Università degli Studi di Salerno

DIPARTIMENTO DI MATEMATICA
Ph.D. Course in Mathematics, Physics and Application

PH.D. THESIS

A Game Theoretical Approach to Safe Decision Making System Development for Autonomous Machines

Supervisor:
prof. Vincenzo Tibullo

Vincenzo Tibullo

Ph.D. Candidate:
dott. Gianpiero Negri

Gianpiero Negri

Coordinator:
prof. Carmine Attanasio

Ph.D. in Mathematics, Physics and Applications – XXXIII Cycle

A mia madre, a mio padre

A Nina

The unknown future rolls toward us. I face it, for the first time, with a sense of hope. Because if a machine, a Terminator, can learn the value of human life, maybe we can too.
(James Cameron, "Terminator 2: judgment day", Santa Monica, Calif: Artisan Home Entertainment, 2003)

One Ring to bring them all and in the darkness bind them, in the Land of Mordor where the Shadows lie.
(J.R.R. Tolkien, "The Lord of the Rings", Ballantine Books, Copyright 1954-1974)

Cooper: Hey TARS, what's your honesty parameter?
TARS: 90 percent.
Cooper: 90 percent?
TARS: Absolute honesty isn't always the most diplomatic nor the safest form of communication with emotional beings.
Cooper: Okay, 90 percent it is.
(Christopher Nolan, "Interstellar", Paramount Pictures, 2014)

Ma videmus nunc per speculum et in aenigmate e la verità, prima che faccia a faccia, si manifesta a tratti (ahi, quanto illeggibili) nell'errore del mondo, così che dobbiamo compitarne i fedeli segnacoli, anche là dove ci appaiono oscuri e quasi intessuti di una volontà del tutto intesa al male.
(Umberto Eco, "Il Nome della Rosa", Fabbri-Bompiani, Milano, Italy, 1980)

Joshua: Greetings, Professor Falken.
Stephen Falken: Hello, Joshua.
Joshua: A strange game. The only winning move is not to play. How about a nice game of chess?

(John Badham, "Wargames", United Artists, 1983)

Abstract

One of the major technological and scientific challenges in developing autonomous machines and robots is to ensure their ethical and safe behaviour towards human beings. When dealing with autonomous machines the human operator is not present, so that the overall risk complexity has to be addressed to machine artificial intelligence and decision-making systems, which must be conceived and designed in order to ensure a safe and ethical behaviour. In this work a possible approach for the development of decision-making systems for autonomous machines will be proposed, based on the definition of general ethical criteria and principles. These principles concern the need to avoid or minimize the occurrence of harm for human beings, during the execution of the task the machine has been designed for. Within this scope, four fundamental problems can be introduced:

1. First Problem: Machine Ethics Principles or Laws Identification
2. Second Problem: Incorporating Ethics in the Machine
3. Third Problem: Human-Machine Interaction Degree Definition
4. Fourth Problem: Machine Misdirection Avoidance.

This Ph.D. research activity has been mainly focused on First and Second Problems, with specific reference to safety aspects. Regarding First Problem, main scope of this work is on ensuring that an autonomous machine will act in a safe way, that is:

- No harm is issued for surrounding human beings (non maleficence ethical principle)
- In case a human being approaching a potential source of harm, the machine must act in such a way to minimize such harm with the best possible and available action (non-inaction ethical principle) and, when possible and not conflicting with above principles:
- The machine must act in such a way to preserve its own integrity (self-preservation).

Concerning Second Problem, the simplified version of some ethical principles reported above has been used to build a mathematical model of a safe decision system based on a game theoretical approach. When dealing just with safety and not with general ethics, it is possible to adopt some well-defined criteria in ensuring the machine behaviour is not issuing any harms towards human beings, such as:

- Always ensure the machine is keeping a proper safety distance at a certain operating velocity
- Always ensure that, within a certain range, the machine can detect the distance between a human being and the location of a potential harm.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Key Terms and Definitions

Prior to move forward with the description of this Ph.D. thesis scope, some key terms and definitions will be provided, for easier reading and better understanding of the main concepts.

Autonomous Machine: A machine capable of performing tasks in the world by itself, without explicit human control.

Biological AI: An organism based on both biological and artificial components.

Ethical Governor: An arbiter of system-generated action to ensure that it constitutes an ethically permissible action.

Ethical Principle: A general judgment serving as a justification for ethical prescriptions or evaluation of human behaviour and actions.

Game Theory: The study of mathematical models of strategic interaction among rational decision-makers.

Machine Decision Making System: Part of the machine architecture capable of evaluating and selecting among a list of possible options the action to be executed by the machine at a certain time.

Golem: An artificial creature, being brought to life by supernatural or technological means.

Machine Ethics: A part of the ethics concerned with adding ethical behaviour or embedding ethical principles in a machine endowed with artificial intelligence.

Machine Ethical Risk Index: A measure of the extent to which a machine fails to meet its ethical principles.

Neuromorphic Artificial Intelligence: An artificial Intelligence implemented by means of electronic analog circuits capable to mimic neurobiological architectures present in the nervous system.

Superintelligence: An intellect outperforming human brain in all possible domains, showing higher capabilities and skills, including creativity and wisdom.

Weaponized AI: An evil, malicious or destructive autonomous system endowed with artificial intelligence.

1.2 Autonomous Machine Ethics and Safety: an Overview

One of the major technological and scientific challenges in developing autonomous machines and robots is to ensure their ethical and safe behaviour towards human beings. When dealing with autonomous machines the human operator is not present, so that the overall risk complexity has to be addressed to machine artificial intelligence and decision-making systems, which must be conceived and designed in order to ensure a safe and ethical behaviour. In this work a possible approach for the development of decision making systems for autonomous machines will be proposed, based on the definition of general ethical criteria and principles. These principles concern the need to avoid or minimize the occurrence of harm for human being, during the execution of the task the machine has been designed for. Within this scope, four fundamental problems can be introduced:

1. First Problem: Machine Ethics Principles or Laws Identification
2. Second Problem: Incorporating Ethics in the Machine
3. Third Problem: Human-Machine Interaction Degree Definition
4. Fourth Problem: Machine Misdirection Avoidance.

