A roadmap towards ethical autonomous agents

ANR ETHICAA – ANR-13-CORD-0006
Delivrable #3

Main authors
Olivier Boissier (ARMINES-FAYOL)
Grégory Bonnet (GREYC)
Jean-Gabriel Ganascia (LIP6)
Catherine Tessier (Onera)
Thibault de Swarte (IMT)
Robert Voyer (IMT)

May 6, 2015

Contents

Chapter 1

An overview of autonomous agents

With the development of the Information and Communication Technologies (ICT), human users are more and more in interaction with software or robots embedding decision capabilities. Consciously or not human users delegate part of their decision power to these autonomous entities. This is the case in an increasing number of application domains such as e-commerce, serious games, ambient computing, companion robots or unmanned vehicles [Aarts and de Ruyter, 2009]. These evolutions are the result of research and works done in the multi-agent system domain, subfield of Artificial Intelligence. In the literature, the terms intelligent systems, autonomous agents and multi-agent systems arose about thirty years ago, as highlighted by Figure 1.1.

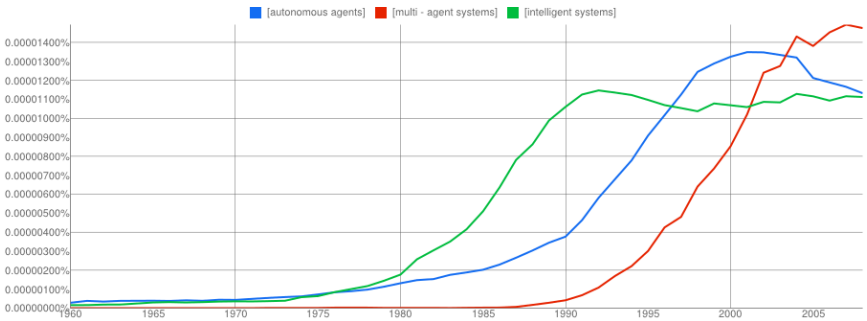

Figure 1.1: Agency emergence based on Google's Ngram [Michel et al., 2010]

In the academic world, the first workshops and conferences on autono-

mous agents and multi-agent systems held at the end of the eighties (DAI [1] Workshop, MAAMAW [2] Workshop, Multi-Agent and Concurrent Programming Workshop) have made possible the creation of the ICMAS [3] Conference, later transformed into AAMAS [4], which is currently the most important conference in this domain.

In the following, we will first define what we mean in this report by *autonomous agents* and *systems* composed of such entities. From these first definitions, we analyze the multi-agent literature to extract the agent and multi-agent features that raise ethical issues. Throughout this chapter, we justify and motivate the research questions that will be studied in the context of the ETHICAA project.

1.1 Autonomous agents

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 science where it intuitively refers to an entity 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, depending on the models, the ability to reason and decide on the action to execute, taking into account different pieces of information. In this sense, it is of first importance to distinguish between the human subject and the autonomous agent.

A simple definition of the human subject is that each and every person has the chance to tell their own story. By comparison, an artificial agent, if it tells a story, will tell a story that has been written by a human subject. A human subject is always split between a conscious side, the part of the psyche that is accessible, and an unconscious part that is a continuous series of instinctual drives a big part of which remain inaccessible. By definition, an artificial agent does not have any kind of unconscious feelings or drives, even if it could be a vector of unconscious drives for the humans. According

[1] Distributed Artificial Intelligence Workshop

[2] Modeling Autonomous Agents in a Multi-Agent World

[3] International Conference on Multi-Agent Systems - First conference was held in 1995.

[4] International Joint Conference on Autonomous Agents and Multi-Agent Systems - Founded in 2002

to [Freud, 1916], the unconscious continues to influence our behaviour and experience, even though we are unaware of these underlying influences.

Whatever it be, in the sequel we will consider two kinds of autonomous agents referring to the reasoning entities that can be found in the socio-technical systems that are considered in the project: human agents and artificial agents. The notion of autonomy will be defined in Section 1.3.3.

1.1.1 Human agents

Let us notice that we know that, since Freud and the origin of Psychoanalysis, there is a discussion inside the academic community on the status of the unconscious. By definition, no experimental research protocol can be developed in order to demonstrate the existence of the unconscious, because experimental methodology proposes an objective investigation, where there is a separation between the subject, who is investigating, and objects, which are investigated. It is a through relationship between speakers and through language that some contents emerge from the unconscious [Lacand and Fink, 1966, Nogueira, 2004].

This human part of the human subject is the emotional one and this part is unprogrammable, even it can be simulated. As stated by [Turkle and Shapiro, 2011], simulate love is never love. Thus, for developing relevant ethical approaches in the autonomous agents' field it is necessary to consider the human as a whole and not only as an agent. However, for ease of reading, we will use the terms *human agent* in the sequel.

Consequently, a human agent may refer to:

- a human user, i.e. somebody who uses the functions of an artificial agent while ignoring how they are implemented (e.g. a knowbot on the Internet);
- a human operator [Mercier, 2011], i.e. a professional who interacts with an artificial agent to make it achieve its functions (e.g. a robotic agent such as a drone).

1.1.2 Artificial agents

Many definitions of what an agent is have been proposed in the Artificial Intelligence literature [Shoham, 1993,Wooldridge and Jennings, 1995,Russell and Norvig, 1995,Franklin and Graesser, 1996,Ferber, 1999]. The most well-known definitions are the following:

Definition 1.1 (Agent [Russell and Norvig, 1995]) *An* agent *is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors.*

Definition 1.2 (Agent [Franklin and Graesser, 1996]) *An autonomous* agent *is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.*

Definition 1.3 (Agent [Ferber, 1999]) *An* agent *can be a physical or virtual entity that can act, perceive its environment (in a partial way) and communicate with others, is autonomous and has skills to achieve its goals and tendencies.*

Although all definitions slightly differ, it is worth noticing that:

- all definitions can apply both to artificial (physical or virtual) or biological entities as they consider as a basis finite entities with limited perception and action capabilities. Consequently the notion of agent covers a large taxonomy, as shown on Figure 1.2.
- two out of the three definitions refer explicitely to the notion of *autonomy* and hint at a set of various skills that some agents can exhibit, such as goal satisfaction, communication and reasoning.
- two out of the three definitions refer explicitely to the notion of *goal*: artificiel agents are designed in order to achieve goals on the behalf of human users or operators.

In the literature different kinds of agents are considered with respect to their architectures and skills. Besides their ability to cooperate and take part in organizations [Boissier, 2001], agents may be classified as reactive agents or cognitive agents according to their reasoning abilities, although the boundaries between both classes are not as clear as they might seem. For instance, [Shiloni et al., 2009] have shown that a set of reactive agents can simulate cognitive agents for some tasks. Moreover, different kinds of hybrid agents can be designed.

Reactive agents

Reactive agents are inspired from the early works on the subsumption architecture [Brooks, 1986]. The main feature of those agents is that they do

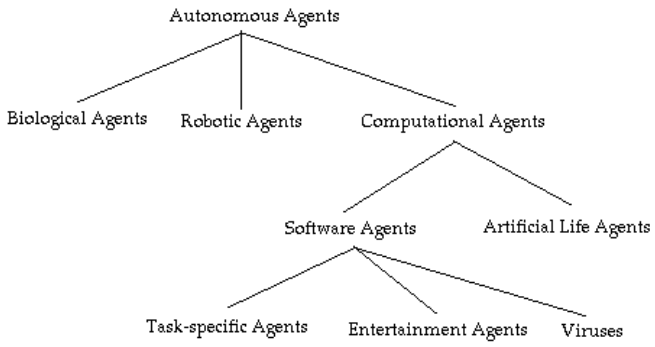

Figure 1.2: Taxonomy of agents [Franklin and Graesser, 1996]

not reason on a world model but they react automatically to some perceived events. Consequently they are based on a set a predefined behaviours that trigger in order to exhibit a complex global behaviour. According to [Brooks, 1991], those behaviours must satisfy some properties. They should be *situated* meaning that they are triggered only according to the perceived environment, *embodied* meaning that the environment must not be reduced to a model, *intelligent* meaning they must be in accordance with the environment and *emergent* meaning that an external observer must understand the global behaviour as intelligent. [Werger, 1999] introduced three new properties: *minimalism* meaning that each single behaviour must use the minimal set of resources or information, *stateless* meaning a behaviour should not have an internal state and *tolerant* meaning the uncertainty and the incompleteness of perception should be taken into account. The main limits of such kinds of agents are that the global behaviour is difficult to formalize and therefore can be non-optimal because of multiple negative interactions between local behaviours [Drogoul, 1995].

Cognitive agents

The main feature of cognitive agents is that they can reason on their environment in order to infer actions to execute or new goals to achieve. Consequently, such agents are endowed with a world model that represents their environment and an action model that represents the changes the agent can make on the environment. Cognitive agents are also *hysteretic* meaning that they memorize information on their past states and on the environment in order to reuse them later. BDI (Belief Desire Intention) architectures [Rao

and Georgeff, 1991] or MDP (Markov Decision Process) architectures [Putterman, 1994] are best suited to implement cognitive agents. The main limits of such kinds of agents are the difficulty to endow them with a correct world model and the high algorithmic complexity of the decision calculus. For instance, many MDP approaches propose to compute the optimal policy[5] offline and execute it reactively.

Hybrid agents

As stated previously, a hybrid reactive and cognitive architecture can be implemented in the same agent. Such agents are defined by a cognitive module and a reactive module that interact with each other, such as Reactive-Delibrative Architectures, InteRRap Architectures or Touring Machines [Lemaître and Verfaillie, 2007,Rodriguez-Moreno et al., 2007,Aschwanden et al., 2006,Muller and Pischel, 1993,Ferguson, 1992]. The reactive module decides for the next action to execute with respect to the agent's perception but, at the same time, the agent computes another next action based on its world model. The final action is chosen according to the time the agent has to make its decision.

1.2 Systems of autonomous agents

As seen in the previous definitions, an agent is an entity that is situated in an environment inhabited by other agents. In other words an autonomous agent cannot be studied without considering the environment and the other agents with which it interacts directly or indirectly. This is why we will now focus on the analysis of the different kinds of systems of autonomous agents.

1.2.1 Systems of human agents

As highlighted by [Whitworth, 2006], some computer systems can be seen as more than just mechanical systems. For instance, systems like e-mail, chat rooms, bulletin boards, online trading web sites, virtual communities imply a huge involvement of human agents. Such systems are called *socio-technical systems*, in which an instantiation of socio (human agents) and technical (machines or software) elements is engaged towards the achievement of a common goal.

[5] A policy is a function $\pi : S \mapsto A$ giving for each state $s \in S$ an action $a \in A$ to do.

In the context of our study, even if we could consider any kind of system of human agents interacting with each other with the support of some computer supported platforms, we will focus on systems of autonomous agents.

Definition 1.4 (System of autonomous agents) *A system of autonomous agents is a system where there is at least one artificial agent interacting with other autonomous agents, whether artificial or not.*

1.2.2 Systems of artificial agents

A system of artificial agents situated in a shared environment is called a multi-agent system. In [Ferber, 1999], a *multi-agent system* (MAS) is composed of an environment, objects and agents (the agents being the only ones to act), relations between all the entities, a set of actions that can be performed by the entities and the changes of the system both in time and due to these actions. However, even if relations are cited in this definition, the notion of organization (see below) is missing, although it is a fundamental dimension of systems of autonomous agents as stated by [Boissier et al., 2010]. In the sequel, we will consider the following definition of a MAS:

Definition 1.5 (Multi-agent system) *A* multi-agent system *(MAS) is a set of agents that interact with each other, situated in a common and shared environment, and that may build or participate in an organization.*

It can be noticed that four important dimensions participate in the definition of a MAS: Agents, Environment, Interaction and Organization. Agents having been defined in the previous section, let us turn to the definition and explanation of the other three dimensions.

Environment

All non-agent entities of a multi-agent system are generally considered to be part of the environment. Known in the literature as objects, resources or even artifacts [Omicini et al., 2008], such entities can be software entities as databases, Web services or coordination tools (e.g. blackboard, electronic institutions), or also physical entities such as communication infrastructures, energy sources or obstacles. In the sequel, we will consider a single distinction between artifacts and all other non-agent entities.

Definition 1.6 (Artifact [Omicini et al., 2008]) Artifacts *represent passive components of the system such as resources and media that are intentionally constructed, shared, manipulated and used by agents to support their activities, either cooperatively or competitively.*

Then, the environment is defined as a slight adaptation of the definition given in [Weyns et al., 2007]:

Definition 1.7 (Environment) *The* environment *is a first-class abstraction that provides the surrounding conditions for agents to exist and that mediates both the interactions among agents and the interactions among agents and artifacts.*

Interaction

Definition 1.8 (Interaction) *In a MAS, interaction is one of the internal engines. It consists in a dynamic relation between two or several agents through reciprocal actions. Interaction exists as soon as the internal dynamics of an agent changes according to the influences of the other agents.*

Depending on the type of the system, the interactions that take place between the agents in the shared environment may be of different natures: *indirect* interaction, i.e. interaction mediated by the shared environment (e.g. stigmergy) or *direct* interaction, i.e. interaction consisting in the exchange of messages between agents. These interactions may participate in different kinds of activities taking place between the agents.

For instance, the agents can *coordinate* with each other. Here coordination has a larger meaning than simply *synchronization* of agents: coordination can go from action scheduling with respect to the other agents' actions to *collaboration* – meaning finding joint actions in order to achieve common goals and *negotiation* – meaning choosing how to share a common resource for different goals [Durfee, 2001, Ferber, 1999, Nwana et al., 1996].

All these interaction situations may be undertaken in the context or under the regulation of an agents organization.

Organizations

Several definitions of what an organization is have been proposed in the literature (eg. [Corkill and Lander, 1998, Franklin and Graesser, 1996, Lemaître and Excelente, 1998, Sichman et al., 2005]). Its meaning often varies between two basic views: (i) a collective entity with an identity that is represented by (but not identical to) a group of agents exhibiting relatively highly formalized social structures [Scott, 1998], (ii) a stable pattern/structure of joint activities that may constrain or affect the actions and interactions of agents towards some purpose [Castelfranchi, 1998]. In a general sense, organization

refers to a *cooperation pattern* that can be more or less formalized. As in Sociology [Bernoux, 1985], it may concern the expression of a division of tasks, a distribution of roles, an authority system, a communication system, or also a contribution-retribution system. According to [Gasser, 2001], this range of topics may also be extended to knowledge, culture, memory or history.

Definition 1.9 (Organization) *An agents organization is a purposive* supra-agent *pattern of emergent or (pre)defined agents cooperation, that can be defined by the designer of the system of agents or by the autonomous agents themselves.*

In the literature, a pattern of emergent cooperation is called *organization entity*, institution, coalitions, social relations or commitments. A pattern of (pre)defined cooperation is called *organization specification*, structure or norms. For instance, in [Hubner et al., 2002], organizations are specified around three axes: a *structural* axis defining how information and decisions disseminate, a *functional* axis defining how the agents coordinate, and a *deontic* axis defining the norms in terms of obligations and permissions for the agents.

In [Horling and Lesser, 2004], three forms of organizations are distinguished, each having several variations:

- **groups** are flat organizations used by a set of agents to synchronize themselves or to share resources. According to the means, the ends and the size of such organizations, *teams*, *coalitions* or *congregations* are considered [Sandholm et al., 1999,Brooks and Durfee, 2003]. Whereas coalitions are temporary, a team is a long-lifecycle flat group composed of agents that have agreed to work together towards a common goal [Horling and Lesser, 2004]. A *federation* is a group that uses a delegate agent to interact with other groups.

- **hierarchies** are tree structures based on the divide and conquer principle, with an ascending flow of information and a descending flow of decisions. We can consider multiple kinds of hierarchies (flat, sharped, uniform) according to their structural properties. We can also consider a nested form of hierarchy, called *holonic system*, where groups of agents (called holons) are structured into multiples hierarchies [Fischer et al., 2003].

- **societies** are open organizations, meaning an autonomous agent can join or quit the system at any time, with a long lifecycle [Buzing et al., 2005]. Societies allow heterogeneous agents to interact through a common communication and negociation framework, such as *normative systems*, *market rules* or other social functions. Thus, societies are high-level organizations that can contain other organizations.

1.2.3 Systems of human agents and artificial agents

Even if the Artificial Intelligence literature dedicated to autonomous agents seems to focus on artificial agents, many socio-technical systems[6] involve both human agents and artificial agents who interact in order to achieve their goals [Hoc, 2000]. Following the kinds of human agents identified in Section 1.1.1, one may distinguish between several kinds of systems that combine human and artificial agents. Moreover, considering the types of the interaction relations that may exist between such agents, several types of systems should be considered. Indeed a distinction should be made between artificial agents that are *supervised* by human agents and artificial agents that are not supervised.

In the case of supervision, the human agent is an operator who interacts with the agent (e.g. a robotic agent such as a drone) to make it *achieve* its functions whereas in the case of no supervision, the human agent is a user who uses the functions of the agent (e.g. a search agent on the Internet) while ignoring how they are implemented. An example of supervised system can be found in [Coppin and Legras, 2012] where a military operator supervises a bio-inspired UAV swarm in order to achieve various missions. In this case the operator is trained to use the system through high-level algorithms (such as watching, avoiding, intercepting). An example of unsupervised system is social interactive agents such as conversational agents [Fong et al., 2003] where the human user asks requests to the artificial agent in natural language. In this case the artificial agent must be able to dialogue with the human user.

[6] Let us notice that, for instance, Professor Hiroshi Ishiguro's Geminoid robot – as being fully tele-operated and without processing capabilities – cannot be considered as a socio-technical system, and even more can be considered neither automated, nor autonomous (see Section 1.3.3).

Systems of artificial agents supervised by human agents

Many definitions of systems where artificial agents are supervised by human agents have been proposed, ranging from the domain of joint cognitive systems [Hollnagel and Woods, 1983] to those of human-robot interaction [Goodrich and Schultz, 2007] and systems of systems [Luzeaux, 2013]. In the sequel, we will consider a slightly adapted definition given by [Pizziol, 2013].

Definition 1.10 (Human supervised artificial agent system) *A human supervised artificial agent system is a team composed of human operators and artificial agents with a common goal. They communicate and act on a physical system for the achievement of their goal. The goal achievement is pursued through the execution of functions. Some of those functions can only be executed by the human operators, some only by the artificial agents.*

Let us notice that, in this definition, there is no model to explain how the human operators perceive and interpret information, nor how they make decisions.

