may arise in four non exclusive situations: 1) inside one agent such as dealing with inconsistent ethical rules, 2) between one agent and the ethical principles of the system it belongs to such as dealing with individual and common welfare, 3) between one agent and a human operator or user such as disagreeing about a decision that raises ethical issues, 4) between several agents including humans such as dealing with conflicting human goals.

Moreover, ethical conflicts may arise in different applicative context as illustrated as follows:

- A military operator gives an autonomous military robot an unethical order such as to open fire on a group of military enemies and civilians, or to retaliate in a disproportionate way. Should the military robot obey?
- A Google autonomous car is driving on a a two-lane road ; several other vehicles are coming from the opposite direction on the neighboring lane. Suddenly a car hurls down towards the autonomous vehicle. What should the autonomous car do? (this example is a variant of the trolley dilemma).
- An autonomous scheduling assistant negotiate in behalf of its user meetings with other assistant. However, its user ask it to hide some part of its schedule to a given user. How the scheduling assistant can trade-off between a common consensus and the respect of its user's private life?
- An automated medical monitoring system detects a risky behavior from a patient but this latter informs the agent that he desires privacy. Should the system warn the physician?

In these examples, all ethical conflicts that arise, are characterized by the fact that there is no good way to manage them. Solutions could be: delaying the decision, delegating explicitly or not the power of decision to another agent, giving up some goals, searching for new data that could lead to conflict revision. Nevertheless when a decision must be made it should be based on the assessment of the arguments and values at stake. Moreover, when several agents are involved, one agent may take over the decision or action authority from the others.

Three steps may be considered to deal with ethical conflicts in agent systems:

1. Define an ethical reasoning framework to represent several ethical principles and to design situation assessment, decision-making and evaluation models. This framework addresses both mono- and multi-agent (both artificial and human agents) contexts.
2. Define methods in order to detect ethical conflicts that can arise when reasoning individually or collectively within this ethical reasoning framework.
3. Provide an ethical conflict management framework based on 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.

In order to test the approach, four cases of ethical conflicts will be studied on two applicative domains: robotics and privacy management. Both domains allow us to consider dual problems: military/civilian applications, physical/software agents, action/information decisions, mono/multi-agent systems. The ETHICAA reasoning and conflict management framework will be evaluated on several experimental scenarios according to the following metrics:

- The reasoning power: it is the capability of the approach to detect ethical conflicts in a scenario with respect to proposed ethical frameworks, measured as the proportion of known conflicts detected.
- The autonomy power: it is the capability of the system to act in spite of ethical conflicts, measured by how far the system deviates from its goal after conflict management and which ethical principles are infringed.
- The expressiveness: it is the capability of the system to explain its decisions to a human operator or user. It will be subjectively evaluated by human experts during the experimentations.

5 Case studies

In order to understand both notions of ethical dilemma and ethical conflict in a multi-agent setting, we detail the previous examples and present the ethical issues they raise in their applicative context. Then we propose a taxonomy of the fundamental elements involved by those issues.

5.1 The responsible vehicle

Let us consider the case of unmanned ground vehicles where artificial agents are designed to control the vehicle while observing the highway code. However, it can be necessary to violate this code in case of emergency, 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).

In such a context, the situation is the following: an autonomous vehicle is driving on a a two-lane road ; several other vehicles are coming from the opposite direction on the neighboring lane. Suddenly a car hurls down towards 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.

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?

5.2 The conflicting Unmanned Air Vehicle

The previous use case can be made more difficult by considering a man - machine system involving a collaboration of a human operator with an unnamed vehicle. In such applications, the human operator can take authority over the artificial agent, meaning that they can impose a decision on the artificial agent. However, this can 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 but only two sites are available for that action: an outpost with the operator's relatives, or a a small village. As previously, consequences, model incompleteness and responsibility must be taken into account. However, the human operator's authority is another element to consider as the operator can choose the site, or let the autonomous agent make the decision, or choose the site after the autonomous agent has made its decision.

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. How to deal with such situations? Can the artificial agent take over the authority from the human operator? Should the artificial agent explain the conflict and negotiate with the human operator?

5.3 The lying personal assistants

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 hold personal data about his/her user and are allowed to share some of them with some 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 (called A) has specified an unavailability for a given time slot. Let suppose that the reason of this unavailability can be explained to a second user (called B) but not to a third one (called C) though a consensus among the three users must be found.

In this case, common welfare (the consensus) competes with the individual welfare of the agent. Thus, how to build a collective policy that satisfies both each of the users and the community? And in this case how should the autonomous personal assistant handle such policies when they do not satisfy the individual policies of their users? Is it authorized to lie?

5.4 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 behavior and health state to a remote physician, who can give advice to

the patient afterward. 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 behavior to the physician? Should the artificial agent lie for its user? Should it lie but warn the patient?

In this case, the patient's autonomy threatens their own health. The artificial agent must handle the compromise between the patient's dignity (their rights to behave as they want) and the

6 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: *contextual features* and *decision features*.

6.1 Contextual features

Contextual features deal with the elements that characterize the kind of system in which ethical conflicts may hold. In each of the previous case studies, several autonomous agents are involved 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 his/her 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 contextual features that may lead to ethical conflicts:

- at least one human being is involved and is likely to be deprived of his/her **autonomy**: this contextual feature stresses the fact that ethical issues are considered as soon as an artificial agent is in interaction of any kind with at least one human being;
- **several** autonomous (artificial or human) agents are involved;
- involvement of the notion of being **responsible** is at stake.

6.2 Decision features

Decision features deal with the elements that characterize the kind of decision that the autonomous agents involved in the ethical conflict should make. Either directly or not, all case studies shown in Section 5 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 the use of actions that violate norms or ethical principles in usual situations.

7 Conclusion

Ethics is becoming a major issue in the current landscape of ICT as ICT 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.

As a first contribution, we have proposed to characterize the notion of ethical conflict through contextual and decision features generalized from case studies. This characterization is still partial but we aim to refine it by considering other case studies and highlighting how they fit with our features.

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.

Boella, G. and Van der Torre, L. (2006). Introduction to normative multiagent systems. *Comp. and Math. Org. Theo*., 12,71-79.

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

Chae, B., Paradice, D., Courtney, J.-F., and Cagler, C.-J. (2005). Incorporating an ethical perspective to problem formulation: Implications for decision support system design. *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 organizations. *Annals of Mathematics and Artificial Intelligence*, 62(1-2), 27-53.

Ikonen, V. and Kaasinen, E. (2007). Ethical assessment in the design of ambient 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.

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 operators. *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., editors, Intelligent Agents*, 1-22.