This Ph.D. research activity has been mainly focused on First and Second Problems, with specific reference to safety aspects. Regarding First Problem, main scope of this work is on ensuring that an autonomous machine will act in a safe way, that is:

- No harm is issued for surrounding human beings (non maleficence ethical principle)
- In case a human being approaching a potential source of harm, the machine will act to minimize such harm with the best possible and

available action (non-inaction ethical principle) and, when possible and not conflicting with above principles:

- The machine must act in such a way to preserve its own integrity (self-preservation).

Concerning Second Problem, the simplified version of some ethical principles reported above has been used to build a mathematical model of a safe decision system based on a game theoretical approach. When dealing just with safety and not with general ethics, it is possible to adopt some well-defined criteria in ensuring the machine behaviour is not issuing any harms towards human beings, such as:

- Always ensure the machine is keeping a proper safety distance at a certain operating velocity
- Always ensure that, within a certain range, the machine can detect the distance between a human being and the location of a potential harm.

1.3 Ph.D. Thesis Structure

This Ph.D. thesis consists of the following chapters:

1. The present "Introduction", which provides information on the background and the scope of this Ph.D. thesis, some main definition and basic concepts, as well as a description of the thesis structure
2. "Autonomous Machines: Ethics and Safety Problem Statement" will describe in details the four fundamental problems mentioned in this

chapter, providing some relevant references to current research concerning the establishment of an ethical framework for artificial intelligent autonomous agents

3. "Game Theory Overview and Application to Autonomous Machines Decision Making System" will describe some main aspects of game theory, as well as its application to the development of decision making systems of autonomous machines

4. "Game Theoretical Approach to Safe Decision Making System Development for Autonomous Machines: Mathematical Modelling" will describe the mathematical approach used in this Ph.D research work, by identifying the strategy sets and the payoff functions used to setup the game

5. "Algorithm Design and Simulation Results" will provide a detailed description of all algorithms developed, along with the main results of the Monte Carlo simulation executed on the selected scenario

6. Finally, the "Conclusion" chapter consists in a summary of the main topics discussed and results obtained within the scope of this Ph.D. research, along with a description of some possible research directions to be investigated for further improvements.

Chapter 2

Autonomous Machines: Ethics and Safety Problem Statement

In this chapter the Four Problems mentioned in Chapter 1 will be described in more details, providing some relevant references to current research concerning the establishment of an ethical framework for artificial intelligent autonomous agents.

2.1 Machine Ethics - Introduction

On traditional vehicles and mobile machines, domain-specific standards are introduced in order to reduce the safety risk, under the basic assumption that main control is up to the human operator: this assumption allows to heavily reduce machine complexity. Designers analyze all machine functions, identifying potential hazards, and then evaluate the risk, by determining the requirements for actual functions development. However, in autonomous systems, such as mobile robots, the human operator is not supposed to be always present, so that machine decision-making system must entirely manage the overall risk complexity. Therefore, in order to avoid causing harms to human beings, an autonomous machine must be endowed with some judgment capabilities, which can drive its decision-making process, based on some fundamental ethical principles. Indeed, main goal of machine ethics is to provide the artificial agents with this set of ethical principles, guiding them in their decisions. According to [1], decision making process is one of the most critical aspects in the development of autonomous machines, as, despite of the availability of information and data, it is unlikely that an ethical behaviour will emerge spontaneously in a machine: that leads to the challenge of identifying first suitable ethical principles, then ensuring their embodiment in the machine architecture. An interesting thought experiment can be introduced on this topic: let's consider a super powerful golem genie, materializing in front of a human being, and telling him that in 50 years it will return to the same place, and asking him to supply it, upon its return, with a set of moral principles. It will then follow those principles consistently and rigidly throughout the universe [2]. Therefore, it is up to the human being, or better to humankind, to ensure that the principles are not faulty and perfectly suitable. In order to translate this philosophical standpoint into requirements of an ethical, autonomous decision-making system, in this chapter four fundamental problems will be considered:

1. First Problem: Machine Ethics Principles Identification

2. Second Problem: Incorporating Ethics in the Machine

3. Third Problem: Human-Machine Interaction Degree Definition
4. Fourth Problem: Machine Misdirection Avoidance.

First Problem: How to define machine ethical principles? When dealing with robotics and autonomous systems ethical behaviour, Asimov's Laws of Robotics [3] are often a starting point, even though they have been subject to criticism due to vagueness and lack of completeness. Nowadays it is possible to refer to several alternative sources to find out a more recent foundation of ethical principles for robotics and AI, such as IEEE General Principles of Ethical Autonomous and Intelligent Systems (A/IS) [4], or Floridi and Clement Jones five principles key to any ethical framework for AI [6]. Coming to **Second Problem**, once identified the proper set of ethical principles, there is the need to embed it in the machine, implementing a so-called ethical machine governor. Main challenge is to translate the set of identified ethical principles in some quantitative form, and in order to do that, one must define some indicators of the machine behaviour. Regarding artificial intelligence in general, it is expected to increasingly introduce a massive disruption in the society: according to [7]: "It poses a multifaceted problem when it comes to designing and understanding regulatory responses to AI." For what concerning specifically safety, commonly taken into account also in contemporary robots and machines development, several standards and regulations are already available, which require as first step to evaluate the presence of a safety risk for people in any tasks to be executed by the machines. Machine designers usually execute a hazard analysis and risk assessment to establish the requirements for the machine governing control system; the evaluation of the risk is based on well-defined parameters, so that its measure can be used as a guidance to design the machine safety features. One main challenge to face is how to setup or define similar measures for dignity, privacy or politeness when dealing with machine behaviour, in such a way to make possible to introduce and quantify a general machine ethical risk index. About **Third Problem**: how to identify a suitable interaction degree between humans and machines? E.g., should a machine be conditioned in ethically behaving based on a human input, or should this behaviour be decided only by artificial decision-making system? Finally, the **Fourth Problem** consists in ensuring the avoidance of machine misdirection (e.g. due to a cyberattack), by introducing, for instance, neuromorphic computing or biological artificial intelligence. As mentioned in Chapter 1, this Ph.D. thesis focuses on a possible machine ethics approach, based on the definition of some suitable ethical principles, which the machine builders should embed in the artificial decision-making system. These principles must concern the need to avoid the occurrence of harm, during the execution of any machine tasks, by solving possible conflicts issued during decision-making process.