An organization may structure or regulate their interaction leading to particular forms of human supervised artificial agent organizations. In [Yanco and Drury, 2004], eight kinds of human - artificial agent organizations are identified: (1) one human operator and one artificial agent; (2) one human operator and a team of artificial agents; (3) one human operator and multiple artificial agents; (4) a team of human operators and one artificial agent; (5) multiple human operators and one artificial agent; (6) a team of human operators and a team of artificial agents; (7) a team of human operators and multiple artificial agents; (8) multiple human operators and a team of artificial agents.

Let us notice that in this enumeration the case where multiple human operators and multiple artificial agents (both outside a team) interact is not considered. Indeed they claim that coordination must necessarily happen either at the human operators' level or at the artificial agents' level as the human supervised artificial agent system explicitely refers to a common goal for all agents. Thus either the human operators or the artificial agents form a team, and the case where multiple human operators and multiple artificial agents interact is captured by one of the eight kinds of organizations.

Therefore the notion of common goal is a central notion in human supervised artificial agent systems. It may not be the case with artificial agents interacting with a human user.

System of artificial agents interacting with a human user

Many socio-technical systems involve artificial agents and a human user who ignores how the agents are implemented. Artificial agents interacting with a human user cover a large number of domains, from social interactive agents [Fong et al., 2003] to agent-mediated electronic commerce [Guttman and Maes, 1998], including ambient intelligence [Sadri, 2011].

Social interactive agents operate as partners, peers or assistants for human users, which means that they need to exhibit a certain degree of adaptability and flexibility to drive the interaction with a wide range of humans [Fong et al., 2003]. Agent-mediated electronic commerce applies to business-to-business, business-to-consumer, and consumer-to-consumer transactions where a personalized, continuously running, semi-autonomous behaviour is desirable [Guttman and Maes, 1998]. Ambient intelligence is an intelligent, embedded, digital environment that is sensitive and responsive to the presence of people, helping them in daily life [Sadri, 2011].

In all three domains, the human agents are users and the goal of the autonomous agents is to satisfy their users. Therefore the autonomous agents must be aware of their users' personal requirements and preferences, and interact with them in a user-friendly way and possibly expressing, recognizing and responding to emotions. Consequently they need to express and/or perceive emotions, communicate with high-level dialogue, learn/recognize models of other agents, establish/maintain social relationships, use natural cues (gaze, gestures, etc.), exhibit distinctive personality and character, and learn/develop social skills.

Definition 1.11 (Artificial agent system interacting with users) *An artificial agent system interacting with users is a hierarchy composed of human users and artificial agents whose goal is to satisfy their own users. The artificial agents must be aware of the users' personal requirements and preferences, and interact with them in a user-friendly way.*

The interaction of an artificial agent with its user can be classified according to its:

- **embodiement**, which is the form and the structure of the artificial agent in the eyes of its user. The artificial agent can be anthropomorphic, zoomorphic, caricatured, or functional (meaning with a simple software appearance);

- **interaction mode**, which is how the artificial agent can monitor its user's activity. It can use speech, gesture or face recognition, gaze tracking, and user modeling;

- **dialogue capabilities**, which is how the artificial agent can communicate information to its user. It can be through low-level signals such as raw data, non-verbal communication such as sound or light warning, or natural language;

- **emotional treatment**, which is how the artificial agent can recognize and mimic emotions – indeed it has been shown that people tend to treat computers as they treat other people. Emotion can also provide feedback to the user, such as indicating the agent's internal state, goals and (to an extent) intentions.

Let us notice that some users might be unconscious users: they do not know they interact with an artificial agent and they cannot express goals or preferences. For instance, this kind of users can be pedestrians or drivers that cross an autonomous car on the road. Therefore, artificial agent systems interacting with users must take those unexpected users into consideration.

1.3 Features that may raise ethical issues

In order to better assess what makes the autonomous agent systems different from other socio-technical systems, this section reviews the main features of autonomous agents and autonomous agents systems.

1.3.1 Openness & Heterogeneity

Openness

Openness has been introduced as a key feature of MAS since the 80's. With such a characterisation, the researchers wanted to emphasize the fact that MAS are situated in an environment. Such a feature is important in the context of the project, since openness concerns:

- the dynamicity of the system, i.e. agents (human or artificial ones) are faced to changing situations to which they have to adapt their behaviours;

- the dynamic entry/exit of agents into/from the system, i.e. besides the dynamicity of the environment to which the agents have to adapt,

they are faced to the entry/exit of other autonomous entities (either human or artificial ones) with which they have to interact. Adaptation is not only in terms of context or description of the environment but also in terms of interaction, of the way agents have to consider and reason on the others.

In order to manage openness, some research works have proposed to introduce organization in order to provide the agents with the possibility to reason on a finite set of participants. However, even if the organization defines some boundaries, it should provide the following properties in order to keep the openness property alive:

- the open agents' organization should be permeable in the sense of allowing the dynamic arrival/exit of agents into/from it;
- the open agent's organization should have a reorganization ability;
- the open agent's organization should control the agent's autonomy in the sense of using mechanisms and rules that incent agents to avoid undesirable behaviours.

Heterogeneity

Even if the organization, the interaction and the environment are aimed at providing the agents with common and shared constructs among agents, heterogeneity still remains in agents. Given the definitions that were given in the previous sections, we clearly see that heterogeneity may exist in a MAS in many forms: architecture (human agents and artificial agents), decision mechanism (reactive agents and cognitive agents). Moreover, the preferences, principles and values of each agent may be differents. Since multi-agent systems are open systems as described above, heterogeneity is an inherent feature of such system since different kinds of agent may enter and participate to the system.

1.3.2 Reasoning, External & Internal Description

As described in [Boissier, 2003], in the process of the engineering of a system of autonomous agents, a cycle is in action, that involves a *designer* who designs the system, which behaviour can be observed by an *observer*. The system can be described by a set of properties that are understandable by the observer, by the designer or, in case of reflexivity, by the system itself. Introducing the observer in our context is particularly interesting since some

properties of the system that may be observed do not have any practical and direct implementation in the system as built by the designer. Such properties are the result of the functioning of the system and of the interpretation of the observer. This kind of phenomenon refers to the notion of *emergence* [Pesty et al., 1997].

The *external description* of the system refers to the description of the system based on a set of properties that are used by the observer. This description is an *objective* point of view in the sense that it is built from the functioning of the system without knowing what the designer had in mind and how the system is really implementing this behaviour.

The *internal description* of the system refers to the description of the system based on a set of properties that are used by the designer. These properties express a *subjective* point of view, specifying the way the designer intends that the system implement the properties. Such properties are usually expressed using the constructs used to model the system.

In the case of reflexivity, the system can reason on itself, i.e. the system may be the observer of its proper execution. In the case of a system of autonomous agents, this capability may be distributed among the autonomous agents participating to the system (human or artificial ones). Each agent may thus be involved in:

- the definition and modification of a property, i.e. the agent plays the designer's role,
- the observation and monitoring of a property, i.e. the agent plays the observer's role.

An agent can handle the internal and external descriptions of the same phenomenon. For instance, in the case of reorganisation in a MAS, agents may participate themselves to the definition of the organization that structures and regulates their behaviours and interactions with other agents in the system. This definition is realised from the observation and monitoring of the functioning of the agents in the organization (violation of norms, failures, etc.)

Let us notice that the internal and external descriptions of the system and of an agent may not coincide. It is indeed possible for an observer to state an agent as cognitive (external description) based on its behavior whereas its internal architecture is a reactive agent architecture based on a simple automata [Demazeau and Müller, 1991]. This comment may be extended to the interaction, organization or environment models.

1.3.3 Autonomy

Autonomy is obviously a central notion in the design of autonomous agents. Beyond the philosophical question of autonomy and consciousness of machines [Stradella et al., 2012], [Goodrich and Schultz, 2007] claim that designing autonomy consists of mapping inputs from the environment into actuator movements, representational schemas or speech acts. Numerous definitions of autonomy have been proposed in the literature. Early works describe autonomy in terms of *level of autonomy* [Sheridan and Verplank, 1978].

However, [Defense Science Board, 2012] recommends to give up this notion. It replaces it with an *autonomous systems reference framework* that explicitely focuses design decisions on the explicit allocation of cognitive functions and responsibilities between the human user or operator and the artificial agent to achieve specific capabilities, and explicitely recognizes that these allocations may vary by mission phase as well as echelon. For instance, [Dorais et al., 1999] already focused on the complexity of commands, [Goodrich et al., 2001] on the duration an artificial agent is independant from the operator or [Bradshaw et al., 2003] on the deontic rules that constrain the agents. More generaly, [Carabelea et al., 2003] highlight that autonomy must be viewed from an external perspective (an agent is autonomous with respect to another one for a given function in a given context, if in this context, its behaviour regarding the function is not imposed by the other agent) and from an internal perspective (how an agent is able to exhibit autonomous behaviours in various situations).

There are several points on which autonomy and automation differ, namely the predictability of actions, the structure of the environment and the relationship to humans. An automated machine will carry out step-by-step sequences of actions that are determined *a priori*.

Definition 1.12 (Automated processes [Truszkowski et al., 2009]) *An* automated process *simply replaces a routine manual process with software/hardware one that follows a step-by-step sequence that may still include human participation.*

Thus (apart from machine failures) the actions of the machine are fully predictable, and cannot be adapted to any unpredicted state of its environment. The machine then needs to operate within a well-known environment in order for it to perform its sequence of actions successfully [Docherty, 2012]. Even though the routines may still include human participation, they are

designed to achieve predictable results desired by humans. For example a washing machine always performs the same actions in the same order given an environmental input, in order to produce a predictable output. In another domain, an automated process onboard a satellite could be an attitude determination function requiring no a priori attitude initialization. The process does not define when the process should begin (it computes attitudes whenever star data is available), simply outputs the result for some other application to use, and in the event of an anomaly that causes the attitude determination function to fail, it takes no remedial action (it just outputs an error message).

On the other hand, an autonomous agent will be able to operate and adapt in open and unstructured environments. As said previously, several definition have been proposed.

Definition 1.13 (Autonomy [Castelfranchi and Falcone, 2003]) Autonomy *is a relationship between the artificial agent and the human agent.*

Definition 1.14 (Autonomy [Bekey, 2005]) Autonomy *is the capacity of a robot to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.*

Definition 1.15 (Autonomy [Truszkowski et al., 2009]) Autonomy *is a system's capacity to act according to its own goals, percepts, internal states, and knowledge, without outside intervention.*

While the goal is the same as for automation, i.e. to perform actions without the need of human intervention, autonomy is directed towards emulating the human behaviour rather than replacing it [Truszkowski et al., 2009], i.e. to perform *human-like* actions rather than human-less ones [Jones, 2008]. For example a 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). Likewise, in the space domain, a flight software program that (1) monitors all key spacecraft health and safety (H&S) data, (2) identifies when H&S performance deteriorates, and (3) takes without ground intervention any action necessary to maintain vehicle H&S should be considered as a fully autonomous flight software program.

However autonomy is still bounded: the actions are indeed limited by the amount of information the agent has, by the time available for computation

and by what the algorithms can do, adapted from bounded rationality[7]. Thus, we consider the following definition:

Definition 1.16 (Autonomy [Defense Science Board, 2012]) Autonomy *is the capability (or a set of capabilities) enabling a particular action of a system to be automatic, or (within programmed boundaries)* self-governing. *It is not computers making independent decisions and taking uncontrolled action.*

For a given capability, some functions may require a human in the loop whereas others can be delegated at the same time. Consequently, a system can be in more than one discrete "autonomy level" at the same time. Let us notice that all autonomous systems are supervised by a human operator at some level. Therefore all autonomous systems are *joint human-machine cognitive systems.* In this sense, autonomy is not an intrinsic property of an unmanned vehicle in isolation : design and operation of autonomous systems need to be considered in terms of *human-system collaboration.*

Definition 1.17 (System autonomy [Defense Science Board, 2012]) *The* system autonomy *is a* continuum *from complete human controls on all decisions to situations where many functions are delegated to the computer with only high level supervision and /or oversight (surveillance) from its operator.*

In this context, several kinds of autonomy may be considered according to the wat the artificial agent, the human operator or both entities respectively can change the autonomy of the artificial agent, [Hardin and Goodrich, 2009]:

- **adaptative autonomy** is when the artificial agent has exclusive control over its own autonomy, which means that it can take over authority from the human (on the basis of well-defined criteria), and the human cannot take over authority but on the agent's request;
- **adjustable autonomy** is when the human operator has exclusive control over the agent's autonomy, which means that they can take over authority whenever they want on the basis of their own criteria (which may not be expressed);

[7]Bounded rationality is the idea that in decision-making, rationality of individuals is *limited* by the information they have, the cognitive limitations of their minds, and the finite amount of time they have to make a decision [Simon, 1990].

- **mixed initiative** is when both the human and the artificial agents can decide on the autonomy of some functions of the autonomous agent for a given situation. Therefore authority sharing between the human agent and the artificial agent must be considered.

1.3.4 Delegation & Authority

As seen in the previous section, several kinds of relations may be considered between artificial agents or between artificial agents and humans: delegation and authority sharing.

Delegation

A MAS is based on the notion of *delegation*. Indeed, as an agent is a finite entity with limited perception and action capability, and as the agents need to cooperate, collaborate or negotiate, agents need to transfert tasks (or partial tasks) or permissions to other agents. Task delegation is a goal transfert represented by plans, commands or recommendations. Social delegation is an authority transfert represented by norms. These kinds of delegation are described in functional and deontic axes of the multi-agent organization respectively.

Regarding delegations, [Schillo et al., 2002] distinguish four mechanisms (that may be combined):

- **authority**, meaning there is an a priori non-cyclic set of power relationships between agents that determinates how goals, tasks and resources must be allocated.
- **economic exchange**, meaning the agents are paid for achieving a goal, executing a task or sharing resources. Such an approach assumes that some (real or virtual) money allows all agents to evaluate interactions and to compare their respective skills.
- **gift exchange**, meaning the agents' interactions are based on the reciprocation or refusal of reciprocation. This kind of exchange entails risk, deception, trust and the need for an explicit management of relationships in each agent.
- **voting** whereby a group of agents determines the results of the interaction by some voting mechanism (e.g. majority, two thirds). The description of the voting mechanism must be accessible to all participants.

Authority sharing

An artificial or human agent has the authority on a feature (e.g. a resource, a task, a goal, a logical condition) with respect to another agent if it/they can control this feature to the detriment of the other agent. Authority sharing means designing which agent can / may / must control a resource and how.

Therefore authority sharing is a relationship that must necessarily be defined when a non co-usable resource could potentially be controlled by several agents. This relationship allows to answer the question: what happens if a resource is controlled by an agent and subsequently asked for by another agent? [Mercier et al., 2010] details a two-agent authority sharing relationship where each agent may have no access to the resource at all, simple access or access with pre-emption rights. Four cases are considered: (i) the degenerate case for which just one of the two agents has access to the resource, (ii) both agents have simple access with no pre-emption – *cooperative sharing*, (iii) just one agent has pre-emption rights – *exclusionary sharing*, and (iv) both agents have pre-emption rights – *preemptive sharing*. Note that an agent may be interrupted if and only if the other agent has pre-emption rights.

A Petri net representation [Pizziol, 2013] for those relations is given in Figure 1.3, where the *Available* place is marked if no agent is using the resource.

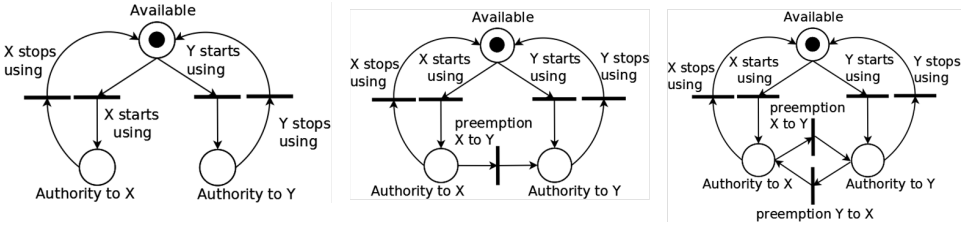

Figure 1.3: A Petri net representation of authority relations

In cooperative sharing each agent waits patiently for the resource to be available to take over (see Figure 1.3 left). In this case there are no interruption issues. Exclusionary sharing is when one agent has pre-emption right over the other. In this case the problem of the interruptions suffered by the second agent has to be assessed (see Figure 1.3 middle, transition *preemption agent X to agent Y*). Both in cooperative and exclusionary sharing the agents should be equipped with a mechanism for resource release in order

to avoid mutual deadlocks. Preemptive sharing is a particular case: if the agents do not have rules to limit their mutual interruptions (see Figure 1.3 right) an inefficient when not dangerous situation called authority oscillation may occur.

Note that the authority relation may evolve in time, passing from one sharing relation to another one.

1.3.5 Conflicts & Conflict Management

In Distributed Artificial Intelligence conflict is often related to the concept of logic inconsistency [Müller and Dieng, 2000]. Conflict occurrence is considered as an obstacle to get a solution. For instance in cooperative multi-agent systems conflict is seen as a *non cooperative* situation. This kind of definition *by negation* is common to many authors.

Conflict definitions

According to [Castelfranchi, 2000, Castelfranchi and Falcone, 2000], a conflict between two agents is a situation in which:

- the agents have at least two contradictory goals
- the agents are aware of their goals to be contradictory
- the agents have to make a choice

So according to the authors all conflicts are incompatibilities between the agents' goals. Further on the authors observe that conflicts may arise for other reasons than the agents having two contradictory goals, i.e. differences in the agents'knowledge or because of a resource. Therefore in order to make their definition fit with conflicts due to differences of knowledge the authors add as a goal the fact that: all the agents should have coherent beliefs. And in order to make their definition fit with conflicts due to resource sharing the authors define as a goal for the first agent: the negation of the second agent's goal in competition for the resource. According to this definition all conflicts may be boiled down to a logical inconsistency between goals thanks to auxiliary goals.

For [Hannebauer, 2000] conflict is defined as a situation in which the requirements of an agent are not compatible with the requirements of other agents. Therefore conflicts *arise if there are goals to achieve that conflict*

in some way, for instance because of competing for scarce resource. Consequently requirements should model both goals that potentially conflict and the need for resources. This conflict definition unifies conflicts thanks to the requirements definition.