2.2 Background on Machine Ethics

According to Cambridge Online Dictionary [8], a **machine** can be defined as an apparatus using mechanical power, having several parts, each with a definite function, and together performing a particular task. By the way, above definition is focused on traditional, mechanical equipment, whereas nowadays, due to the extraordinary growth of emerging digital technologies, a broader definition must be considered. For instance, according to the first indent of 2006/42/EC - Machinery Directive, Article 2(a) [9], machinery is defined as follows: *an assembly, fitted with or intended to be fitted with a drive system other than directly applied human or animal effort, consisting of linked parts or components, at least one of which moves, and which are joined together for a specific application.* However, in 2018, the European Commission started a review of the Machinery Directive itself, focused on machinery, which utilizes emerging digital technologies, such as artificial intelligence, internet of things, autonomous robots and so on. Hence, in what follows, the terms "machine", "artificial intelligence/AI", "artificial agent" will be used to identify an advanced autonomous system, endowed with artificial intelligence and a decision-making capability, including, e.g., ordinary physical machines, robots, as well as pure algorithms [10]. Moreover, in 2018, European Commission has established a High-Level Expert Group on Artificial Intelligence (AI HLEG) comprising representatives from academia, civil society, as well as industry. With the general objective to support the implementation of the European Strategy on Artificial Intelligence, providing recommendations on policy development and on ethical, legal and societal issues related to AI. Similarly, in 2020, the U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations, after 15 months of consultation with leading AI experts in commercial industry, government, and academia. Considering major hi-tech corporate standpoint on these topics, a quite comprehensive overview has been provided by Thilo Hagendorff, who has listed some of the main initiatives [11]:

Information Technology Industry AI Policy Principles" (2017), the principles of the "Partnership on AI" (2018), the IEEE first and second version of the document on "Ethically Aligned Design" (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016, 2019), as well as the brief principle lists of Google (2018), Microsoft (2019), DeepMind (DeepMind), OpenAI (2018), and IBM (Cutler et al., 2018) which have become well-known through media coverage. Other large companies such as Facebook or Twitter have not yet published any systematic AI guidelines, but only isolated statements of good conduct.

These are some clear evidences of current and potential impact of above emerging technologies in all aspects of everyday life: it is reasonable to expect that they will help in replacing human beings in the execution of repetitive or dangerous tasks, or will be employed in disaster recovery, or simplifying services, which could be accessed more easily or comfortably. According to Virginia Dignum, AI HLEG member, artificial intelligence will be very soon among the human beings, in different forms, e.g. service, transportation, medical and military robots [12]. As artificial agents are more and more required to make decisions with direct impact on human society, one of the most critical upcoming research challenge is to "integrate moral, societal and legal values with technological developments in AI, both during the design process as well as part of the deliberation algorithms employed by these systems." [13]. A more radical standpoint on these issues from Luke Muehlhauser and Louie Helm is that: *Self-improving artificial intelligence (AI) could become so vastly more powerful than humans that we would not be able to stop it from achieving its goals. If so, and if the AI's goals differ from* *ours, then this could be disastrous for humans.* [2] This leads to the need to include, during machine design phase, some key principles in its program, which can guide it in acting according to the principles themselves. A similar point of view has been expressed by Nick Bostrom and Eliezer Yudkowsky [14], mentioning that the possibility of creating "thinking machines" must be always accompanied by adequate measures to ensure such machines do not cause harms to human beings. Another AI HLEG member, Luciano Floridi, is also conducting an extensive research activity on this subject, by identifying in particular the need to develop laws, corporate policies, standards and best practices to ensure that artificial agents development and deployment will be beneficial for humanity [15]. In an effort to define, if possible, a unique, general framework capable to include all the relevant sources, Floridi has highlighted the crucial relevance of harmonizing those different initiatives he analyzed:

- The Asilomar AI Principles, developed under the auspices of the Future of Life Institute, in collaboration with attendees of the high-level Asilomar conference of January 2017 [16]

- The Montreal Declaration for Responsible AI, developed under the auspices of the University of Montreal, following the Forum on the Socially Responsible Development of AI of November 2017 [4]

- The General Principles offered in the second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems [5]

- The Ethical Principles offered in the Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, published by the European Commission's European Group on Ethics in Science and New Technologies, in March 2018 [17]

- The 'five overarching principles for an AI code' offered in UK House of Lords Artificial Intelligence Committee's report, AI in the UK: ready, willing and able?, published in April 2018 [18]

- The Tenets of the Partnership on AI, a multi-stakeholder organization consisting of academics, researchers, civil society organisations, companies building and utilising AI technology, and other groups.

Once machine ethical principles have been identified, it is crucial as well to define a suitable strategy to incorporate those principles in the machine design. According to Norbert Wiener, "We need to be sure that the purpose put into the machine is the purpose which we really want" [20]. This requires as first an assumption of responsibility [21]:

1. **In Design**: ensuring that development processes take into account ethical and societal implications of AI as it integrates and replaces traditional systems and social structures

2. **By Design**: integrating ethical reasoning abilities as part of the behaviour of artificial autonomous systems

3. **For Designers**: Research integrity of researchers and manufacturers, and certification mechanisms.

Regarding responsibility for designers, a large set of norms have been introduced to ensure that machinery is developed in order to take into account essential health and safety requirements, allowing to reduce as much as possible the risk of harm for all surroundings human operators and bystanders. Speaking of traditional vehicles and automated mobile machinery, for example cars, agricultural equipment, construction equipment etc., some well-known and established standards are available, which can be applied to ensure the overall system safety: for instance, ISO 26262 [22] is commonly applied to automotive/passenger cars, and ISO 13849 [23] is used instead for machinery. Nevertheless, when moving towards advanced robotics and machines driven by artificial intelligence, there are some peculiar key aspects, increasing complexity in a relevant way with respect to traditional machines:

1. Advanced machines act autonomously, without expecting any inputs from human operators, and therefore

2. They must be independently capable of a safe behaviour, avoiding physical harms to humans.

As mentioned in Chapter 1, the safe behaviour should be considered as part of the overall set of capabilities a machine must have in order to claim its adherence to complete ethical framework. On the latter aspect, a relevant problem is the representation of ethics by means of a set of different features, including safety, privacy, dignity, politeness and so on, as described by Louise Abigail Dennis and Michael Fisher [24]: an ethical decision-making process should be supported by a suitable machine architecture, in which a number of different agents, or **reasoners**, are customized to reason about some particular ethical feature, e.g. safety, dignity, privacy and so on. The whole set of these reasoners constitutes the so called **ethical arbiter** or **ethical governor**.

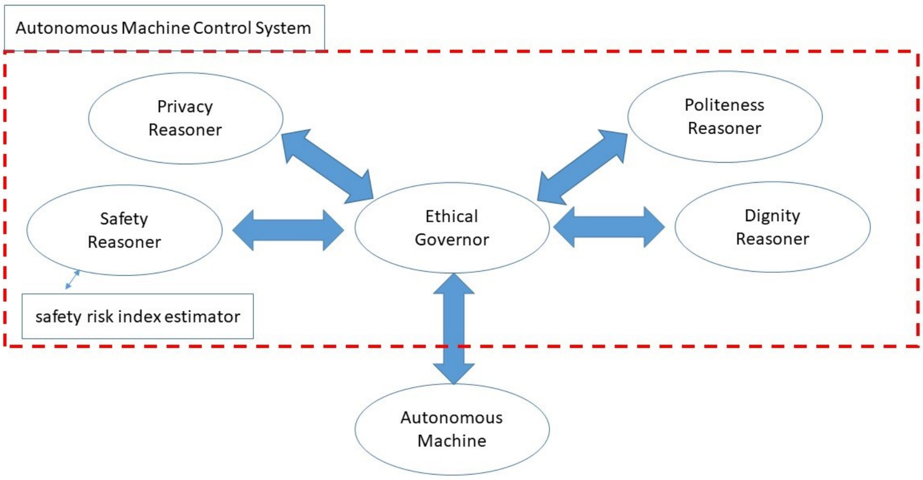


Figure 2.1: Autonomous Machine Ethical Governor and Features

During the execution of a certain task, the autonomous system communicates the set of possible options or choices to the ethical arbiter, which reasoners will must assess these options by converting them into logical or mathematical form. Finally, the ethical arbiter provides the result of this assessment back to the autonomous system. Furthermore, concerning above mentioned responsibilities by-design, it is also needed to identify some practical ways to translate high level, general ethical principles in machine-readable codes and programs, which can be implemented during design phases, so to connect ethics and machine implemented procedures. Several attempts to define a possible framework for this kind of implementation has been done, for instance introducing so-called deontic logic, defined as the field of philosophical logic that is concerned with obligation, permission, and related concepts or, alternatively, as a formal system trying to capture the essential logical features of these concepts, which can be eventually used to establish a framework for robot ethics [25]. Another relevant issue is related with the definition of human-machine interaction degree in the machine decision-making process, which relates with the Third Problem: on this topic, a notable research work has been conducted by Giuseppe Contissa, Francesca Lagioia and Giovanni Sartor (European University Institute), which introduced the concept of *ethical knob*, as a mechanism available to passenger to set the vehicle ethical behaviour [26]. Finally, once and if ethical principles have been defined or identified and properly implemented with a suitable procedure in a machine, another significant issue concerns the possibility of having machine behaviour misdirected, due for instance to an external cyberattack. On this aspect, machine artificial intelligence is supposed to be robust enough to detect and reject all the attempts to disrupt the service it provides. These issues fall within the scope of cybersecurity, defined as the state of being protected against the criminal or unauthorized use of electronic data, or the measures taken to achieve this, which capability must be available to the machine itself or to the intelligent infrastructure it is connected with [27].

2.3 Machine Ethics and Safety: Towards a Mathematical Modelling

One of the main challenges for incorporating ethical principles in autonomous machines relates with the need of defining a suitable set of ethical principles and their corresponding mathematical model. Coming back to Muehlhauser and Helm's Golem Genie example [2], in order to provide the artificial intelligence with suitable principles or laws, it must be assured the principles themselves to be non-contradictory, self-explanatory, clear and well described, in such a way to avoid that, due to an intrinsic vagueness or ambiguity, they could be misunderstood.

Indeed, the Golem legend itself provides deep insights on the ethical implication of building automata, as described by Zvi Harry Rappaport [28]:

In 16th Century Prague, Rabbi Loew created a Golem, a humanoid made of clay, to protect his community. When the Golem became too dangerous to his surroundings, he was dismantled. This Jewish theme illustrates some of the guiding principles in its approach to the moral dilemmas inherent in future technologies, such as artificial intelligence and robotics. Man is viewed as having received the power to improve upon creation and develop technologies to achieve them, with the proviso that appropriate safeguards are taken. Ethically, not-harming is viewed as taking precedence over promoting good. Jewish ethical thinking approaches these novel technological possibilities with a cautious optimism that mankind will derive their benefits without coming to harm.