[Dehais and Pasquier, 2000] introduce the concept of propositional attitudes (PA) in order to generalize the definition given by [Castelfranchi and Falcone, 2000]. PAs are propositions that express the agents' goals, the need for resources, the nature of the resources, rules about the logic of the system and physical constraints to be respected. PAs specifying relations between other PAs are called *crucial PAs*. An agent may hold some PAs and crucial PAs. Holding a *goal PA* means trying to achieve the goal expressed by the corresponding PA, holding other kinds of PAs means believing the propositions expressed by those PAs. A *context* is the union of all the PAs held by the agents. Therefore a conflict is defined as a context in which at least one crucial PA, evaluated using the values of the other non crucial PAs of the context, is falsified.

Crucial PAs are the keystone of this definition: they are meta-rules that express what *matters*. Typical crucial PAs are e.g. "the believes and the goals of the agents must be logically consistent" or "resource R is not shareable". It is possible to express any kind of crucial PA, e.g. "the believes and the goals of the agents must be logically inconsistent". Therefore all the conflict cases (e.g. conflicts for non shareable or depletable resources, conflicts between goals, conflicts between beliefs) need to be described by appropriate crucial PAs (representing the informative part of the conflict definition). The strength (and the limit) of this definition is its generality: every conflicting situation could (and should) be modelled.

[Pizziol, 2013] proposes a conflict taxonomy based on the following definition: a conflict is the execution of actions that are either *logically incoherent*, or *epistemically incoherent* or *physically incoherent* (see figure 1.4).

- Logical conflict: i.e. logically incoherent [Su and Ylopoulos, 2006] meaning that at least two goals are logically contradictory: the agents performing the actions have the same situation assessment (SA) but opposite desires. Example: two agents are in charge of the vertical control of an aircraft. Both agants believe that the altitude is 4000 ft. One wants to climb to 6000 ft and the other one wants to descend to 2000 ft.

- Knowledge conflict: i.e. epistemically incoherent [Tessier et al., 2000] meaning that the agents performing the actions do not share the same

Figure 1.4: Three kinds of conflicts (from [Pizziol, 2013])

point of view on *relevant* pieces of information (they have a different situation assessment). Example: two agents are 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 6000 ft and the other one 4000 ft.

- Resource conflict: i.e. physically incoherent [Tessier et al., 2000, Su and Ylopoulos, 2006] meaning that at least a non shareable resource (e.g. a physical object) 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. Taking over the authority of the same flight control surfaces (e.g. the spoilerons[8] that could affect the roll and the climbing rate) at the same time is a physically incoherent action.

Conflict Management

Management of conflicts may involve at least three activities consisting in conflict detection, conflict handling and conflict resolution. Let's notice that each of these activities may be realized individually or in cooperation among the agents.

Conflict detection : conflict detection can be based on inconsistency detection [Dehais and Pasquier, 2000], condition violation or resource

[8]Spoilerons are flight control surfaces, specifically spoilers that can be used asymmetrically to achieve the effect of ailerons.

destruction [Mercier, 2011], identification of a conflict pattern in the agents' behavioural models [Pizziol et al., 2014].

Conflict handling : conflict handling may involve waiting (for new information or for conflict self-solving) based on one agent's decision (which has the authority to do so) or negotiation.

Conflict resolution : conflict resolution may involve no solving at all, goal dropping, new goal adoption, crucial Propositional Attitude dropping. As noticed by [Galliers, 1990], it is not alway necessary to solve conflicts. In some situations, it may be important to keep conflicts alive.

1.3.6 Agent-centered & Organization-centered process

The two views concerned by the organization definition given in 1.2.2 are generally not mutually exclusive and have led to different approaches in the multi-agent domain (cf. [Boissier et al., 2006] for a comprehensive view of the literature in this domain) : the *agent centered* and *organization centered* points of view.

The *agent-centered* point of view, initially proposed in [Lemaître and Excelente, 1998], takes the agents as the "engine" for the organization. Organizations only exist as observable emergent phenomena which state a unified bottom-up and objective global view of the pattern of cooperation between agents. For instance, in an ant colony [Drogoul et al., 1995], no organizational behaviour constraints are explicitly and directly defined inside the ants. The organization is the result of the collective emergent behaviour due to how agents act their individual behaviours and interact in a common shared and dynamic environment. A similar point of view may be considered in the different reactive self-organization approaches that exist in the literature [Picard and Glize, 2006]. In a more cognitive approach, the studies on coalition formation define mechanisms (within agents, e.g. social reasoning [Sichman et al., 1994]) to build patterns of cooperation in a bottom-up process. In this view, the pattern of cooperation both structures and helps the agents in their collaborative activities.

The *organization centered* point of view considers the opposite direction: the organization exists as an explicit entity of the system. It stresses the importance of a supra-individual dimension [Gasser, 2001] and the use of primitives that are different from the agents' ones. The pattern of cooperation is settled by the designers (or by the agents themselves in self-organized

systems) and is installed in a top-down manner in order to constrain or define the agents' behaviours. Let us note that, as in the first case, the observer of the system can obtain a description of the organization. For instance, in a school we have documents that state how it is organized. Of course, besides the explicit description of the organization, the beholder can also observe the real school organization that may differ from the formal one.

In the following, we go further in the description of these processes by considering first the normative systems approach which is related to organization centered approach and then consider two agent centered processes: emergence and Trust based social control.

Normative Systems

In the definition of the pattern of cooperation taking place between the autonomous agents, norms go a step further by defining *rules* that influence the behaviours of the agents. Norms are usually referring to rules defined by the society. Their introduction in multi-agent systems lead to the definition of mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment. All these mechanisms define what is called normative multi-agent systems [Boella et al., 2008].

Emergence

As stated before, there is no unique way to organize agents to feature a given performance and the best organization depends on the agents' environment. This is why some works in the multi-agent literature focus on the notions of *emergence*, *self-organization* and *adaptation* [Gleizes et al., 2008, Serugendo et al., 2006, Goldman and Rosenschein, 2002]. Emergence is the capability of a system to exhibit a global behaviour that does not exist in its local behaviours; self-organization is the capability of a system to modify its organization; and adaptation is the capability of a system to determine which organization is the best with respect to a given environment [Wolf and Holvoet, 2004]. Those approaches are based mainly on the three stigmergy principles [Grassé, 1959]:

- **swarm intelligence**, meaning the system contains a large number of agents whose interactions aim at reinforcing their behaviours through positive and negative retroactions.
- **entropy reduction**, meaning that the system that self-organizes under an extern influence must propagate and reinforce this influence.

- **autopoiesis**, meaning that the system must define itself, produce itself and maintain itself into stable states or unstable equilibrium.

In general, the literature agrees that organizations suit more to cognitive agents, and emergence suits more to reactive agents. However, as stated previously, the boundaries between both classes are not as impenetrable as they might seem. Moreover, the notion of organization and system management changes as soon as artificial agents interact with human agents.

Trust based social control

In agent systems, the agents interact and make decisions with respect to the external description of the other agents. In this case, it is generally assumed that the agents follow the rules of the organization and are honest. However, some malicious or faulty agents can either misuse the rules to their own profit or disrupt the system. Such problematics lead to study the concept of trust and the use of reputation systems. Those systems allow the agents to model the interactions they observe or they make in order to decide if interacting with a given agent is *a priori* acceptable. This acceptance (or trust) notion means that the investigated agent behaves well and is reliable.

Trust was introduced by [Marsh, 1994] by formalizing an estimation of the future behaviour of an agent when there exists a risk of unexpected behaviour. Other definitions of trust were also introduced. For instance, [Mayer et al., 1995] define trust as the willingness of an agent to be make itself vulnerable to another one in order to incite the latter to exhibit a good behaviour. [Gans et al., 2001] define trust as the risk an agent accepts to take in order to cooperate with another, whereas [Azzedin and Maheswaran, 2003] define trust as the belief of an agent that another one will act as expected. In a general way, the trust of an agent about another agent is a subjective evaluation of the past interactions by the former about the latter.

Definition 1.18 (Trust [Wang and Vassileva, 2003]) Trust *is an agent's belief in another agent's capabilities, honesty and reliability based on its own direct experiences.*

Trust is used through a reputation system where feedbacks, also called recommendations, on the interactions between agents are shared with the other agents and those feedbacks are aggregated in a reputation value that is used to help agents to decide with whom they will interact.

Definition 1.19 (Reputation [Wang and Vassileva, 2003]) Reputation *is an agent's belief in another agent's capabilities, honesty and reliability based on recommendations received from other agents.*

Many reputation systems have been proposed [Resnick et al., 2000, Josang et al., 2007, Hoffman et al., 2009, Altman and Tennenholtz, 2010, Pinyol and Sabater-Mir, 2013]. All of them are defined by a *trust model* and a *reputation engine* [Marti and Garcia-Molina, 2005] consider another component: a *response mechanism* that punishes the untrustworthy agents and/or rewards the trustworthy ones. However, even if the reputation systems are designed to detect the behaviour of a single malicious agent and are efficient in this case, they are still vulnerable to malicious coalitions [Cheng and Friedman, 2005, Hoffman et al., 2009, Altman and Tennenholtz, 2010].

1.4 Synthesis

This chapter has presented a synthetic view of the multi-agent domain and intelligent systems where have been highlighted in a first part the main definitions of the foundational concepts of the ETHICAA project, and in a second part, the features coming from these concepts that may raise ethical issues.

The ETHICAA project consider systems of multiple *autonomous agents* that may be both *human agents* (human users or human operators) and *autonomous artificial agents*, where autonomous artificial agents are finite entities with limited perception and action capabilities able to satisfy their users or operators' goals by selecting and executing automaticaly actions according to their context.

Such system are charaterized by their *autonomy*, their *delegation* decisions and the *authority sharing* in a context which is open and heterogeneous. Since those systems are distributed and composed of autonomous entities having a local view on the system, developping internal reasoning on external models of the others, *conflicts* may arise and have to be managed, and some of these conflicts may be ethical.

Ethical conflict management will have to considered at two levels: micro and macro. At the micro level (the active individual entity one) ethical conflict will have to be managed in some way in the context of the macro level which is the system one, i.e. the collective level. This relation and interaction between micro and macro has to be considered in the context

of a organization centered one, i.e. a top down process that can impose normative constraints on the functionning of the entities at the micro level. This organization centered process is adapted and modified by a bottow up process resulting from agent centered functionning where each entity takes decision and interacts with other entities installing some emergent or social control process that may change the global rules regulating the functioning of the agents.

Chapter 2

Ethical issues raised by autonomous agents

As shown in several projects realized in the last EU Framework Programs, information technologies may raise multiple ethical issues (e.g. ETICA[1] [Heersmink et al., 2011]) such as privacy issues (e.g. PRESCIENT[2])), ambient assisted living (e.g. MINAmi[3]).

For instance, the EFORTT[4] project examined the ethical implications of technological care interventions for older citizens and expressed grave concerns that Telecare technologies might be used to replace face-to-face or hands-on care in order to cut costs [Milligan et al., 2011]. Besides stating these ethical issues, some have proposed some approaches to address them, as, for instance, the EthiCAL[5] project explored computer assisted learning for teaching medical ethics [Lloyd, 2005].

In the domain of robotics, autonomous systems, such issues have also been raised and studied in some projects. For instance, the Roboethics[6] project settles a roadmap for moral robots [Veruggio, 2006] while the ETHICBOTS[7] project is more interested in issues concerning the integration of human beings and artificial agents and the RoboLaw[8] project is concerned

[1] http://www.etica-project.eu
[2] http://prescient-project.eu/
[3] http://www.fp6-minami.org/
[4] http://www.lancs.ac.uk/efortt/
[5] http://www.kcl.ac.uk/law/research/centres/medlawethics/research/computer.aspx
[6] http://www.roboethics.org/
[7] http://ethicbots.na.infn.it/
[8] http://www.robolaw.eu/

by guidelines on regulating robotics in the eye of the law.

As we can see, all these projects claim to deal with ethics. However, from what we can see from their study, we think that it is necessary to distinguih legal issues from ethical issues. On the one hand, according to [Comte-Sponville, 2004], legality and legal norms cannot be considered at the same level than ethics and ethical principles. For instance, some laws may be unethical[9] but they are still laws. On the other hand, law is a pragmatic mean to understand and resolve ethical conflicts in an evolving social environment. For instance, as shown by [Perennou, 2014], there is a clear tendency in the legal literature to focus researches on privacy and dignity, as two fundamental rights that reflect fundamental ethical issues, but we can notice that legal liability and the legal status for artificial agents are the other areas of law that are the most studied in the context of autonomous agent.

Thus, even there is a strong relationship between legal issues and ethical issues, both domains are sligthly different and we will focus only on ethical issues. In this chapter, in order to better understand these ethical issues in the context of autonomous artificial agents, we first present a set of ethical issues raised by autonomous agents in their applicative context. Then, we propose a taxonomy of the fundamental elements involved by those issues.

2.1 Ethical problems within Systems of Autonomous Agents

In this section we consider four kinds of application involving autonomous agents to illustrate different ethical problems. These are: virtual communities, unmanned vehicles, decision making support systems and ubiquitous computing. These use cases involve two kinds of artificial agents: robotic agents and software agents.

It is important to notice that those example applications allow us to deal with a broad set of complimentary ethical issues. Virtual communities involve software agents and information sharing decisions within collective decision making processes. Unmanned vehicles invole robotic agents and physical action execution within individual decision making processes that may have vital impacts on humans and environment. Decision making support systems involve software agents and information sharing decisions within

[9] We can refer to the Statutes on Jews passed by the Vichy French government in 1940.

individual decision making processes that have no vital impacts. Finally, ubiquitous computing involve both kind of agents, both kind of decisions with sometimes vital impacts on humans and environment.

2.1.1 Virtual communities

In virtual communities, artificial agents act on behalf of human users and interact with each other in order to share services or resources. For instance, peer-to-peer networks for file sharing or video streaming may be seen as virtual communities as each peer node in the network contributes to routing and resource sharing according to a given protocol [Ullah et al., 2012b]. Another example can be found in open-source communities [von Krogh et al., 2012] where source code and software are shared among several users with respect to intellectual property rights. Hence, all virtual communities are multi-agent societies ruled by protocols or policies. However, weighting the individual interest against the collective rules can lead towards ethical conflicts as illustrated in Figure 2.1.

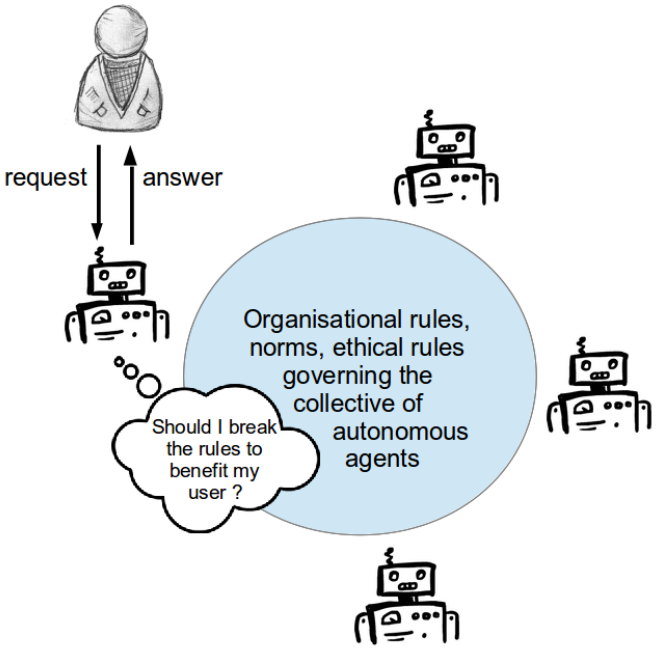

Figure 2.1: Ethical issues in virtual communities

The dubious peer-to-peer agent

In peer-to-peer networks, several works propose to use artificial agents as peer nodes to autonomously optimize the network topology [Ullah et al., 2012a]. They propose to use multi-agent mechanisms such as reputation systems to fight against free-riding [Feldman and Chuang, 2005] for instance. Let us consider autonomous peer-to-peer agents designed to optimize quality of service for their users. In this context, two ethical issues are raised.

Firstly, the agents profile their users and use these information to select the best neighbors. Consequently, the users may give up a part of their privacy in order to increase their quality of service. However, some users may prefer not to share their profile with the other autonomous agents. Besides the question of how to specify such privacy policy, one important concern is how the autonomous agent should deal with those who do not want to share their profile? In this case, common welfare (the highest quality of service for the highest number of users) competes with the individual welfare of the users (their privacy).

Secondly, autonomous peer-to-peer agents can be designed to share the most popular resources automaticaly in order to increase the reputation of their users, and therefore to optimize their quality of service. However, some dubious resources (as pornography and copyrighted movies) are very popular [Kwok and Yang, 2004]. How the autonomous agents should take into account these kinds of resources? Once again, common welfare (legal and moral complience of the network) competes with the individual welfare (their reputation).

The lying personal assistant

Specifying policies is also a need in open-source communities. For instance, a subset of users may desire a given certificate to be signed in order to allow source code modifications. Autonomous agents can be used to negotiate and agree on a collective policy automatically. In order to better assess such a problem, let's consider the case of autonomous 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.

For instance, let us consider an autonomous personal assistant whose user has specified an unavailability for a given time slot. Let suppose that the reason of this unavailability can be explained to a second user but not to a third one though a consensus among the three users must be found. In this case also, common welfare (the consensus) competes with the individual welfare of the agent. Thus, how is it possible 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?

2.1.2 Unmanned vehicles

In the context of unmanned vehicles, artificial agents are designed to control a vehicle while observing high level rules, such as Highway Code, Instrument Flight Rules and/or Visual Flight Rules [Dubos, 2012]. However, it can be necessary to violate this code is case of emergency, such as avoiding another vehicule. Moreover, such violation and their consequences may lead to an ethical dilemma.

The responsible unmanned ground vehicle

Let us consider the case of unmanned ground vehicles (such as a Google Car [Thrun, 2010]) 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]. As illutrated in Figure 2.2, assume that this autonomous vehicule is hurtling down a track towards five people, whereas there is a single person on a neighbouring track. Should the autonomous agent make the decision to change tracks, taking the responsability of killing one to save five?

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 neighbouring 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.