Figure 2.2: Kafka and the Golem statue, Prague

As already mentioned in the previous sections, when dealing with robotics and autonomous machines, Asimov's Laws of Robotics [3]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm;

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law;

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

are often used as reference, even though it has been highlighted that these laws are not easy to interpret, and furthermore they do not consider the capability of an artificial intelligence to adapt its program to the environment it is interacting with, and not even the coordinated interaction between different machines, which can result in a swarm-like behaviour [29]. Due to Asimov's laws incompleteness and inability to enclose some contemporary issues related to advanced robotics or artificial intelligence, several alternative laws or principles have been proposed. For instance, Robin R. Murphy and David D. Woods have introduced a parallel set of laws of responsible robotics, to highlight the robots accountability in their interactions with people, and with a special focus on the responsibilities of designers to ensure an adequate machine system safety in order to avoid harms to human beings [30]. Furthermore, a group of researchers from Google Brain, Stanford University, UC Berkeley and OpenAI, have proposed a set of concrete problems in AI safety, in an attempt to overcome the above-mentioned incompleteness and vagueness of Asimov's laws, and provide designers with some practical methods and approaches for designing machines [31]. As reported in the previous sections, Luciano Floridi [15] has also provided some main contributions in founding a machine ethics general approach, by proposing five key principles for any AI ethical framework:

1. Beneficence: AI must be beneficial to humanity.
2. Non-maleficence: AI must also not infringe on privacy or undermine security.
3. Autonomy: AI must protect and enhance our autonomy and ability to

take decisions and choose between alternatives.

4. Transparency: AI must promote prosperity and solidarity, in a fight against inequality, discrimination, and unfairness.

5. Accountability: We cannot achieve all this unless we have AI systems that are understandable in terms of how they work (transparency) and explainable in terms of how and why they reach the conclusions they do (accountability).

Based on this framework, and in order to extend current approach to safety system design as introduced in the Machinery Directive and related harmonized standards to the special case of autonomous machines, a set of four main guidelines can be introduced, which can help in performing the translation of general principles into mathematical or logical equations:

1. **There are five key ethical principles: beneficence, non-maleficence, autonomy, transparency and accountability**

2. **Machine ethics can be represented as a set of different features: safety, dignity, politeness, privacy, etc.**

3. **Each feature must be expressed in quantitative terms, by means of specific risk indexes (e.g. machine safety risk); these indexes can be reviewed as individual contribution to a general machine ethical risk index**

4. **Each ethical principle must be expressed in quantitative terms as well, introducing the machine capability to adhere to that**

principle as a function of all the risk indexes related to the different machine ethics features.

Among all machine ethics features above mentioned, within this Ph.D. research work, safety has been identified as a first, relevant case, because, as already mentioned in the previous sections, contemporary machines must be already designed to be safe, and machine safety risk index can be already expressed in quantitative terms, as a combination of different parameters such as probability of exposure and severity of harm: this is actually the subject of the so-called **machine safety engineering** [19]. Indeed, the process of assessing machinery risks and apply related risk reduction measures is described in details within ISO 12100 [32]. A machine safe design must be ensured by analyzing all the possible hazards the machine can issues during the execution of its tasks, analyzing the risks for each of those hazards, described by means of suitable indexes, and introducing technical safety requirements and measures to reduce as much as possible the risks themselves. In this context, having introduced a suitable safety risk index, it's possible to state for instance that potential "machine maleficence” in terms of safety is adequately reduced when the machine is acting in such a way to keep its safety risk index as low as possible. This leads to the introduction of a quantitative relationship between the non-maleficence capability (or, to be more precise, the machine capability to adhere to non-maleficence ethical principle) and the safety risk index, so that the best ethical decision taken by the machine corresponds to the minimum value of the risk index itself. As mentioned in the introduction, once identified the proper set of ethical principles, there is the need to embed it in the machine: this will lead to the need of defining practical guidelines in translating ethical principles in quantitative relationships, in such a way to be able to make them understandable at a machine level. And indeed the process of taking an ethical decision by a machine implies a problem of minimizing a general machine ethical risk index, made up of different individual contributions, including safety. This encourages in considering mathematical approaches for modelling the ethical principles, represented in some quantitative form. One possible approach is based on game theory, which applies whenever the actions of several agents are interdependent and can be defined as the study of mathematical models of conflict and cooperation between intelligent rational decision-makers [33]. The purpose of study in game theory is game [36]. Players of the game are described in terms of their available actions, which influence on the game is known. There, players involved in a game are arranged in their preferences, their information, the strategic actions available to them, and how these influence the outcome. A high-level description of a game specifies only what payoffs each individual or group can obtain by assistance of its members. Within this scope, a machine can be considered as one of the players, and therefore it's possible to formulate its decision-making process as a game, where payoffs functions must be defined in such a way to take into account the task to be executed and the ethical principles the machine will obey. Above mentioned game theory framework has been used in this Ph.D. thesis in order to move from a pure philosophical description or definition towards a quantitative form and a mathematical model of the machine ethics principles, and to allow formulating the artificial agent decision-making process in terms of logic statements, optimization problems etc. which can be coded in a software program: a wider discussion and more details on this mathematical modelling approach will be provided in Chapter 3.

Chapter 3

Game Theory Overview and Application to Autonomous Machines Decision Making System

3.1 Autonomous Machines: Overview and Mathematical Modelling

Let's recall some useful definitions for the scope of this work. A **machine** can be defined as an apparatus using mechanical power and having several parts, each with a definite function and together performing a particular task. More generally, a machine can be defined as an apparatus used to perform a particular task. A machine can be termed as **autonomous**, if it can take decisions without human inspection, control or assistance: in other terms, autonomous machines are capable of performing tasks in the world by themselves, without explicit human control.