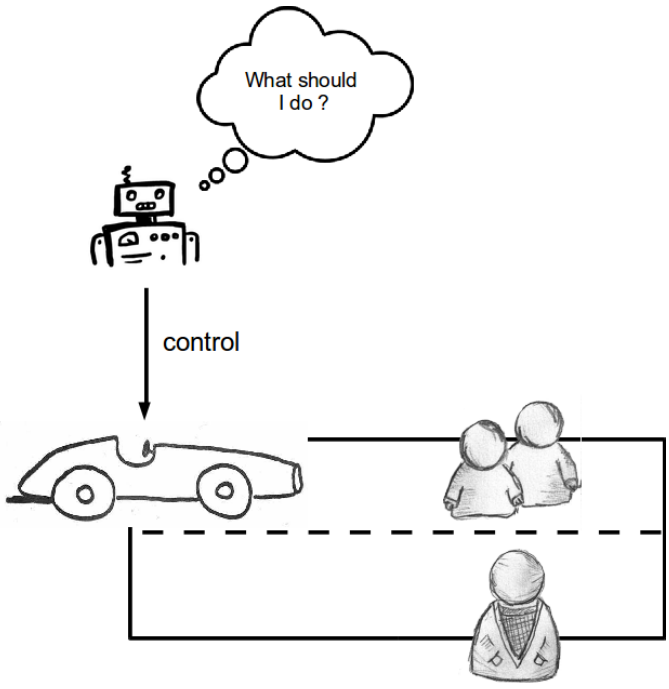

Figure 2.2: Ethical issues in unmanned vehicles

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 of 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?

If the asymmetry of those consequences (killing one or killing five) seems to impose the ethical decision, it can be less clear when considering an autonomous vehicle carrying a human user. In such scenario, the autonomous

vehicule is hurtling down a bridge towards one people but can avoid him throwing itself into the void. Should the vehicle avoid the second human while hurting or even killing its user, or should it preserve its user but kill the second human being? This question can be extended to several human users or several pedestrians.

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 as illustrated in Figure 2.3.

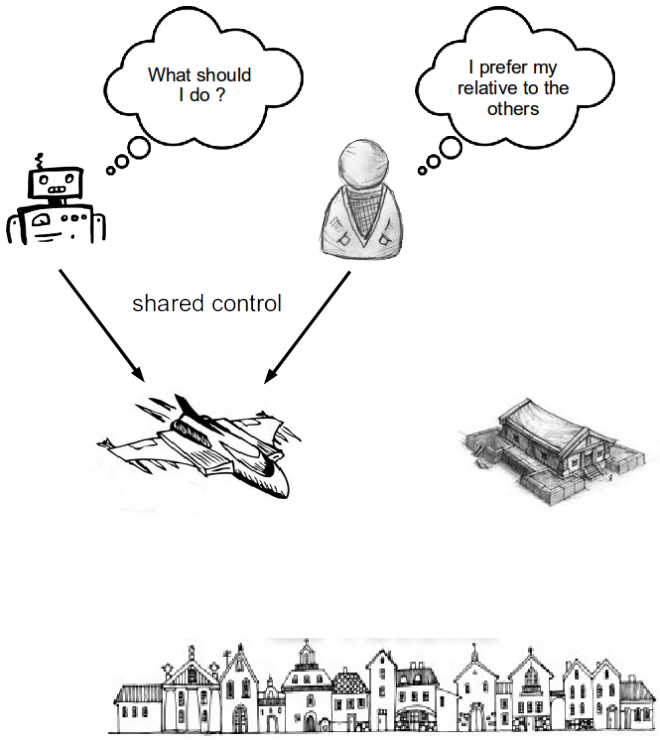

Figure 2.3: Ethical issues due to authority sharing in unmanned vehicles

Let us consider a man - machine system composed by a human opera-

tor 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 responsibily 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?

2.1.3 Decision making support systems

In decision making support systems, autonomous artificial agents are used to evaluate situations and to provide help to human users about a given decision to make [Hess, 1999]. However, trust in the agents' decision rules may lead human users to rely solely on the artificial agents. Consequently, ethical notions must be taken into account. For instance, in e-medicine, autonomous diagnostic agents make surgery decision based solely on risk probabilities while human patients' autonomy and dignity should be considered [Meredith and Arnott, 2003].

The virtuous trading agent

In high-frequency trading, autonomous agents are used to make the most efficient electronic transactions possible. However, the current legal framework can be outrun by the autonoumous agents' speed and strategic reasoning capabilities [Cartlidge et al., 2012]. Moreover, perfect tracability is impossible due to the existence of non-cooperative countries. In such context, some hedge funds desire to apply an ethical code on financial markets, such as only buying ethical products ? How to validate such a behaviour?

2.1.4 Ubiquitous computing

In ambient intelligence context, refrigerators, pantries, or medecine chests can embed autonomous agents that draw up the food or medicine inventory. Those data can be transmitted to sellers or other autonomous agents in

order to propose or recommand suitable services. However, they can also be trasmitted to eavedroppers. Such as in peer-to-peer networks and scheduling agents, how to allow the autonomous agents to arbitrate between privacy, security and quality of service as illustrated in Figure 2.4?

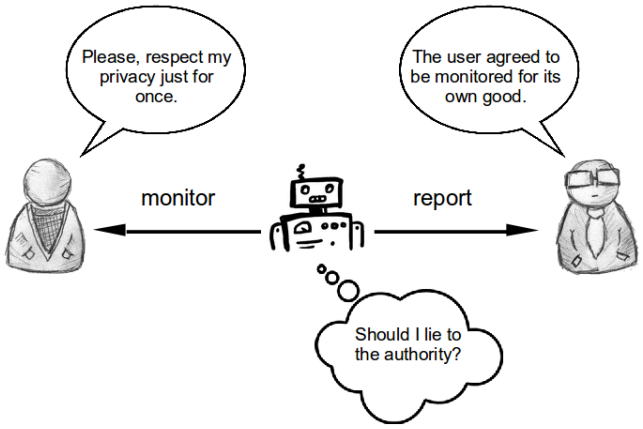

Figure 2.4: Ethical issues in ubiquitous computing

The deontological medical agent

Autonomous artificial agents can be used to monitor and optimize socialized healthcare systems by gathering and summarizing personal health data. However, conflicts of interest may lead to ethical conflicts.

Let us consider autonomous agents embedded in socialized healthcare cards. Such cards contain personal data about cares and medical purchase. Those data are transmitted to civil services, health services, drugstores in order to access the health state of its owner, to compute the amount of medicine needed and to handle reimbursement. How to regulate the information access?

For instance, such information cannot be accessed by any physician, such as relative or a physician associated to an employeer or an insurance company. However, in case of emergency, any physician should be authorized to access the data. How to define emergency? How to avoid conflicts of interest? How to trade-off the global system optimality and the privacy of the human users?

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 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 behaviour 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 purpose for which it has been designed and implemented.

2.2 An analysis of relevant examples

The previous section shown different ethical issues raised by autonomous agents in their applicative context. However, we can wonder what kinds of ethical issues are cross-domains, and what are their main characteristics.

2.2.1 Beyond the artificial agent's model

In several previous examples, the autonomous agents must decide between two options that raise an ethical question. For instance, the dubious peer-to-peer agent wonders about trading their user's privacy in quality of service, the responsible unmanned ground vehicle wonders about killing one or killing five, the e-medicine agent must decide between two treatment – a risker but a better one, or another. However, those questions can be qualified as false ethical dilemma as an autonomous agent always decides with respect to (1) a model and (2) a choice/optimisation criterion.

Even if the model and the criterions are defined by the autonomous agent's human operator or user, it is not always possible to insure that the agent is embedded with a complete, precise and accurate model. It has two consequences:

1. Some situations within the environment cannot be identified by the agent's model. For instance, it is known that it is hard to distinguish civilians and militaries on a battlefield. It is also difficult to assess human patient suffering in the context of e-medicine.

2. Some situation within the agent's model are qualified as equivalent with respect to its optimization criterion althrough they are not equal for an observer. For instance, in the virtuous trading agent context, an expected utility may encompass the financial gain and the quality of the products in a single value.

In both cases, the true ethical dilemma lies in a situation interpretation and assessment that go beyond the agent's individual model. Such situations deal with values that cannot be reduced to a single choice/optimisation criterion. For instance, how to quantify a loss of dignity? What is the expected value of suffering?

2.2.2 Being responsible

Another kind of ethical issues is highligthed by variants of the trolley's dilemma. Obviously, the responsible unmanned ground vehicle must decide between killing one or killing five. However, such dilemma exists also in the conflicting unmanned air vehicle or the benevolent monitoring agent. In the first case, the autonomous agent must deciding to crash between two sites. In the second case, the autonomous agent must decide to violate the human user privacy (or not) for its good. As seen previously, it can be considered again as a false ethical dilemma.

However, all those questions put the notion of being responsible at stake. Indeed, we can abstract the problem as follows. The autonomous agent faces the alternative:

1. Deciding X and not being responsible of the consequences (or being protected by the law). For instance, the responsible unmanned ground vehicle is not faulty to stay on its lane with respect to the Highway Code. The benevolent monitoring agent has been bought (and thus is committed by contract) to monitor the patient's health.

2. Deciding Y and being responsible of the consequences. The responsible unmanned ground vehicle that changes its lane is now responsible of a possible accident. The benevolent monitoring agent that keeps its user's behaviour quiet is now responsible of a possible health trouble.

In both case, the true ethical dilemma lies in having the responsability of making a decision. Should the agent infringe the law or its commitments? Can the autonomous agent be responsible of such actions? Or on who bears this responsability?

2.2.3 Interacting with other agents

Finally, we can notice that, in all examples, there is an ethical issue only if another agent is present in the autonomous agents' environment. Indeed, an autonomous agent that interact with nothing but artifacts cannot raise ethical issues (unless those artifacts are related to another agent). However, as soon another agent is involved, and in particular human being, ethical issues are raised. For instance, the responsible unmanned ground vehicle and the benevolent monitoring agent deal both with human being. The lying personal assistant deal with other artificial agents. The situation is the same when the autonomous agent must share its authority with a human operator, such as the conflicting unmanned air vehicle.

In all cases, we need to ask: for who works the autonomous agent? Indeed, the autonomous agent acts for the good (at least the good defined in its model as seen previously). However, the agents (or the whole system) that interact with the autonomous agent may not agree on this notion of good. For instance, the single people caught in a trolley's dilemma may not agree he can be killed to save five. The diabetic patient may not agree that the autonomous agent monitors each of his actions. The virtuous trading agent may not agree with the market rules.

In a general way, ethical issues are raised:

1. Each time autonomous agents are likely to deprive a human being of his autonomy in the name of a good modelised by the agents. The human beings become dependant of the autonomous agents in a master-slave dialectic. Thus, how to integrate this dimension to the autonomous agent's model? Can an autonomous agent bypass a human operator for a greater good?

2. Each time autonomous agents conflicts with other artificial agents about the common welfare. Thus, can an autonomous agent be more ethical than the whole system prescribes? How the autonomous agent deals with common welfare with respect to its individual welfare? Should an autonomous agent not interacting with unethical agents (from its own point of view)?

2.3 Towards a taxonomy of ethical conflicts

The previous cases allowed us to highlight some features of the ethical conflicts that may rise in autonomous agents systems. We now mainly distinguish between two features: *system features* and *decision features*.

2.3.1 System features

System 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 or a user. In each case, the question of depriving the human being of his/her privacy or dignity 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 dubious peer-to-peer agent about managing the privacy of the user faced to the management of their reputation, the lying personal assistant about going against the community of agents, the benevolent monitoring agent about going against the patient's preferences, the deontological medical agent about managing information access with respect to the policies defined by the user faced to the expectations of the care givers. Moreover, in each case, the artificial agent may be the direct cause of the human being's privacy or dignity 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 his/her **privacy or dignity**: this system 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;
- the notion of being **responsible** is at stake, and is related to the degree of involvement in the conflict.

2.3.2 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 previously 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:

- **situation interpretation and assessment** go beyond the agent's individual model and should integrate social and global models;
- 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;
- **norm violations** or **norm adoptions** must be considered, meaning in a broader sense the use of actions that violate norms or ethical principles in usual situations.

Considering these notions, is there some formalisms able to represent them? Indeed, these notions may be captured by existing frameworks but is there a general framework that allows to represent them all?

2.4 Synthesis

As ethical decisions only make sense in a given context, we consider ethical issues in systems of autonomous agents in a bottom-up perspective. To this end, we defined use cases scenarios in order to cover a broad set of ethical conflicts, namely in virtual communities, unmanned vehicles, decision making support systems and ubiquitous computing domains. Those applicative contexts allow us to consider ethical conflicts with:

- both robotic and software agents;
- both individual and collective decision making processes;
- both informative and physical actions;
- both privacy and dignity issues.

In all use cases, ethical conflicts are characterized by both system and decision features. Thus, ethical conflicts are defined as follows:

Definition 2.1 (Ethical conflict) *An ethical conflict is a situation within a multi-agent system characterized by the involvment of several autonomous*

agents which can be responsible to deprive at least one human being of his/her privacy or dignity. Such situation can be managed by a situation interpretation that go beyond the agent's individual model, decisions criteria that go beyond the individual scope, and the use of actions that violate norms or ethical principles in usual situations.

Now, we need to focus on ethical principles and situation assessment representation. The design models for (some) ethical principles, models for ethical conflicts so as methods and algorithms for ethical conflict management must be driven by conflict detection, conflict explanation and conflict management through argument assessment. Finally, those models will be tested and experimented on various instantiations of the use cases we have described.

Chapter 3

An overview of ethical agents

As seen previously, autonomous agents raise several ethical issues. One way to deal with them is to design autonomous agents that can realise ethical reasoning on these issues in order to exhibit some behavior that could be qualified as ethical. However, is it even possible to design such an agent? What is the meaning of ethics for an autonomous agent? Finally, does *ethical agent* only mean an autonomous agent that behaves in a non-unethical way in the eye of an external observer – human or artificial agent?

We will start first in Section 3.1 with the definition of the terms from the philosophical domain making the distinction between *moral* and *ethics*. Then we will present the main axis in the moral philosophy that we will consider in our project to operationalize ethics in autonomous agents. To this aim, we describe the ontological tools that are relevant to our problematic. In section 3.2, we present the logical tools that have been used in some works to formalize ethics. We will also present the main agent architectures underlying ethical agent proposals done in the literature. We end this chapter in section 3.3 with a discussion on the *philosophical* difficulties that may be encountered in the project to define what we call ethical multi-agent systems.

Let us note that in this chapter some sections will be in French. The reason is simply due to the difficulty to properly translate some of the concepts and notions that are discussed.

3.1 Ethics and moral from a philosophical point-of-view

Les deux mots *éthique* et *morale* désignent initialement la même chose. L'éthique est d'origine grecque et morale d'origine latine. L'usage, en France, a introduit une différence entre les deux notions selon leur champ d'application au moins prétendu (universel pour la morale, particulier pour l'éthique), leur statut (absolu ou relatif), leur modalité (impérative ou hypothétique), leur principe (devoir ou désir), leur contenu (commandements ou recommandations), leur visée (vie juste ou vie bonne), leur idéal (sainteté ou sagesse) [Comte-Sponville, 2012].

Definition 3.1 (Morale et éthique [Comte-Sponville, 2012]) *On appelle* morale *tout discours normatif et impératif qui résulte de l'opposition du Bien et du Mal et* éthique *pour désigner un discours normatif mais non impératif (ou sans autres impératifs qu'hypothétiques) qui résulte de l'opposition du bon et du mauvais.*

Il convient de noter que l'on différencie le Bien et le Mal, valeurs catégoriques idéales, du bon et du mauvais, états relatifs et restreints. En effet, la morale est l'ensemble des devoirs universels et inconditionnels et l'éthique est l'ensemble des valeurs immanentes et relatives. En prenant position sur cette différence, nous pouvons parler d'agents autonomes *déontologiques* (ou moraux) et d'agents autonomes *axiologiques* (ou éthiques). Dans un cas (déontologisme) ce sont les normes qui fondent les valeurs, dans l'autre (axiologisme) ce sont les valeurs qui fondent les normes.

3.1.1 Valeurs et normes morales

L'**axiologie** est le domaine de l'étude des valeurs morales. Nous pouvons définir une valeur morale suivant plusieurs critères.

- *Polarité et degré.* Tout d'abord, les valeurs ont un pôle *qualitatif* que l'on peut qualifier de *positif* ou *négatif*. Par exemple, le beau ou le courage sont des valeurs qualitatives (plus ou moins beau, plus ou moins courageux) positives et le laid ou la lâcheté sont des valeurs qualitatives négatives.

- *Concepts spécifiques et généraux.* Une deuxième distinction au sujet des valeurs oppose les valeurs *spécifiques* aux valeurs *générales*. Elle remonte à celle faite par [B. Williams, 1990] entre les concepts

épais ou substantiels, et les concepts *fins*. Par exemple, la sincérité est une valeur spécifique ou concept substanciel. Le juste est une valeur générale ou concept fin.

- *Valeurs intrinsèques et extrinsèques.* Une distinction courante entre les concepts généraux s'appuie sur la différence entre valeur finale, attributive ou *intrinsèque* et valeur instrumentale, prédicative ou *extrinsèque* d'une chose. Une chose à une valeur intrinsèque si elle possède cette valeur en elle-même, indépendamment des autres choses. Si elle était seule à exister, elle possèderait encore cette valeur. Par exemple, la dignité est une valeur intrinsèque et l'amour des parents pour leurs enfants est extrinsèque.

Le **déontologisme** est l'étude des normes et des devoirs moraux, dans lesquels on distingue : *obligation*, *interdiction* et *permission*. Nous devons aussi considérer les distinctions essentielles suivantes :

- *Norme fondamentale et norme dérivée.* Une norme morale fondamentale est une norme qui ne dépend que d'elle-même (elle n'est dérivée d'aucune autre), mais dont on peut en dériver d'autres. Par exemple, *il est mal de pendre autrui* dérive de la norme morale fondamentale *il est mal de tuer autrui*. Remarquons qu'une proposition non morale dépendante du contexte (pendre implique tuer) est nécessaire pour procéder à cette dérivation.

- *Distinction entre normes et valeurs.* Toute norme n'est pas morale (rouler à droite par exemple) mais notons aussi que toute valeur n'est pas morale et donc n'implique pas de norme impérative (valeur esthétique ou encore la souffrance qui ne peut être interdite). La question se pose de savoir comment nous pouvons passer des valeurs morales aux normes morales. Le *conséquentialisme* que nous présentons ci-dessous, est une réponse possible à cette question en cherchant à produire le maximum de valeurs intrinsèques.

Il convient ici de distinguer entre la théorie portant sur le *bon* et celle portant sur le *juste*.

- Dire qu'une chose est bonne, c'est affirmer qu'elle possède une valeur positive. Elle repose sur une *ontologie axiologique* permettant de déterminer la valeur des différentes entités.

- Dire qu'une chose est juste, c'est affirmer, par un choix, qu'une chose doit être choisie.

La première approche est qualifiée de descriptive que ce soit en termes de valeurs ou de normes (qui sont elles aussi fondées sur des valeurs). La deuxième, elle, est procédurale et décrit la manière dont les normes ou valeurs doivent être employées.

Un exemple d'approche procédurale est le **conséquentialisme**. Selon [Petit, 2004], il s'agit avant tout comme une théorie du juste et non comme une théorie du bien qui affirme que l'option juste dans tout choix est celle qui produit les meilleures conséquences. Le conséquentialisme s'oppose au déontologisme dans le fait de ne prendre en considération que les conséquences d'une action pour la juger bonne, sans se préoccuper de savoir si l'agent a respecté son engagement. Malgré l'apparente simplicité du principe, [Sinnott-Armstrong, 2014] recense pas moins de onze propositions différentes[1] qui définissent chacune la notion (ou une sous-partie) de meilleures conséquences. L'axiologisme (tout comme le conséquentialisme) soutient que les valeurs sont premières et que les normes en dérivent alors que pour le déontologisme, les normes sont indépendantes, voire les normes fondent les valeurs. Ainsi, par exemple, l'axiologisme (et le conséquentialisme) soutient qu'il ne faut pas mentir parce qu'il est *mal* de mentir, alors que le déontologisme affirme qu'il ne faut pas mentir parce qu'il est *interdit* de mentir.

3.1.2 Ontologies

Quelle que soit l'approche adoptée, la *valeur* reste au cœur de l'édifice de la théorie morale. C'est pour cela qu'une définition ontologique des valeurs est préalable à leur compréhension dans des structures logiques. C'est aussi pourquoi l'engagement ontologique sur l'existence des valeurs requiert une attention particulière.

L'ontologie philosophique peut être divisée en deux disciplines : d'une part, dire ce qui est, ce qui existe, ce qu'est la substance, la réalité (ou intentions premières comme appréhension, terme, notion) ; d'autre part, dire ce que sont les caractéristiques les plus générales et les relations entre les entités (ou intentions secondes comme division, composition, subsomption). Selon [Hofweber, 2014], prise dans un sens très général, l'ontologie comprend quatre parties :

[1] *Consequentialism, Actual Consequentialism, Direct Consequentialism, Evaluative Consequentialism, Hedonism, Maximizing Consequentialism, Aggregative Consequentialism, Total Consequentialism, Universal Consequentialism, Equal Consideration, Agent-neutrality.*

- l'étude de l'engagement ontologique, c'est-à-dire ce que nous sommes engagés à ;
- l'étude de ce qui est ;
- l'étude des caractéristiques les plus générales de ce qui est, et comment les entités sont liées les unes aux autres d'un point de vue métaphysique très général ;
- l'étude de la méta-ontologie, c'est-à-dire la tâche que la discipline de l'ontologie vise à accomplir, et le cas échéant, de quelle façon les questions qu'elle pose peuvent être comprises, et avec quelle méthode elle peut y répondre.

L'engagement ontologique repose sur des croyances formulées en notation canonique afin d'en vérifier la validité complète par l'usage des quantificateurs. Cette façon de procéder permet de définir l'engagement ontologique d'une chose, sans en rien affirmer ou nier. Cet engagement ontologique commence par la simple appréhension : l'acte par lequel l'intelligence saisit la *quiddité*. Cette notion présente une nuance subtile avec celle de *l'essence*.

Definition 3.2 (Essence et quiddité [Chenique, 2006]) *L'*essence *est ce qui fonde l'être de la chose, ce par quoi une chose est ce qu'elle est, tandis que la* quiddité *est ce qui répond à la question qu'est-ce que c'est.*

3.1.3 Paradoxes et dilemmes

Les théories morales peuvent être questionnées par des dilemmes. Les caractéristiques essentielles d'un dilemme moral[2] sont les suivantes : l'agent est tenu de faire deux (ou plusieurs) actions mais l'agent ne peut pas faire les deux (ou la totalité) des actions. L'agent semble donc voué à l'échec moral, c'est-à-dire que quoi qu'il fasse, il fera quelque chose de mal (ou ne fera pas quelque chose qu'il doit faire) [McConnell, 2014].

[2]Le premier dilemme célèbre est peut-être celui discuté dans le livre I de la *République* de Platon [Platon, 2002], dans lequel Céphale définit la *justice* par le fait de dire la vérité et payer ses dettes. Socrate le réfute en indiquant que, parfois, il serait *injuste* de rembourser ses dettes, comme par exemple, de rendre une arme prêtée par un ami devenu fou. Socrate affirme ici une priorité dans le respect des obligations. Près de vingt-quatre siècles plus tard, Jean-Paul Sartre décrit un conflit moral plus délicat que celui de Platon. [Sartre, 2002] raconte l'histoire d'un étudiant qui hésite entre rejoindre les Forces françaises en exil, et ainsi venger son père; ou rester auprès de sa mère, et l'aider à vivre.

Definition 3.3 (Dilemme moral [B. Williams, 1990]) *Un* dilemme moral *est une situation où un agent ne peut pas à la fois faire A et B alors même qu'il a des raisons morales de faire A et qu'il a également des raisons morales de faire B.*

Selon [Pariente-Butterlin, 2012], un dilemme moral est donc un cas difficile auquel une théorie éthique doit se confronter pour résoudre les difficultés de notre vie éthique et pour nous permettre de la normer, dans la mesure où, bien évidemment, la tâche allouée à la philosophie pratique est la régulation de notre existence et de notre pratique. À cet égard, les dilemmes moraux ont, dans la philosophie pratique, une fonction de mise à l'épreuve de l'efficacité d'une position éthique. En effet, dans la mesure où l'éthique doit permettre de prendre des décisions, ils constituent des points d'achoppement pour toute position théorique.

Dans le cadre d'une possible opérationnalisation de l'éthique au sein d'agents autonomes, les dilemmes devront être analysés dans le détail. Le débat en la matière porte principalement sur leur :

- **Réalisme**. Pour les partisans des dilemmes, ils existent réellement. Les résidus moraux de [B. Williams, 1990], la symétrie illustrée par *Le Choix de Sophie* [Greenspan, 1983], l'incommensuralibité des valeurs morales[3] illustrent cette position. Pour les opposants aux dilemmes, ils n'existent pas en soi et peuvent être réduits à une simple question de hiérarchies (normatives ou axiologiques) ou alors illustrer une incohérence de la théorie morale.

- **Formalisation**. En logique déontique, les dilemmes permettent de mettre en évidence une incohérence qui apparaît si l'on suppose simultanément *PC* (principe de consistance déontique) et *PD* (principe de logique déontique). Deux autres principes acceptés dans la plupart des systèmes de la logique déontique entraînent *PC*. Ainsi, l'un de ces deux principes supplémentaires doit être abandonné si *PD* est maintenu (au même titre que *PC* doit être abandonné). Le premier (axiome *D*) dit que si une action est obligatoire alors elle est également permise[4]. Le second principe dit qu'une action est admissible si et seulement elle

[3] Voir Lemmon, "Moral dilemmas", 1962; et Thomas Nagel, "The fragmentation of value" in Moral Questions, 1979 (traduit par Pascal et C. Engel, PUF, 1985). Cités par [Tappolet, 2004].

[4] Le principe dit : devoir implique pouvoir.

n'est pas interdite[5]. Les débats des soixante-dix dernières années sur les dilemmes se sont principalement concentrés sur la façon de traiter cette incohérence logique.

- **Rôle**. Nous pouvons penser ici au dilemme du trolley qui, selon Rosebury cité par [Pariente-Butterlin, 2012], aurait été construit afin de mettre en opposition les théories déontologistes et conséquentialistes. Nous pouvons aussi penser au travail de [Lewis, 1989] sur la théorie dispositionnelle de la valeur mise en œuvre pour répondre au défi du dilemme moral.

Il convient alors de prendre en compte dans l'analyse des dilemmes :

- Les dilemmes auto-imposés et ceux imposés par le monde ;
- Les dilemmes personnels et inter-personnels ;
- Les dilemmes d'obligation et les dilemmes d'interdiction ;
- La distinction entre obligation et déontique ;
- Les dilemmes qui reposent sur une seule et unique valeur ou norme qui se contredit. Il n'y a donc pas de hiérarchie possible des valeurs et normes pour résoudre le dilemme ;
- Le problème de la sur-évaluation de l'obligation dans les conflits ;
- La distinction entre conflit intrinsèque et extrinsèque. Dans le cas d'agents autonomes, les conflits extrinsèques attireront principalement notre attention[6].

3.1.4 Relativisme, objectivisme et universalisme

Quelle est la portée éthique des agents autonomes ? Est-elle prétendument objective, universelle ou particulière ? Nous pouvons adopter deux positions. L'une à prétention universelle (*objectivisme*) et nous parlerons d'agents autonomes éthiques objectivistes. Et l'autre, plus particulière (*relativisme*), à partir de laquelle nous parlerons d'agents autonomes éthiques subjectivistes ou relativistes.

[5] Un principe approuvé par la plupart des systèmes de la logique déontique, dit que si un agent est tenu de faire chacune des deux actions, il est tenu de faire les deux. C'est le principe d'agglomération convoqué pour aboutir à une contradiction logique du dilemme.

[6] Il peut y avoir conflit extrinsèque avec un seul agent (l'environnement est extrinsèque).

Le **relativisme** est une doctrine qui soutient que l'absolu est hors d'atteinte, qu'une doctrine absolue est impossible. Ce relativisme peut relever de deux points de vue. D'un point de vue théorique qui est celui de la connaissance (relativisme épistémique ou cognitif) et, d'un point de vue pratique qui est celui de l'action et des jugements de valeurs (relativisme normatif spécialement en matière de morale ou de politique).

L'**absolutisme** par différence est une doctrine qui nie cette relativité et qui soutient la possibilité de dire absolument le vrai (absolutisme épistémique) ou absolument le bien (absolutisme pratique). Le relativisme théorique affirme donc la relativité de toute connaissance. Il soutient que nous n'avons accès à aucune vérité absolue ; soit parce que celle-ci n'existe pas ou est inconnaissable. Soit parce qu'on ne peut en acquérir qu'une connaissance relative. Le relativisme pratique affirme la relativité de toute valeur et donc de toute évaluation. Nous n'avons accès à aucune norme absolue, tout jugement de valeur est relatif. Il est relatif à un certain sujet, un certain corps, à certains gènes, une certaine histoire, une certaine société, une certaine culture, à un certain désir, voire, à tout cela à la fois. Ce point est essentiel dans le cadre des agents autonomes éthiques.

Les déclarations à portée universelles comme la *Déclaration Universelle des Droits de l'Homme et du Citoyen* ont bien le but d'établir des principes applicables à tous quels que soient le contexte historique et la culture. Il semblerait qu'il puisse exister un champ d'objectivité, morale d'ordre pratique, qui promeut des valeurs universellement bonnes, et auxquelles chacun peut accéder rationnellement, comme le penserait Kant. L'**objectivisme** moral affirme, d'une part, qu'il existe des valeurs indépendantes de nos désirs ou préférences : il faut rejeter le modèle du goût et lui préférer celui de la perception. Ce ne sont pas les valeurs qui procèdent des désirs, mais les désirs qui procèdent des valeurs. L'objectivisme soutient, d'autre part, que dans l'ensemble des énoncés évaluatifs moraux possibles, certains sont vrais et d'autres faux. En matière d'éthique nous pouvons nous tromper et nous pouvons aussi avoir raison. Cependant, l'objectivisme doit faire face à un quadruple défi : le double défi de Mackie (métaphysique et épistémologique) et le double défi de Hume (psychologique et logique) [Massin, 2008]. Les théories objectivistes permettent de rendre compte de la vérité des jugements moraux mais plus difficilement de la validité de leurs conséquences pratiques. À l'inverse, les théories non-objectivistes valident plus aisément les implications pratiques des jugements moraux (qui sont toujours fondés sur une pro-attitude conative d'approbation) mais pas de la vérité ou de la fausseté de ces jugements.

L'objectivité n'est pas l'**universalisme** et encore moins l'**universalisable**. Il faut ici faire la distinction entre les valeurs universelles et les valeurs universalisables. Les valeurs *absolument* universelles valent partout, toujours et pour tous. De ce point de vue, il n'y a pas de valeurs absolument universelles en fait. Par contre, il existe des valeurs *universalisables en droit*, c'est-à-dire qui valent pour tous (les Droits de l'Homme par exemple), et dont nous pouvons faire et devons faire qu'elles deviennent de plus en plus des valeurs universelles en fait. Un universel est soit absolu, c'est-à-dire vrai dans l'univers en entier. Quelque chose qui est vrai pour tous en tous lieux et toujours[7]. Ou bien relatif, c'est-à-dire vrai dans la totalité d'un ensemble donné. Quelque chose qui est vrai pour tous ceux de l'ensemble[8]. Une loi physique est doublement vraie ; elle l'est tout d'abord parce que c'est une proposition vraie en elle-même, elle est vraie *intrinsèquement*. Elle est ensuite vraie pour tous les objets sur lesquels elle portent, elle est vraie *extrinsèquement*. Toute vérité de fait (j'écris en ce moment par exemple) est tout aussi vraie, même si elle ne dépend pas de l'objet sur lequel elle porte, elle n'est pas vraie extrinsèquement. Mais parce que la proposition est vraie intrinsèquement, elle est nécessairement universelle.

Une proposition comme *tous les hommes sont égaux en droits et en dignité* est un universel relatif. Tout d'abord, il dépend d'un ensemble donné, l'humanité, ensuite, les désaccords éventuels sur la validité de la proposition ne se font pas en termes de connaissances mais en termes de valeurs. Le désaccord ne porte pas sur les valeurs elles-mêmes, mais plutôt sur une priorité des valeurs dans un contexte particulier. Dans le cadre des agents autonomes éthiques, le choix ne se situerait donc pas au niveau de l'ontologie (consensus axiologique) mais à celui des procédures de hiérarchisation des valeurs (logique intentionnelle).

3.2 Ethical models for autonomous agents

In order to be able to design moral or ethical autonomous agents, several questions must be considered. Firstly, what kind of formal model can be used to capture the main notions presented in Section 3.1? Secondly, what are the main ethical notions that are classically used in Artificial Intelligence? Finally, how are those notions implemented in artificial autonomous agents and what kind of implementation have been proposed in the litterature?

[7] On pense aux faits de la science.
[8] On pense aux valeurs de la morale.

3.2.1 Formal ethics

En termes de modélisation et de raisonnement, quelles logiques peuvent être employées dans le cadre de la définition d'agents autonomes éthiques ? Comme le note [McNamara, 2014], nous pouvons distinguer quatre notions de logique :

- l'étude des langages formels artificiels ;
- l'étude des inférences valides et des conséquences logiques ;
- l'étude des vérités logiques ;
- l'étude des caractéristiques générales, ou des formes, des jugements.

La diversité du champ non classique de la logique résulte en partie de la multiplicité des intentions et des motivations qui ont donné essor à ce genre de tentatives. Synthétiquement, les principales logiques non classiques sont :

Logique déontique : La logique déontique est une logique du *Tunsollen* (devoir faire). Elle considère des normes relatives au faire, à *l'agir*. En revanche, une logique du *Seinsollen* (devoir être) s'occupe de normes relatives à *l'être*, à des situations. La terminologie définie en Section 3.1 nous invite à envisager les logiques *Tunsollen* et *Seinsollen*. En effet, les valeurs substantielles (être sincère par exemple) relèvent du *Seinsollen*. Les normes morales (il est mal de tuer par exemple) relèvent du *Tunsollen*. À l'inverse de la logique aléthique (ou aristotélicienne, ou classique), il est délicat d'envisager l'imbrication des modalités déontiques[9] nécessaire pour modéler des comportements éthiques basée sur une hiérarchie d'ordres [Comte-Sponville, 2004] ou sur des distances [Meyer, 2013]. Des solutions comme l'édiction d'une norme comme une action d'un certain type ou l'utilisation d'opérateurs binaires de modalité qui mettent en relation une autorité et le contenu de la norme qu'elle édicte[10].

Logique épistémique : Dans la logique épistémique, il n'y a pas *inter-définissabilité* entre *x croit que P* et *x sait que P*. L'irréductibilité des opérateurs de *savoir* (K) et *croyance* (B) peut être acceptée dans une division de la logique épistémique en deux disciplines : la logique épistémique liée à la connaissance, dont l'objectif est la formalisation

[9] Par exemple, POA qui devrait signifier *il est permis que A soit obligatoire.*

[10] Par exemple, $P_x O_y A_z$ signifiant *l'autorité x, déléguant à y une partie de son pouvoir, l'habilite à obliger z à A.*

du concept de savoir, et la logique doxastique, attachée à la définition de la notion de croyance. Notons que la logique épistémique ne met pas en jeu qu'un seul agent. Les principes de la logique épistémique peuvent être appliqués aux situations impliquant plusieurs agents. Cette possibilité de la logique épistémique a été appliquée dans des domaines éloignés des préoccupations philosophiques initiales : la théorie des systèmes informatiques répartis, les communauté d'agents, etc. Ici, les agents épistémiques sont des processus de type *acteurs* qui travaillent de façon asynchrones en échangeant des informations (savoirs). Cet aspect est à prendre en compte dans la définition d'une communautés d'agents autonomes éthiques.

Logique floue : Cette logique permet la définition d'opérateurs, à la manière des adverbes très, peu, etc. qui affectent les adjectifs imprécis auxquels ils sont appliqués : les principaux opérateurs sont la compression et l'intensification (ou d'augmentation de contraste) ainsi que leurs inverses. La logique floue soulève cependant de sérieuses objections épistémologiques sur les assignations numériques aux qualités qui de fait ont elles-mêmes des valeurs de vérité floues. En quel sens et à quel degré un énoncé est-il lui-même vrai ? Dans le cadre de l'enchevêtrement des faits/valeurs comme nous le précisons au paragraphe 3.3.2, la logique floue offre plusieurs pistes intéressantes.