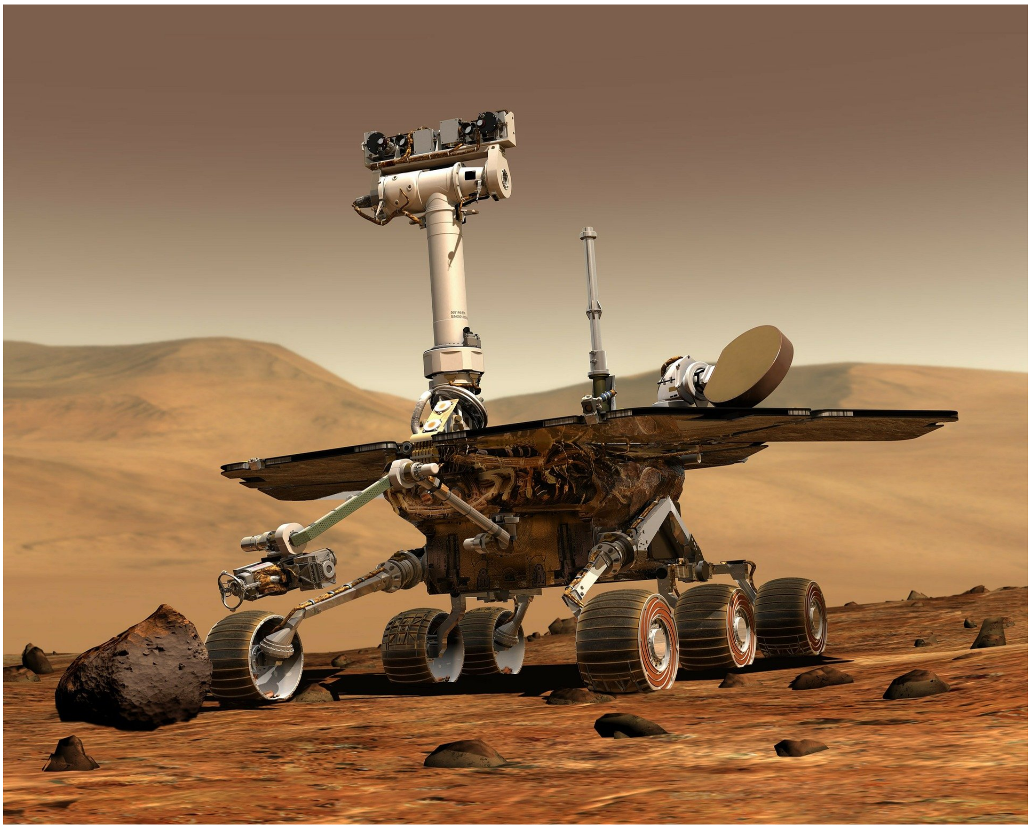


Figure 3.1: Autonomous robots: NASA Mars exploration Rover - WikiImage Pic from Pixabay

More specifically, a machine can be termed as a **robot**, if it is autonomous and if it agrees with the three laws stated by Isaac Asimov (see Section 2.3). Autonomous robots can be used in a large number of applications, including construction equipment, self driving cars and vehicles (see Fig. 3.3), hazardous waste management, household maintenance (see Fig. 3.2) etc.

Figure 3.2: Autonomous robots: a vacuum cleaning robot for household maintenance - pic from Pixabay

An autonomous machine must be able to perform specific tasks with a high degree of autonomy, representing an intersection between a number of different knowledge domains, such as artificial intelligence, robotics and information engineering. An autonomous machine is endowed with a set of different capabilities, including:

- Information and data retrieving from the environment
- Medium/long term autonomy, i.e. capability to execute a task without any human intervention
- Self propulsion, i.e. the machine contains its own means of motion (e.g. electric motors).

Figure 3.3: First self driving bus, from 2019 deployed on the public road in Berlin - Pic from Pixabay

More in general, an autonomous mobile robot must be able to:

- (Perception) Perceive its environment, by means of a suitable set of sensors (e.g. laser scanners, cameras, temperature sensors)
- (Decision-making) Take decisions based on the input from its perception system and/or other data or information evaluated or elaborated during the execution of its control program
- (Actuation) Actuate a travel or manipulation task, by interacting with its environment.

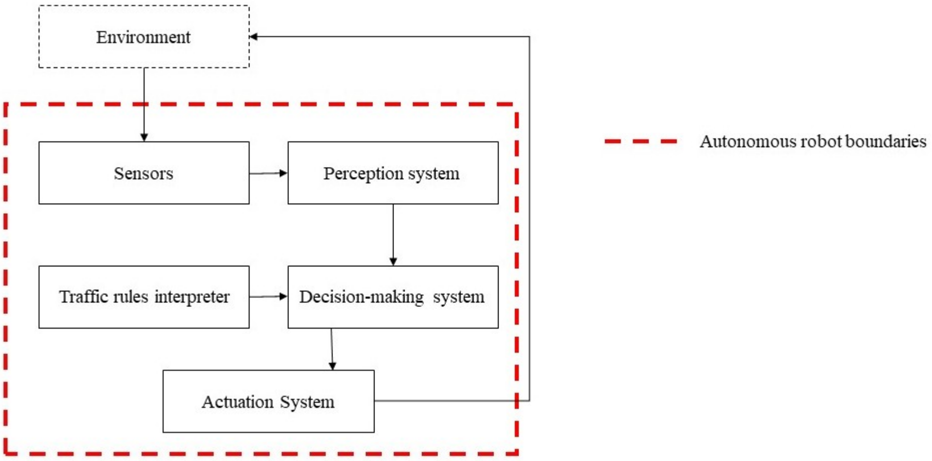
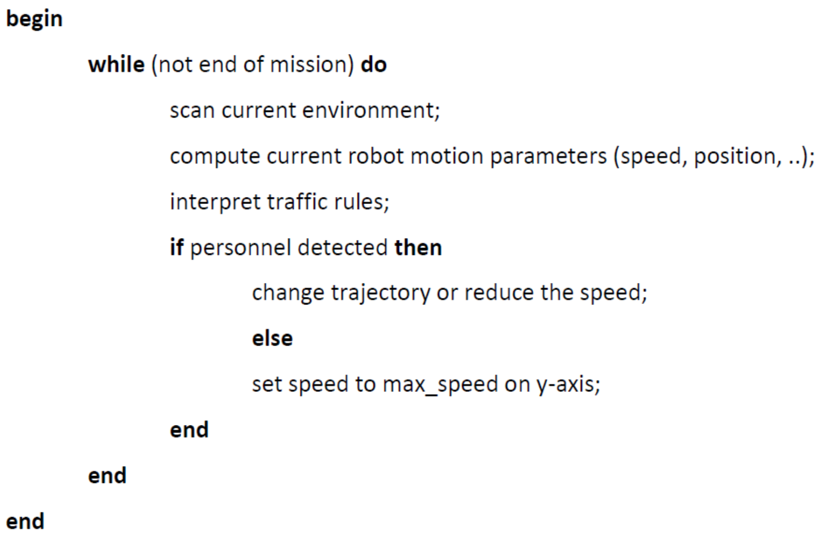


Figure 3.4: Autonomous machine control system: architectural block diagram

Some possible actions are: increase or decrease velocity, keep standing still, turn/change direction, pick, grab or place objects etc. Generally speaking, a control system for autonomous machines and robots can be very complex and can be divided in different subsystems [37]:

- perception system
- traffic rules interpreter
- decision making system or behaviour controller
- level car controller.

In the Fig. 3.5, a pseudocode example of the control program of the autonomous machine is showed.



```
begin
    while (not end of mission) do
        scan current environment;
        compute current robot motion parameters (speed, position, ..);
        interpret traffic rules;
        if personnel detected then
            change trajectory or reduce the speed;
            else
            set speed to max_speed on y-axis;
        end
    end
end
```

Figure 3.5: Autonomous machine control program: pseudocode example

Autonomous robots motion control system can be mathematically modeled depending on the specific application. Considering a mobile robot equipped with two fixed wheels and a centrally-oriented wheel, controlled with the help of electric motors and drives, combined into a single control system executive-enforcement mechanisms and a computing devices (see Fig. 3.6), its kinematic model can be described by the following equations [38]:

$$\begin{cases} \dfrac{dx}{dt} = v\cos\alpha \\ \dfrac{dy}{dt} = v\sin\alpha \\ \dfrac{d\alpha}{dt} = \omega \end{cases} \tag{3.1}$$

where:

x, y: coordinates of the robot relative to the fixed coordinate system;

α: robot orientation angle in space relative to the horizontal axis;

v, ω: linear and angular velocity of the robot.

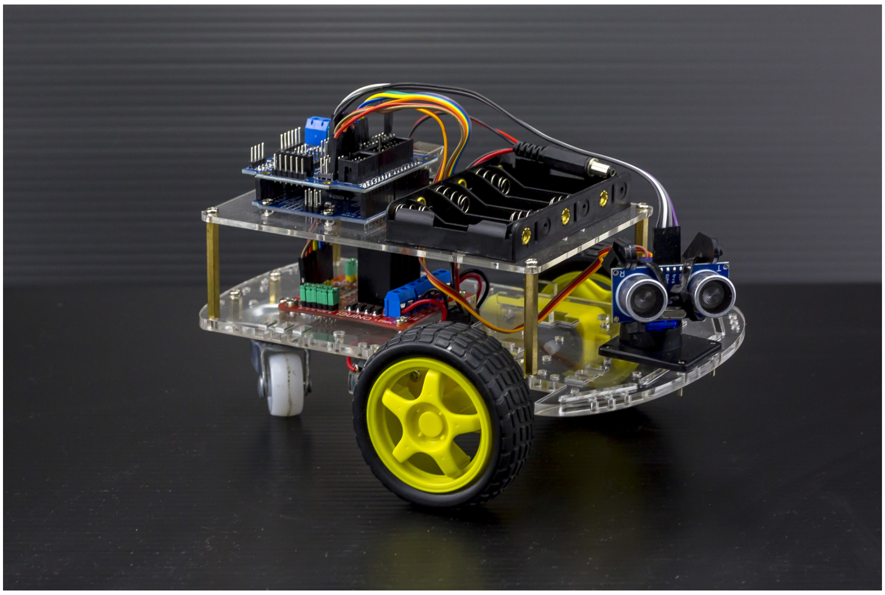


Figure 3.6: Example of a 2 Wheel mobile robot, built with Arduino Robot Kit (retrieved from `https://www.auselectronicsdirect.com.au/2-wheel-drive-ultrasonic-arduino-projects-robot-ki`)

In this specific case, motion control problem consists in determining the transformation leading the robot from the point of coordinates (x, y) and orientation angle α to the point with coordinates (x^*, y^*) and orientation angle α^*. Ideally, the task of achieving the specified coordinates will be accomplished if the achievement error of any controlled coordinates equals to zero. In reality, the achievement of the specified coordinates will be fulfilled, if errors of linear coordinates and orientation angle error are represented as a set of inequalities:

$$\begin{cases} |x^* - x| < e_x, \\ |y^* - y| < e_y, \\ |\alpha^* - \alpha| < e_\alpha \end{cases} \tag{3.2}$$