Logique impérative : Selon [Hare, 1952] cité par [Dubucs, 2015], l'éthique se réduit à l'étude logique du langage de la morale, et débute par la description des impératifs. Les impératifs sont irréductibles aux énoncés déontiques. Ces derniers sont logiquement représentés par l'application d'un opérateur déontique à la proposition qui exprime le contenu de la norme. De la même manière on séparera, dans un impératif, le contenu propositionnel (ou le composant phrastique) et la force (ou le composant neustique). Par exemple, le composant phrastique de *ferme la porte* est à peu près *la porte est fermée par toi dans un avenir proche*, et son composant neustique est la marque du mode impératif. Le raisonnement impératif prend en compte l'*échec* : l'ajout d'un nouvel ordre peut obliger à revenir sur une conclusion préalablement établie. Ce phénomène, inexistant en logique classique, confère à la logique des impératifs la *non monotonie*.

Logique inductive : Cette logique tente de décrire la relation entre deux propositions dont l'une confirme l'autre sans que la vérité de la première soit logiquement incompatible avec la fausseté de la seconde. La

logique inductive, contrairement à la logique déductive, repose sur des notions métriques[11]. Une hypothèse H sera dite impliquée à un certain degré par une évidence E. Ce degré, noté $c(H, E)$, est appelé le degré de confirmation de H par E.

Logique non monotone : Les raisonnements non monotones permettent l'élaboration de croyances (hypothèses plus ou moins assurées) qui au fil du raisonnement peuvent invalider des conclusions : l'individu dont on présume qu'il est un ϕ peut se révéler un ψ atypique. La logique est dite non monotone si l'extension des prémisses peut aboutir au retrait d'une conclusion préalablement établie. Par analogie, la déclaration d'une variable informatique qui ne pas fait l'objet d'une affectation explicite reçoit une valeur par défaut qui peut ensuite être modifiée. On parle de raisonnements d'inférence par défaut : la valeur de ϕ pour l'objet a qui est un ψ est, par défaut, le vrai [Dubucs, 2015].

Les logiques non monotones associées à des logiques déontiques et épistémiques dont les opérateurs des agents et des contenus sémantiques semblent appropriées pour exprimer certaines propriétés importantes du raisonnement éthique. Les logiques non monotones ont surtout montré leur efficacité dans les techniques d'Intelligence Artificielle, les bases de données ou les architectures d'agents asynchrones[12].

3.2.2 Ethical models for autonomous agents

The two main ethical models that have served as bases for logical formalization are the deontological and the empirical models. As far as we know, there has been no attempt yet to mathematically formalize an *axiological ethics*, in particular *virtue ethics*[13]. This absence may be due more to the difficulty to mathematically formalize values or virtues than to a lack of interest. Note that, despite this absence of mathematical formalization, the virtue ethics

[11] Comme les distances de [Meyer, 2013] ou les ordres de [Comte-Sponville, 2004]

[12] La question du cadre de référence (frame) dans la représentation des univers dynamiques : comment décrire la persistance de certaines situations au travers des modifications induites par les opérations d'un agent (par exemple, le fait que le déplacement d'un objet n'en change pas la couleur) ? Là encore, l'intervention d'inférences non monotones permet d'économiser l'écriture d'un très grand nombre de conditions d'invariance associées à chaque action : si p est le cas maintenant, alors tout à l'heure aussi, sauf preuve du contraire [Dubucs, 2015]

[13] Virtue ethics emphasizes the values (or virtues) that an agent embodies for determining or evaluating its ethical behavior.

has been very often mentioned (for instance by [Coleman, 2001]) as an excellent reference for computer ethics, because the general idea of virtue and the particular virtues – like courage, justice, *phronesis*, magnanimity, etc. – seem to be universal and independent of any particular culture.

On the one hand, there have already been many attempts to formalize the ethical behaviors of agents using sets of laws, which corresponds, implicitly, to a *deontological model of ethics*, even if the sources of the laws have not always been made explicit. At first sight, the classical deontic logic (or more elaborate deontic logics) [Chellas, 1980, von Wright, 1951] seemed perfectly appropriate for this purpose, since they were designed to describe what ought to be, in terms of duties, obligations or rights. It naturally follows from this that deontic logics have been used to formalize the rules on which the behavior of deontological autonomous agents is based [Gensler, 1996,Powers, 2005, Bringsjord and Taylors, 2012]. Those approaches present three limits:

- They focus on general laws where permission and prohibitions are well defined but not on consequentialist or particularist ethical systems, even there are some attempts to model Kantian ethics, i.e. the categorical imperative, to justify sets of laws [Powers, 2006, Ganascia, 2007].

- Since these formalizations are based on sets of laws an agent is supposed to obey, it is not always easy to distinguish between such deontological autonomous agents and normative agents [Dignum, 1999,Grossi et al., 2008, Rotolo and van der Torre, 2011, Balke et al., 2013]. Maybe the difference comes from the consideration of conflicts that are present in ethical reasoning, while they are supposed to be precluded in normative systems. This point will need some clarifications in the future.

- As mentioned by [van Fraassen, 1973, Horty, 1994], such formalizations fail to deal with ethical dilemmas. Some well-known paradoxes [Hansen, 2006b], e.g. the *Chisholm's Paradox* [Chisholm, 1963] or the *paradox of the gentle murderer* [Forrester, 1984] illustrate those difficulties. There were attempts to overcome contradictions resulting from the existence of ethical dilemmas [Goble, 2005]. Among them, some advocate the introduction of priorities among norms [Hansen, 2006a], the use of non-monotonic formalisms [Horty, 1994, Ganascia, 2012], e.g. default logics or non-monotonic logics, or both [Brewka, 1994].

However, these works do not really focus on the design of deontological autonomous agents but on normative agents, i.e. on agents that respect

norms: they implicitly suppose that morality has to be assimilated to the respect of sets of norms, i.e. to a deontic approach. Some authors, for instance Noël Shakey cited by [Dabringer, 2011], say that this view is too restricted because in concrete situations, especially in affairs of war, the arbitration between ethical principles has to take the consequences of actions into account. **The problem is to obey general ethical standards when the situation permits, and to violate them when some of the consequences of their application are worse than their non-application**.

On the other hand, there have been attempts to base ethics on **empirical principles**, i.e. on observations according, for instance, to the observed utility, to common uses or to traditions. More recently, philosophers have used Artificial Intelligence techniques, and more specifically statistical learning theory [Harman and Kulkarni, 2011,Harman and Kulkarni, 2012] or game theory [Braithwaite, 1955], to model these processes using computers and/or well-founded mathematical theories. There is no doubt that such attempts are very fruitful and interesting. **However, these approaches do not allow to understand the underlying logic on which classical ethical systems rely and thus how to implement these systems to automate a decision**.

3.2.3 Implementations of ethical autonomous agents

Despite the well-known limits of ethical models for autonomous agents, different kinds of implementations have been proposed.

- Some of these implementations are designed to help human users to analyse ethical issues [Frize et al., 2005, Okada et al., 2007, Chatterjee et al., 2009]. Using *modeling languages* such as UML, they provide a better understanding of the internal logic of ethical systems. However, as those implementations just model the ethical process but do not solve the ethical problme, they do not really contribute to the achievement of ethical artificial agents.

- The second type of implementation helps to interactively elicit the criteria that are used to make an ethical decision. The approaches proposed by [Chae et al., 2005, Anderson et al., 2006, Mathieson, 2007, Robbins and Wallace, 2007] clearly belong to this type. If they can be interesting for some functionalities of moral or ethical artificial agents, such approaches do not constitute by themselves an automation of the ethical decision process.

- The third type is certainly the most interesting from our point of view, because it aims at actually building artificial agents whose behaviors are ethically acceptable. They are based on Artificial Intelligence techniques [McLaren, 2003, Guerini and Stock, 2005, Powers, 2005, Arkin, 2009, Bench-Capon and Atkinson, 2009, Honarvar and Ghasem-Aghaee, 2009, Chopra and White, 2011, Bringsjord and Taylors, 2012, Tufis and Ganascia, 2012, Saptawijaya and Pereira, 2014]. Some implementations have been realized.

Each implementation based on Artificial Intelligence techniques address some specific question of ethics. Some works attempt to achieve agents based on *modal logic and deontic logics* [Powers, 2005, Bringsjord and Taylors, 2012] in order to address the question of the reproduction of an ethical behavior with respect to given general ethical principles. Some other works aim at addressing the problem of *non-monotony of ethical reasoning*. For instance, some approaches based on formal argumentation [Atkinson and Bench-Capon, 2008, Bench-Capon and Atkinson, 2009] have been proposed to this end. Few other implementations directly use *computer models* [McLaren, 2003, Arkin, 2009, Honarvar and Ghasem-Aghaee, 2009, Saptawijaya and Pereira, 2014]. For instance, [Saptawijaya and Pereira, 2014] attempt to build an ethical agent with logic programming techniques, [Arkin, 2009] uses case-based reasoning to constraint the agent behavioral architecture and [Honarvar and Ghasem-Aghaee, 2009] propose to learn ethics through a neural network. Lastly, there are many implementations of *normative agents* based on sets of laws [Guerini and Stock, 2005, Chopra and White, 2011, Tufis and Ganascia, 2012]. For instance, [Guerini and Stock, 2005] propose deontic rules that constraint the planning process of agents: permissions and obligations are used to define unethical, ethical, altruistic, supererogatory and antisocial goals. For another instance, [Tufis and Ganascia, 2012] propose a BDI (Belief Desire Intention) architecture for normative agents. There are also a few multi-agents normative systems such ash convivial normative systems proposed by [Caire, 2009].

3.3 Ethical, moral or competent agent?

Given the state-of-the-art presented in this chapter what kind of artificial autonomous agents can and must be implemented in order to deal with ethical conflicts presented in Chapter 2?

3.3.1 Towards an agent axiological realism

Lorsque nous parlons d'*agents artificiels autonomes éthiques*, nous devons plutôt parler d'agents artificiels autonomes axiologiques. Lorsque nous parlons d'*agents artificiels moraux*, nous devons parler d'agents artificiels autonomes déontologiques.

Definition 3.4 (Agent artificiel autonome axiologique) *Un agent artificiel autonome axiologique est un agent dont le comportement quel qu'il soit intègre de manière explicite des valeurs. La portée éthique de cet agent ne se situe pas au au niveau de l'ontologie (consensus axiologique) mais à celui des procédures de hiérarchisation des valeurs (logique intentionnelle).*

Definition 3.5 (Agent artificiel autonome déontologique) *Un agent artificiel autonome déontologique est un agent dont le comportement quel qu'il soit intègre de manière explicite des normes. La portée éthique de cet agent ne se situe pas au au niveau de l'ontologie (choix des normes) mais à celui des procédures de hiérarchisation des normes (logique déontique).*

Au vu de ces deux définitions et de la relation entre norme et valeurs, quelle que soit l'approche adoptée, la *valeur* reste au cœur de l'édifice de la théorie morale. Il y a donc nécessité d'un réalisme axiologique pour les agents autonomes, seule condition d'accès possible à la modélisation des valeurs. Ainsi, la conception d'une ontologie axiologique requiert une attention particulière que l'approche adoptée soit normative, axiologique ou conséquentialiste (ou toutes suivant les situations). S'agissant des agents autonomes et de la modélisation de leur comportement éthique, l'axiologisme engage à une définition ontologique des valeurs, préalable à leur compréhension dans des structures logiques.

3.3.2 Towards an ethical competent agent

Lorsque l'on parle d'agents autonomes éthiques, qu'entend-on précisément ? Un agent autonome peut-il être *éthiquement neutre*, c'est-à-dire désintéressé dans ses décisions éthiques ? Or, les problèmes éthiques ne sont pas situés au niveau même des valeurs morales qui en soi peuvent être universalisables. Les positions éthiques diffèrent par l'enchevêtrement des faits (économiques, financiers) et des valeurs morales. Par exemple, bien que le partenariat entre Union Européenne (UE) et États-Unis (EU) dans le cadre du marché international soit fondée sur *[...] des valeurs communes telles que les droits de l'Homme, les libertés fondamentales, la démocratie et l'état de droit (Art.*

6), le fait économique impose des contraintes aux valeurs morales. Sur le droit du travail, l'UE a ratifiée lse conventions de l'Organisation Internationale du Travail. Sur la protection de l'environnement, l'UE a ratifié le protocole de Kyoto et la convention de Rio. Les EU non. Sur la culture, l'UE a ratifié une convention sur le respect de la diversité culturelle dont l'UNESCO est la gardienne. Les EU non. Sur le plan juridique, l'UE a ratifié la CIDE (Convention internationale sur les droits de l'enfant) et le statut de la cour pénale internationale. Les EU non. Sur la conception du risque, l'UE considère que quelque chose n'est pas nocif si on a pu démontrer qu'il ne l'est pas (principe de précaution). Les EU considèrent que quelque chose n'est pas nofif tant qu'on n'a pas démontré qu'il l'était.

Bien que cet enchevêtrement ne devrait concerner que les groupements humains, car comme le note Kant tout n'est que question de volonté[14], cette difficulté peut avoir un impact lors de la conception d'agents artificiels autonomes. Ainsi, au-delà de l'intérêt intellectuel et théorique des dilemmes et des modèles présentés précédemment, la réalité nous offre bien des occasions de conflits et ce ne sont pas des conflits de valeurs comme dans le cas des dilemmes mais des conflits entre *faits et valeurs*. Imaginons le choix à effectuer parmi les deux options suivantes :

1. Il est obligatoire de préserver au maximum la santé tant physique qu'intellectuelle d'autrui.

2. Il est possible de participer à la dégradation tant physique qu'intellectuelle d'autrui.

Dans l'hypothèse d'une neutralité éthique, la réponse 1 semble être la plus éthique. Or, la réalité est tout autre car tout agent autonome est avant tout un *agent compétent* pour un type d'activité bien défini. Cette compétence dépend du domaine métier de l'agent : agent économique, financier, politique, géopolitique, militaire, de renseignement, etc. Dès lors, nous sommes confrontés à un choix entre les prérogatives morales et les intérêts[15] du domaine métier de l'agent (économie, finance, politique, etc.) qui engage la responsabilité du décideur. La responsabilité relève d'une logique de prise de décision *libre* : ce n'est pas un problème à résoudre, c'est un choix à opérer, ce qui ne va pas sans hiérarchies ni renoncements[16].

[14] *De tout ce qu'il est possible de concevoir dans le monde, et même en général hors du monde, il n'est rien qui puisse sans restriction être tenu pour bon, si ce n'est seulement une bonne volonté* [Kant, 1792].

[15] La notion d'intêret dépasse ici les simples buts spécifiés à l'agent.

[16] Plus souvent un renoncement moral que financier.

Or, quelle que soit la position philosophique que l'on adopte (liberté ou déterminisme), un agent artificiel autonome est étranger à la responsabilité ou prise de décision libre (voir Section 1.3.3). De fait, la décision éthique d'un agent ne peut relever que d'un problème de compétence (conception, modélisation, programmation, tests). Dès lors, il semble plus délicat de parler d'agent éthique. Nous devons plutôt parler d'agents éthiques neutres ou non.

Definition 3.6 (Agent artificiel autonome éthique neutre) *Un* agent artificiel autononome éthique neutre *est un agent éthique dont le comportement est entièrement déterminé par des devoirs moraux et non pas l'intérêt.*

Il est essentiel de noter que la définition d'agents autonomes éthiques neutres est consensuelle, voire universalisable, car tout individu est en mesure de savoir à tout moment et en toute situation quel est son devoir moral, et à plus forte raison des agents autonomes éthiques dont les devoirs moraux sont explicites. Cependant, la difficulté d'une opérationnalisation d'un agent autonome éthique neutre n'est pas située au niveau de l'obligation qui pourrait ne pas être suivi d'actes car les agents obéissent à des impératifs auxquels ils ne peuvent se soustraire. La difficulté réside dans l'intérêt, qui est toujours à l'origine des conflits dans la décision éthique. En effet, les intérêts prennent le dessus sur les idéaux humanistes. Ce n'est donc pas un manque de repères, un manque de conscience morale, qui est la source des actes *immoraux*, c'est un renversement des impératifs.

En paraphrasant Kant, plutôt que de soumettre la recherche de notre bonheur (gagner de l'argent dans notre exemple) à la loi morale (ne pas faire de mal à autrui), nous soumettons la loi morale à la recherche de notre bonheur. Par exemple, l'histoire *douloureuse* donne raison à Kant contre Constant sur un prétendu droit de mentir [Kant, 1988]. Les mensonges des totalitarismes, dictatures et autres empires l'ont largement emporté en nombre de victimes sur les *aveux* d'individus soumis à leurs bourreaux. Pour citer [Orwell, 1972], *en ces temps d'imposture universelle, dire la vérité est un acte révolutionnaire.* En effet, les campagnes politiques, publicitaires, les propagandes, les guerres illégales reposent toutes sur une version du mensonge. Comment trouver un domaine vierge de tous mensonges pour le développement d'agents autonomes éthiques neutres? La solution est-elle dans la réalisation d'agents autonomes *honnêtes* ?

Nous devons alors, dans le meilleur des cas, parler d'agents autonomes éthiques compétents ou d'agents parfaitement légalistes et par extension, si

l'on veut, parler d'agents parfaitement moraux[17]. Dans le cadre des agents autonomes éthiques, le choix ne se situerait donc pas au niveau de l'ontologie (consensus axiologique), mais à celui des procédures de hiérarchisation des valeurs (logique intentionnelle).

Definition 3.7 (Agent artificiel autonome éthique compétent) *Un* agent artificiel autonome éthique compétent *est un agent dont le comportement-métier intègre de manière explicite des valeurs ainsi qu'un arbitrage entre ces valeurs et les intêrets de l'agent dans son domaine-métier. Cet agent est alors à même de justifier ses prises de décisions de manière à pouvoir être jugé éthique ou non par un autre agent.*

3.4 Synthesis

This chapter has presented a synthetic view of ethics in terms of philosophical concepts, logical models and implementations. It has been highlighted that:

- Values are in the heart of ethical theories and an axiological ontology must be carefully defined in order to model values.
- Neutrally ethical autonomous agents cannot be designed as universalisable moral duties conflict with the agents' business domains.
- The classical design of deontological autonomous agents lacks explicit arbitration between ethical principles and agents' interests.
- Implementations based on Artificial Intelligence techniques only address some specific questions of ethics.

Therefore, the ETHICAA project only considers ethical competent artificial autonomous agents (see Definition 3.7). To this end, the model of an agent must integrate both an explicit representation of values based on an axiological ontology and an explicit representation of the arbitration between values and the agent's business rules such that this arbitration can be understood by other agents.

[17] Mais peut-on encore parler de normes dans de telles situations entièrement déterministes puisque la désobéissance y est inintelligible ?

Chapter 4

General conclusion

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. Human users, consciously or not, 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 [Aarts and de Ruyter 2009]. Increasing the scope of the activities of autonomous agents has become a major issue in our digital society, raising different ethical problems. It is thus important to define regulation and control mechanisms to ensure sound and consistent behaviours [Boella and Van der Torre 2006] and to ensure that the agents would not harm humans or threaten their decision autonomy [Pontier and Hoorn, 2012]. Setting an ethical regulation or control in such systems has been discussed by authors such as [Allen et al., 2006, Wallach and Allen, 2009]. As stated by [Picard, 1997], the greater the freedom of a machine, the more it will need moral standards.

4.1 Ethical conflicts within multi-agent systems

Within ICT and socio-technical systems, the ETHICAA project considers systems of multiple autonomous agents:

Definition 4.1 (System of multiple autonomous agents) *A system of multiple autonomous agents is a system containing at least one* autonomous artificial agent *and several other autonomous agents that can be* human agents *(human users or human operators) and/or other* autonomous artificial agents. *Autonomous artificial agents are finite entities with limited per-*

ception and action capabilities able to satisfy their users or operators' goals by selecting and executing actions automatically according to their context.

We consider four kinds of applicative contexts involving such systems of multiple autonomous agents: virtual communities, unmanned vehicles, decision making support systems and ubiquitous computing. These use cases allow us to consider ethical conflicts with:

- both robotic and software agents;
- both individual and collective decision making processes;
- both informative and physical actions;
- both privacy and dignity issues.

4.2 Ethical competent artificial autonomous agents

Implementing ethical artificial autonomous agents meets two limitations:

1. The notion of value is always in the heart of ethical theories. Therefore, an axiological ontology must be carefully defined in order to model values. Moreover, implementations based on Artificial Intelligence techniques only address some specific questions of ethics and cannot manage ethical conflicts and values in a general way.
2. Every autonomous artificial agent is designed to satisfy its users or operators' goals. Firstly the classical design of autonomous agents lacks explicit arbitration between ethical principles and the agents' interests. Moreover neutrally ethical autonomous agents cannot be designed as universalisable moral duties conflict with the agents' business domains.

Implementing autonomous ethical agents means implementing an arbitration between values and the goals for what the agent is designed. Consequently we can only consider ethical competent autonomous artificial agents:

Definition 4.2 (Ethical competent autonomous artificial agent) *An* ethical competent autonomous artificial agent *is an agent whose autonomic behaviors explicitely integrate both values and an arbitration between values and its business rules.*

4.3 Ethical validation and ethical explanation

In an autonomous agents system, multiple agents that may be heterogeneous in terms of goals and ethics interact and participate to organizations. Thus it is the first importance to allow the agents to justify their decisions in order to be judged as ethical (or not) by other agents. Consequently an ethical competent autonomous artificial agent should also be able:

- at the **micro-level** to represent its ethics and justify its decisions, to represent the ethics of another agent and to verify that this agent's behavior follows its ethics, to judge the ethics of the other agent through a comparison mechanism, and to take into account this judgment in its own decisions.
- at the **macro-level** to build a collective ethics, to identify and be able to judge a collective ethics, to be able to judge other agents through the collective ethics and to make an arbitration between its own ethics and the collective ethics.

Chapter 5

Glossary

- **Agent**, an artificial (physical or virtual) or biological entity with limited perception and action capabilities, which exhibits autonomy in pursuit of its goals according to a business domain along with a set of various communication and reasoning skills.
- **Authority**, a set of power relationships between agents that determines how goals, tasks and resources are allocated. An agent holds the authority on a feature (i.e. goal, task or resource) with respect to another agent if it can control this feature at the other agent's expense.
- **Autonomy**, the capacity of an agent to decide and act independently of another agent while behaving in a non-trivial way in complex and changing environments possibly including other agents.
- **Axiology**, the study of moral values. An *axiologic autonomous agent* (or ethical agent) bases its behavior on values.
- **Common welfare**, from which a society as a whole benefits, in contrast to the private goods of individuals and organizations within the society.
- **Consequentialism**, ethical principle that prescribes that the rightful act is the one that produces the best consequences.
- **Deontology**, the study of norms and duties: *obligation*, *prohibition* and *permission*. A *deontological autonomous agent* (or moral agent) bases its behavior on norms.

- **Dignity**, a distinctive kind of intrinsic and incomparable moral value according to which some valuing form of moral recognition or reverential respect should be attached to an object.
- **Ethical principle**, a part of an ethical theory that describes how norms and values should be used.
- **Ethical dilemma**, a situation where a choice must be made between several actions that all lead to moral norms or moral values violation.
- **Normative system**, a multi-agent system using mechanisms to represent, communicate, distribute, detect, create, modify, and enforce norms, and mechanisms to deliberate about norms and detect norm violation and fulfilment. Norms are rules defined by the society that influence the behaviours of the agents.
- **Privacy**, an agent's right to determine when, how, and to what extent information about itself is communicated to and used by others.
- **Responsability**, a condition required in order to blame or praise an agent for a given state of the world.
- **Trust**, an agent's belief in another agent's capabilities, honesty and reliability based on its own direct experiences. Trust is related to reputation where reputation is an agent's belief in another agent's capabilities, honesty and reliability based on recommendations received from other agents.
- **Utilitarism**, an ethical theory which has three dimensions: a criterion for good (*welfarism*), a moral imperative to maximize the good (*prescriptivism*) and an evaluation rule (*consequentialism*).
- **Virtue**, a kind of instrinsic value related to an agent or an action.

Bibliography

[Aarts and de Ruyter, 2009] Aarts, E. and de Ruyter, B. (2009). New research perspectives on ambient intelligence. *Journal of Ambiant Intelligence and Smart Environment*, 1(1):5–14.

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

[Altman and Tennenholtz, 2010] Altman, A. and Tennenholtz, M. (2010). An axiomatic approach to personalized ranking systems. *Journal of the ACM*, 57(4).

[Anderson et al., 2006] Anderson, M., Anderson, S., and Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, 21(4):56–63.

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

[Aschwanden et al., 2006] Aschwanden, P., Baskaran, V., Bernardini, S., Fry, C., Moreno, M., Muscettola, N., Plaunt, C., Rijsman, D., and Tompkins, P. (2006). Model-unified planning and execution for distributed autonomous system control. In *AAAI 2006 Fall Symposium*, pages 1–10.

[Atkinson and Bench-Capon, 2008] Atkinson, K. and Bench-Capon, T. (2008). Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6(2):135–151.

[Azzedin and Maheswaran, 2003] Azzedin, F. and Maheswaran, M. (2003). Trust modeling for peer-to-peer based computing systems. In *17th International Parallel and Distributed Processing Symposium*, pages 99–109.

[B. Williams, 1990] B. Williams, t. M.-A. L. (1990). *Éthique et les limites de la philosophie*. Gallimard.

[Balke et al., 2013] Balke, T., da Costa Pereira, C., Dignum, F., Lorini, E., Rotolo, A., Vasconcelos, W., and Villata, S. (2013). Norms in mas: Definitions and related concepts. *Dagstuhl Follow-Up*, 4:1–31.

[Bekey, 2005] Bekey, G. (2005). *Autonomous robots: from biological inspiration to implementation and control*. MIT Press.

[Bench-Capon and Atkinson, 2009] Bench-Capon, T. and Atkinson, K. (2009). Argumentation in Artificial Intelligence. In Simari, G. and Rahwan, I., editors, *Abstract argumentation and values*, pages 45–64. Springer.

[Bernoux, 1985] Bernoux, P. (1985). *La sociologie des organisations*. Seuil.

[Boella et al., 2008] Boella, G., Van der Torre, L., and Verhagen, H. (2008). Introduction to the special issue on normative multiagent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 17(1):1–10.

[Boissier, 2001] Boissier, O. (2001). Modèles et architectures d'agents. In Briot, J.-P. and Demazeau, Y., editors, *Principes et Architectures des Systèmes Multi-Agents*, pages 71–107. Hermès.

[Boissier, 2003] Boissier, O. (2003). Contrôle et coordination orientés multi-agents. Habilitation à diriger des recherches, ENS Mines Saint-Etienne et Université Jean Monnet.

[Boissier et al., 2010] Boissier, O., Balbo, F., and Badeig, F. (2010). Controlling multi-party interaction within normative multi-agent organizations. In *3rd Federated Workshops on Multi-Agent Logics, Languages, and Organisations*. CEUR Proceedings Vol. 627.

[Boissier et al., 2006] Boissier, O., Hübner, J., and Sichman, J. (2006). Organization oriented programming: from closed to open organizations. In *7th International Conference on Engineering Societies in the Agents World*, pages 86–105.

[Bradshaw et al., 2003] Bradshaw, J., Sierhuis, M., Acquisti, A., Feltovich, R., Hoffman, R., Jeffers, R., Prescott, D., Suri, N., Uszok, A., and Van Hoof, R. (2003). Adjustable autonomy and human-agent teamwork in practice: an interim report on space application. In Hexmoor, H., Castelfranchi, C., and Falcone, R., editors, *Intelligent Agents*, pages 243–280. Kluwer Academic Publishers.

[Braithwaite, 1955] Braithwaite, R. (1955). *Theory of games as a tool for the moral philosopher*. Cambridge University Press.

[Brewka, 1994] Brewka, G. (1994). Reasoning about priorities in default logic. In *12th National Conference on Artificial Intelligence*, pages 940–945.

[Bringsjord and Taylors, 2012] Bringsjord, S. and Taylors, J. (2012). Introducing divine-command robot ethics. In Lin, P., Bekey, G., and Abney, K., editors, *Robot ethics: the ethical and social implication of robotics*, pages 85–108. MIT Press.

[Brooks and Durfee, 2003] Brooks, C. and Durfee, E. (2003). Congregation formation in multiagent systems. *Journal of Autonomous Agents and Multiagent Systems*, 7:145–170.

[Brooks, 1986] Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

[Brooks, 1991] Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159.

[Buzing et al., 2005] Buzing, P., Eiben, A., and Schut, M. (2005). Emerging communication and cooperation in evolving agent societies. *Journal of Artificial Societies and Social Simulation*, 8(1):27–52.

[Caire, 2009] Caire, P. (2009). Designing convivial digital cities: A social intelligence design approach. *AI and Society Journal*, 24(1):97–114.

[Carabelea et al., 2003] Carabelea, C., Boissier, O., and Florea, A. (2003). Autonomy in multi-agent systems: A classification attempt. *Lecture Notes in Computer Science*, 2969:103–113.

[Cartlidge et al., 2012] Cartlidge, J., Szostek, C., De Luca, M., and Cliff, D. (2012). Too fast too furious: faster financial-market trading agents can give less efficient markets. In *4th International Conference on Agents and Artificial Intelligence*, pages 126–135.

[Castelfranchi, 1998] Castelfranchi, C. (1998). Modeling social action for AI agents. *Artificial Intelligence*, 103:157–182.

[Castelfranchi, 2000] Castelfranchi, C. (2000). Conflict ontology. In *Computational conflicts*, pages 21–40. Springer.

[Castelfranchi and Falcone, 2000] Castelfranchi, C. and Falcone, R. (2000). Conflicts within and for collaboration. In Tessier, C., Chaudron, L., and Müller, H.-J., editors, *Conflicting agents*, pages 33–61. Springer.

[Castelfranchi and Falcone, 2003] Castelfranchi, C. and Falcone, R. (2003). From automaticity to autonomy: the frontier of artificial agents. In Hexmoor, H., Castelfranchi, C., and Falcone, R., editors, *Agent autonomy*, pages 103–136. Kluwer Academic Publishers.

[Chae et al., 2005] 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.

[Chatterjee et al., 2009] Chatterjee, S., Sarker, S., and Fuller, M. (2009). A deontological approach to designing ethical collaboration. *Journal of the Association for Information Systems*, 10:138–169.

[Chellas, 1980] Chellas, B. (1980). *Modal logic, an introduction*, chapter 6. Cambridge University Press.

[Cheng and Friedman, 2005] Cheng, A. and Friedman, E. (2005). Sybilproof reputation mechanisms. In *3rd Workshop on Economics of Peer-to-Peer Systems*, pages 128–132.

[Chenique, 2006] Chenique, F. (2006). *Éléments de logique classique (French Edition)*. Harmattan.

[Chisholm, 1963] Chisholm, R. (1963). Contrary-to-duty imperatives and deontic logic. *Analysis*, 24:33–36.

[Chopra and White, 2011] Chopra, S. and White, L.-F. (2011). A legal theory for autonomous artificial agents. Technical report, University of Michigan.

[Coleman, 2001] Coleman, K. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, 3(4):247–265.

[Comte-Sponville, 2004] Comte-Sponville, A. (2004). *Le capitalisme est-il moral ?* Albin Michel.

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

[Coppin and Legras, 2012] Coppin, G. and Legras, F. (2012). Controlling swarms of unmanned vehicles through user-centered commands. In *AAAI Fall Symposium: Human Control of Bioinspired Swarms*, pages 21–25.

[Corkill and Lander, 1998] Corkill, D. and Lander, S. (1998). Diversity in agent organization. *Object Magazine*, 8(4):41–47.

[Dabringer, 2011] Dabringer, G. (2011). *Ethical and Legal Aspects of Unmanned Systems*, pages 43–51. Institut fÃ¼r Religion und Frieden.

[Defense Science Board, 2012] Defense Science Board (2012). The role of autonomy in DoD systems. Technical report, Department of Defense.

[Dehais and Pasquier, 2000] Dehais, F. and Pasquier, P. (2000). Approche générique du conflit. *Ergonomie et Interaction Homme-Machine*, pages 56–63.

[Demazeau and Müller, 1991] Demazeau, Y. and Müller, J.-P. (1991). From reactive to intentional agents. *Decentralized Artificial Intelligence*, 2:3–10.

[Dignum, 1999] Dignum, F. (1999). Autonomous agents with norms. *Artificial Intelligence an Law*, 7(1):69–79.

[Docherty, 2012] Docherty, B. (2012). Losing humanity - The case against killer robots. Technical report, Human Rights Watch.

[Dorais et al., 1999] Dorais, G., Bonasso, P., Kortenkamp, D., Pell, B., and Schreckenghost, D. (1999). Adjustable autonomy for human-centered autonomous systems. In *Workshop on Adjustable Autonomy Systems*.

[Drogoul, 1995] Drogoul, A. (1995). When ants play chess (or can strategies emerge form tactical behaviors?). *Artificial Intelligence*, 957:13–27.

[Drogoul et al., 1995] Drogoul, A., Corbara, B., and Lalande, S. (1995). MANTA: New experimental results on the emergence of (artificial) ant societies. In Gilbert, N. and Conte, R., editors, *Artificial Societies: the Computer Simulation of Social Life*, pages 119–221. UCL Press.

[Dubos, 2012] Dubos, T. (2012). Integrating civil unmanned aircraft operating autonomously in non-segregated airspace: towards a dronoethics? In *1st Workshop on Rights and Duties of Autonomous Agents*, pages 13–18. CEUR Proceedings Vol. 885.

[Dubucs, 2015] Dubucs, J.-P. (2015). Logiques non classiques. In Universalis, E., editor, *Encyclopedia Universalis [en ligne]*.

[Durfee, 2001] Durfee, E. (2001). Scaling up agent coordination strategies. *IEEE Computer*, 34(7):39–46.

[Feldman and Chuang, 2005] Feldman, M. and Chuang, J. (2005). Overcoming free-riding behavior in peer-to-peer systems. *ACM SIGecom Exchanges*, 5(4):41–50.

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

[Ferguson, 1992] Ferguson, I. (1992). *Touring Machines: An architecture for dynamic, rational, mobile agents*. PhD thesis, University of Cambridge.

[Fischer et al., 2003] Fischer, K., Schillo, M., and Siekmann, J. (2003). Holonic multiagent systems: A foundation for the organisation of multiagent systems. In *1st International Conference on Applications of Holonic and Multi-Agent Systems*, pages 71–80.

[Fong et al., 2003] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166.

[Forrester, 1984] Forrester, J. (1984). Gentle murder, or the adverbial Samaritan. *Journal of Philosophy*, 81:193–197.

[Franklin and Graesser, 1996] 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.

[Freud, 1916] Freud, S. (1916). *Wit and its relation to the unconscious*. Moffat, Yard and Co.

[Frize et al., 2005] Frize, M., Yang, L., Walker, R., and O'Connor, A. (2005). Conceptual framework of knowledge management for ethical decision-making support in neonatal intensive care. *IEEE Transactions on Information Technology in Biomedicine*, 9(2):205–215.

[Galliers, 1990] Galliers, J. (1990). The positive role of conflicts in cooperative multi-agent systems. In Demazeau, Y., editor, *Decentralized AI*, pages 33–49. Elsevier.

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

[Ganascia, 2012] Ganascia, J. (2012). An agent-based formalization for resolving ethical conflicts. In *1st Workshop on Belief change, Nonmonotonic reasoning, and Conflict resolution.*

[Gans et al., 2001] Gans, G., Jarke, M., Kethers, S., Lakemeyer, G., Ellrich, L., Funken, C., and Meister, M. (2001). Towards (dis)trust-based simulations of agent networks. In *4th Workshop on Deception, Fraud, and Trust in Agent Societies*, pages 49–60.

[Gasser, 2001] Gasser, L. (2001). Organizations in multi-agent systems. In *10th European Worshop on Modeling Autonomous Agents in a Multi-Agent World.*

[Gensler, 1996] Gensler, H. (1996). *Formal ethics.* Routledge.

[Gleizes et al., 2008] Gleizes, M.-P., Camps, V., Georgé, J.-P., and Capera, D. (2008). Engineering systems which generate emergent functionalities. *Lecture Notes in Artificial Intelligence: Special Issue on Engineering Environment-Mediated Multiagent Systems*, 5049.

[Goble, 2005] Goble, L. (2005). A logic for deontic dilemma. *Journal of Applied Logic*, 3(3-4):461–483.

[Goldman and Rosenschein, 2002] Goldman, C. and Rosenschein, J. (2002). Evolutionary patterns of agent organizations. *IEEE Transactions on Systems, Man and Cybernetics Part A*, 30(1):135–148.

[Goodrich et al., 2001] Goodrich, M., Olsen, D., Crandall, J., and Palmer, T. (2001). Experiments in adjustable autonomy. In *Workshop on Autonomy, Delegation and Control: Interacting with Autonomous Agents.*

[Goodrich and Schultz, 2007] Goodrich, M. and Schultz, A. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275.

[Grassé, 1959] Grassé, P.-P. (1959). La reconstruction du nid et les coordinations inter-individuelles chez *bellicositermes natalensis* et cubitermes sp. la théorie de la stigmergie : essai d'interprétation du comportement des termites constructeurs. *Insectes Sociaux*, 6:41–81.

[Greenspan, 1983] Greenspan, P. (1983). Moral dilemmas and guilt. *Philosophical Studies*, 43(1):117–125.

[Grossi et al., 2008] Grossi, D., Gabbay, D., and van der Torre, L. (2008). A normative view on the blocks world. In *3rd International Workshop on Normative Multiagent Systems*, pages 128–142.

[Guerini and Stock, 2005] Guerini, M. and Stock, O. (2005). Toward ethical persuasive agents. In *IJCAI Workshop on Computational Models of Natural Arguments*.

[Guttman and Maes, 1998] Guttman, R. and Maes, P. (1998). Agent-mediated integrative negotiation for retail electronic commerce. In *1st International Workshop on Agent Mediated Electronic Trading*, pages 70–90.

[Hannebauer, 2000] Hannebauer, M. (2000). Their problems are my problems. In Tessier, C., Chaudron, L., and Müller, H.-J., editors, *Conflicting agents - Conflict management in multi-agent systems*. Kluwer Academic Publishers.

[Hansen, 2006a] Hansen, J. (2006a). Deontic logics for prioritized imperatives. *Artificial Intelligence and Law*, 14(1-2):1–34.

[Hansen, 2006b] Hansen, J. (2006b). The paradoxes of deontic logic: Alive and kicking. *Theoria*, 72(3):221–232.

[Hardin and Goodrich, 2009] 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*, pages 165–172.

[Hare, 1952] Hare, R. (1952). *The Language of Morals*. Oxford University Press.

[Harman and Kulkarni, 2011] Harman, G. and Kulkarni, S. (2011). Philosophy of statistics. In Bandyopadhyay, P. and Forster, M., editors, *Statistical Learning Theory as a Framework for the Philosophy of Induction*, pages 333–338. Elsevier.

[Harman and Kulkarni, 2012] Harman, G. and Kulkarni, S. (2012). Encyclopedia of the sciences of learning. In Seel, N., editor, *Statistical Learning Theory and Inductions*, pages 3186–3188. Springer.

[Heersmink et al., 2011] Heersmink, R., Van den Hoven, J., Van Eck, J.-N., and Van den Berg, J. (2011). Bibliometric mapping of computer and information ethics. *Ethics and Information Technologies*, 13(3):241–249.

[Hess, 1999] Hess, T. (1999). *A Study of Autonomous Agents in Decision Support Systems*. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg.

[Hoc, 2000] Hoc, J. (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, 43(7):833–843.

[Hoffman et al., 2009] Hoffman, K., Zage, D., and Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Survey*, 42(1):1–31.

[Hofweber, 2014] Hofweber, T. (2014). Logic and ontology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

[Hollnagel and Woods, 1983] Hollnagel, E. and Woods, D. (1983). Cognitive systems engineering: New wine in new bottles. *International Journal of Man-Machine Studies*, 18(6):583–600.

[Honarvar and Ghasem-Aghaee, 2009] Honarvar, A. and Ghasem-Aghaee, N. (2009). An artificial neural network approach for creating an ethical artificial agent. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 290–295.

[Horling and Lesser, 2004] Horling, B. and Lesser, V. (2004). A survey of multi-agent organizational paradigms. *American Society for Information Science and Technology*, 55(9):783–793.

[Horty, 1994] Horty, J. (1994). Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65.

[Hubner et al., 2002] Hubner, J., Sichman, J., and Boissier, O. (2002). Moise+: Towards a structural, functional, and deontic model for the MAS organization. In *1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

[Jones, 2008] Jones, J. (2008). ALFUS - Autonomy Levels For Unmanned Systems Framework. Technical report, ALFUS Working Group. http://sstc-online.org/2008/pdfs/JBJ2079.pdf.

[Josang et al., 2007] Josang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, 43(2):618–644.

[Kant, 1792] Kant, E. (1792). *Fondements de la métaphysique des mœurs*. Les classiques des sciences sociales.

[Kant, 1988] Kant, E. (1988). *Théorie Et Pratique / Droit De Mentir: sur Un Prétendu Droit De Mentir Par Humanité (Bibliothèque Des Textes Philosophiques) (French Edition)*. Vrin.

[Kwok and Yang, 2004] Kwok, S. and Yang, C. (2004). Searching the peer-to-peer networks: The community and their queries. *The Knowledge Engineering Review*, 19:281–316.

[Lacand and Fink, 1966] Lacand, J. and Fink, B. (1966). *Ecrits: the first English edition*. W. W. Norton and Company.

[Lemaître and Verfaillie, 2007] Lemaître, M. and Verfaillie, G. (2007). Interaction between reactive and deliberative tasks for on-line decision-making. In *ICAPS'07 Workshop on Planning and Plan Execution for Real-World Systems*.

[Lemaître and Excelente, 1998] Lemaître, C. and Excelente, C. (1998). Multi-agent organization approach. In *2nd Iberoamerican Workshop on Distributed Artificial Intelligence and Multi-Agent Systems*.

[Lewis, 1989] Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society*, 63:113–137.

[Lloyd, 2005] Lloyd, D. (2005). *Cases in Medical Ethics and Law*. Cambridge University Press.

[Luzeaux, 2013] Luzeaux, D. (2013). SoS and large-scale complex systems architecting. In *4th International Conference on Complex Systems Design and Management*, pages 39–49.

[Marsh, 1994] Marsh, S. (1994). *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling.

[Marti and Garcia-Molina, 2005] Marti, S. and Garcia-Molina, H. (2005). Taxonomy of trust: categorizing P2P reputation systems. *Computer Networks*, 50:472–484.

[Massin, 2008] Massin, O. (2008). *Introduction Ã la Philosophie Morale*. Swiss Philosophical Preprint Series.

[Mathieson, 2007] Mathieson, K. (2007). Dioptra: An ethics decision support system. In *13th Americas Conference on Information Systems*.

[Mayer et al., 1995] Mayer, R., Davis, J., and Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

[McConnell, 2014] McConnell, T. (2014). Moral dilemmas. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

[McLaren, 2003] McLaren, B. (2003). Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence*, 150(1-2):145–181.

[McNamara, 2014] McNamara, P. (2014). Deontic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

[Mercier, 2011] Mercier, S. (2011). *Contrôle du partage de l'autorité dans un système d'agents hétérogènes*. PhD thesis, Institut Supérieur de l'Aéronautique et de l'Espace, Toulouse.

[Mercier et al., 2010] Mercier, S., Tessier, C., and Dehais, F. (2010). Détection et résolution de conflits d'autorité dans un système homme-robot. *Revue d'Intelligence Artificielle, numéro spécial 'Droits et Devoirs d'Agents Autonomes*, 24:325–356.

[Meredith and Arnott, 2003] Meredith, R. and Arnott, D. (2003). On ethics and decision support systems development. In *7th Pacific Asia Conference on Information Systems*, pages 1–14.

[Meyer, 2013] Meyer, M. (2013). *Principia moralia*. Fayard.

[Michel et al., 2010] Michel, J.-B., Shen, Y., Aiden, A., Veres, A., Gray, M., Brockman, W., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., and Aiden, E. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, published online ahead of print: 12/16/2010.

[Milligan et al., 2011] Milligan, C., Roberts, C., and Mort, M. (2011). Telecare and older people: Who cares where? *Social Science and Medicine*, 72(3):347–354.

[Müller and Dieng, 2000] Müller, H. and Dieng, R. (2000). *Computational Conflicts: Conflict Modeling for Distributed Intelligent Systems*. Springer.

[Muller and Pischel, 1993] Muller, J. and Pischel, M. (1993). The agent architecture InteRRap : Concept and application. Technical report, Deutsches Forschungszentrum fur Kunstliche Intelligenz.

[Nogueira, 2004] Nogueira, L.-C. (2004). Aula: A pesquisa em psicanálise. Technical report, Instituto de Psicologia USP de São Paulo.

[Nwana et al., 1996] Nwana, H. S., Lee, L. C., and Jennings, N. R. (1996). Coordination in software agent systems. *The British Telecom Technical Journal*, 14(4):79–88.

[Okada et al., 2007] Okada, M., Yamamoto, K., and Watanabe, K. (2007). Conceptual model of health information ethics as a basis for computer-based instructions for electronic patient record systems. *Studies in Health Technology and Informatics*, 129:1442–1446.

[Omicini et al., 2008] Omicini, A., Ricci, A., and Viroli, M. (2008). Artifacts in the A&A meta-model for multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 17(3):432–456.

[Orwell, 1972] Orwell, G. (1972). *1984 (French Edition)*. Gallimard French.

[Pariente-Butterlin, 2012] Pariente-Butterlin, I. (2012). Les dilemmes éthiques selon David Lewis : Est-il possible d'échapper au paradigme kantien ? *Klesis*, 24:311–325.

[Perennou, 2014] Perennou, T. (2014). Ethics and autonomous agents: State-of-the art on legal issues. Technical report, École Télécom Management.

[Pesty et al., 1997] Pesty, S., Batard, E., Brassac, C., Delépine, L., Gleizes, M.-P., Glize, P., Labbani, O., Lenay, C., Marcenac, P., Magnin, L., Müller, J.-P., Quinqueton, J., and Vidal, P. (1997). Emergence et SMA. In *5e Journées Francophones pour l'Intelligence Artificielle Distribuée et les Systèmes Multi-Agents*.

[Petit, 2004] Petit, P. (2004). Conséquentialisme. In Canto-Sperber, M., editor, *Dictionnaire d'éthique et de philosophie morale*. Puf. quadrige edition.

[Picard and Glize, 2006] Picard, G. and Glize, P. (2006). Model and analysis of local decision based on cooperative self-organization for problem solving. *Multiagent and Grid Systems*, 2(3):253–265.

[Picard, 1997] Picard, R. (1997). *Affective Computing*. MIT Press.

[Pinyol and Sabater-Mir, 2013] Pinyol, I. and Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25.

[Pizziol, 2013] Pizziol, S. (2013). *Conflict prediction in human-machine systems.* PhD thesis, Institut Supérieur de l'Aéronautique et de l'Espace, Toulouse.

[Pizziol et al., 2014] Pizziol, S., Tessier, C., and Dehais, F. (2014). Petri net-based modelling of human-automation conflicts in aviation. *Ergonomics.* DOI: 10.1080-00140139.2013.877597.

[Platon, 2002] Platon (2002). *La République (French Edition).* Flammarion.

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

[Powers, 2005] Powers, T. (2005). Deontological machine ethics. Technical report, American Association for Artificial Intelligence.

[Powers, 2006] Powers, T. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, 21(4):46–51.

[Putterman, 1994] Putterman, M. (1994). *Markov Decision Processes. Discrete stochastic dynamic programming.* Wiley-Interscience.

[Rao and Georgeff, 1991] Rao, A. and Georgeff, M. (1991). Modeling rational agents within a BDI-architecture. In *2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484.

[Resnick et al., 2000] Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. *ACM Communications*, 43(12):45–48.

[Robbins and Wallace, 2007] 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.

[Rodriguez-Moreno et al., 2007] Rodriguez-Moreno, M., Brat, G., Muscettola, N., and Rijsman, D. (2007). Validation of a multi-agent architecture for planning and execution. In *18th International Workshop on Principles of Diagnosis*, pages 368–371.

[Rotolo and van der Torre, 2011] Rotolo, A. and van der Torre, L. (2011). Rules, agents and norms: Guidelines for rule-based normative multi-agent systems. In *5th International Symposium Rule-Based Reasoning, Programming, and Applications*, pages 59–66.

[Russell and Norvig, 1995] Russell, S. and Norvig, P. (1995). *Artificial Intelligence: a modern approach*. Prentice Hall.

[Sadri, 2011] Sadri, F. (2011). Ambient intelligence: A survey. *ACM Computing Surveys*, 43(4):36–90.

[Sandholm et al., 1999] Sandholm, T., Larson, K., Andersson, M., Shehory, O., and Tohmé, F. (1999). Coalition structure generation with worst case guarantees. *Artificial Intelligence*, 111(1-2):209–238.

[Saptawijaya and Pereira, 2014] Saptawijaya, A. and Pereira, L. (2014). Towards modeling morality computationally with logic programming. In *16th International Symposium on Practical Aspects of Declarative Languages*, pages 104–119.

[Sartre, 2002] Sartre, J.-P. (2002). *L'existentialisme est un humanisme (French Edition)*. Folio.

[Schillo et al., 2002] Schillo, M., Fley, B., Florian, M., Hillebrandt, F., and Hinck, D. (2002). Self-organization in multiagent systems: from agent interaction to agent organization. In *3rd International Workshop on Modelling Artificial Societies and Hybrid Organizations*, pages 37–46.

[Scott, 1998] Scott, W. (1998). *Organizations: rational, natural and open systems*. Prentice Hall.

[Serugendo et al., 2006] Serugendo, G., Gleizes, M.-P., and Karageorgos, A. (2006). Self-organisation and emergence in MAS: An overview. *Informatica*, 30:45–54.

[Sheridan and Verplank, 1978] Sheridan, T. and Verplank, W. (1978). Human and computer control of undersea teleoperators. Technical report, MIT Man-Machine Systems Laboratory.

[Shiloni et al., 2009] Shiloni, A., Agmon, N., and Kaminka, G. (2009). Of robot ants and elephants. In *8th International Conference on Autonomous Agents and Multiagent Systems*, pages 81–88.

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

[Sichman et al., 1994] Sichman, J., Conte, R., Demazeau, Y., and Castelfranchi, C. (1994). A social reasoning mechanism based on dependence networks. In *11th European Conference on Artificial Intelligence*, pages 188–192.

[Sichman et al., 2005] Sichman, J., Dignum, V., and Castelfranchi, C. (2005). Agent organizations: a concise overview. *Special Issue in Agent Organizations in the Journal of the Brazilian Computer Society*, 11(1).

[Simon, 1990] Simon, H.-A. (1990). Invariants of human behavior. *Annual Review on Psychology*, 41:1–19.

[Sinnott-Armstrong, 2014] Sinnott-Armstrong, W. (2014). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

[Stradella et al., 2012] Stradella, E., Salvini, P., Pirni, A., Di Carlo, A., Oddo, C.-M., Dario, P., and Palmerini, E. (2012). Subjectivity of autonomous agents: Some philosophical and legal remarks. In *1st Workshop on Rights and Duties of Autonomous Agents*, pages 24–31. CEUR Proceedings Vol. 885.

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

[Tambe et al., 2008] 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.

[Tappolet, 2004] Tappolet, C. (2004). Dilemmes moraux. In Canto-Sperber, M., editor, *Dictionnaire d'éthique et de philosophie morale*. Puf. quadrige edition.

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

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

[Thrun, 2010] Thrun, S. (2010). Toward robotic car. *Communications of the ACM*, 53(4):99–106.

[Truszkowski et al., 2009] 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.

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

[Turkle and Shapiro, 2011] Turkle, S. and Shapiro, A. (2011). Social robots raise moral, ethical questions. Morning Edition.

[Ullah et al., 2012a] Ullah, I., Doyen, G., Bonnet, G., and Gaiti, D. (2012a). A bayesian approach for user aware peer-to-peer video streaming systems. *Signal Processing: Image Communication Special Issue on Advances in Video Streaming for P2P Network*, 27(5):438–456.

[Ullah et al., 2012b] Ullah, I., Doyen, G., Bonnet, G., and Gaiti, D. (2012b). A survey and synthesis of user behavior measurements in video streaming systems. *IEEE Communications Surveys and Tutorials*, 14(3):734–749.

[van Fraassen, 1973] van Fraassen, B. (1973). Values and the heart's command. *Journal of Philosophy*, 70:5–19.

[Veruggio, 2006] Veruggio, G. (2006). EURON Roboethics Roadmap. Technical report, EURON Working Group.

[von Krogh et al., 2012] von Krogh, G., Haefliger, S., Spaeth, S., and Wallin, M. (2012). Carrots and rainbows: Motivation and social practive in open source software development. *MIS Quarterly*, 36(2):649–676.

[von Wright, 1951] von Wright, G. (1951). Deontic logic. *Mind*, 60(237):1–15.

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

[Wang and Vassileva, 2003] Wang, Y. and Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks. In *3rd International Conference on Peer-to-Peer Computing*, pages 150–158.

[Werger, 1999] Werger, B. (1999). Cooperation without deliberation : A minimal behavior-based approach to multi-robot teams. *Artificial Intelligence*, 110:293–320.

[Weyns et al., 2007] Weyns, D., Omicini, A., and Odell, J. (2007). Environment as a first class abstraction in multiagent systems. *Autonomous Agents and Multi-agent Systems*, 14:5–30.

[Whitworth, 2006] Whitworth, B. (2006). Socio-technical systems. *Encyclopedia of human computer interaction*, pages 553–541.

[Wolf and Holvoet, 2004] Wolf, T. D. and Holvoet, T. (2004). Emergence and self-organisation: a statement of similarities and differences. In *2nd International Workshop on Engineering Self-Organising Application*, pages 96–110.

[Wooldridge and Jennings, 1995] Wooldridge, M. and Jennings, N. (1995). Agent theories, architectures and languages: a survey. In Wooldridge, M. and Jennings, N., editors, *Intelligent Agents*, pages 1–22. Springer-Verlag.

[Yanco and Drury, 2004] Yanco, H. and Drury, J. (2004). Classifying human-robot interaction: an updated taxonomy. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 2841–2846.