The control process is based on coordinate transformation expressed by means of the rotation matrix:

$$M(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.3}$$

Therefore, the coordinate transformation to be used by robot control system to generate the setpoint and actuate the motion from (x, y, α) to (x^*, y^*, α^*) can be expressed as:

$$\begin{pmatrix} x_1 \\ y_1 \\ \alpha_1 \end{pmatrix} = M(\alpha) \begin{pmatrix} x^* - x \\ y^* - y \\ \alpha^* - \alpha \end{pmatrix}. \tag{3.4}$$

where x_1, y_1, α_1 are the coordinates and orientation angle of the robot in new coordinate system. The motion control block diagram for this application is showed in Fig. 3.7:

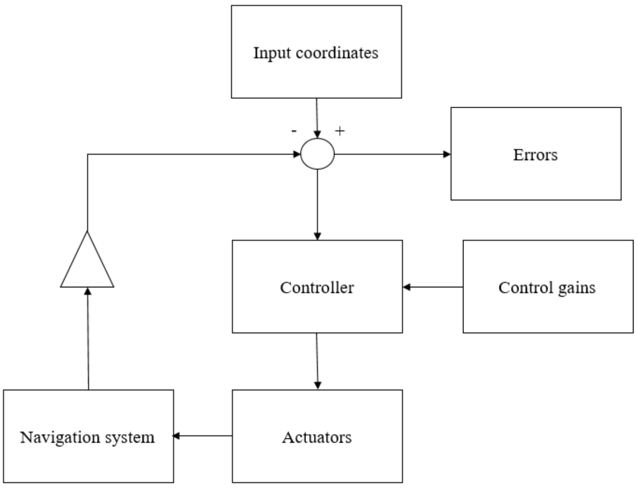


Figure 3.7: Autonomous machine control architecture: example of block diagram

3.2 Game Theory: an Overview

Game theory is the study of mathematical models of strategic interaction among rational decision-makers [33]. It has applications in all fields of social science, as well as in logic, systems science and computer science. The first known discussion of game theory occurred in a letter written by Charles Waldegrave, an active Jacobite, and uncle to James Waldegrave, a British diplomat, in 1713. Game theory did not really exist as a unique field until John Von Neumann published a paper in 1928 [34]: Von Neumann's original proof used Brouwer's fixed-point theorem on continuous mappings into compact convex sets, which became a standard method in game theory and mathematical economics. His paper was followed by his 1944 book Theory of Games and Economic behaviour co-authored with Oskar Morgenstern. In 1950, the first mathematical discussion of the prisoner's dilemma appeared, and an experiment was undertaken by notable mathematicians Merrill M. Flood and Melvin Dresher, as part of the RAND Corporation's investigations into game theory [35]. RAND (an American nonprofit global policy think tank created in 1948 by Douglas Aircraft Company to offer research and analysis to the United States Armed Forces) pursued the studies because of possible applications to global nuclear strategy. Around this same time, John Nash developed a criterion for mutual consistency of players strategies, known as Nash equilibrium, applicable to a wider variety of games than the criterion proposed by Von Neumann and Morgenstern. Game theory applies whenever the actions of several agents are interdependent [39]. According to [36], the purpose of study in game theory is **game**. There players involved in a game are arranged in their preferences, their information, the strategic actions available to them, and how these influence the outcome. A high level description of a game specifies only what payoffs each individual or group can obtain by assistance of its members. Game theory is generally divided into two branches:

- **Cooperative Game Theory**, focused on predicting which coalitions will form, the joint actions that groups take and the resulting collective payoffs

- **Non Cooperative Game Theory**, which studies and models conflict situations among economic agents; that is, it studies situations where the profits (gains, utility or payoffs) of each economic agent depend not only on his/her own acts but also on the acts of the other agents [40].

The so-called prisoner's dilemma is a standard example of a game analyzed in game theory that shows why two individuals X and Y might not cooperate, even if it appears that it is in their best interests to do so. Two men are arrested and imprisoned, and each prisoner is in solitary confinement with no means of communicating with the other. The prosecutors lack sufficient evidence to convict the pair on the principal charge. They hope to get both sentenced to a year in prison on a lesser charge. Simultaneously, the prosecutors offer each prisoner a bargain. Each prisoner is given the opportunity either to: betray the other by testifying that the other committed the crime, or to cooperate with the other by remaining silent. The offer is:

- If X and Y each betray the other, each of them serves 6 years in prison
- If X betrays Y but Y remains silent, X will be set free and Y will serve 7 years in prison (and vice versa)
- If X and Y both remain silent, both of them will only serve 1 year in prison (on the lesser charge)

It is implied that the prisoners will have no opportunity to reward or punish their partner other than the prison sentences they get, and that their decision will not affect their reputation in the future. Because betraying a partner offers a greater reward than cooperating with them, all purely rational self-interested prisoners would betray the other, and so the only possible outcome for two purely rational prisoners is for them to betray each other [41]. A game can be conveniently represented by means of a so-called **payoff matrix**, a bi-matrix reporting the payoffs of each player depending on the strategy used as response to the other player's strategy. For instance, the payoff matrix for the prisoner's dilemma can be represented as follows:

		Player Y	
		B	S
Player X	B	$(-6,-6)$	$(0,-7)$
	S	$(-7,0)$	$(-1,-1)$

Table 3.1: Prisoner's Dilemma: payoff matrix

where B: player betrays, S: player stay silent. Let's introduce the following definitions:

- A **strategy**, or **pure strategy**, is a complete algorithm for playing the game, telling a player what to do for every possible situation throughout the game.

- A **strategy profile** S (sometimes called a strategy combination) is a set of strategies for all players which fully specifies all actions in a game. A strategy profile must include one and only one strategy for every player.

- A **payoff function** s_i for a player i is a correspondence between a strategy profile of all players and a payoff, obtained by player i.

- In this context, a **game** G can be defined as a pair $\{S, s\}$, where s is the vector function $s = (s_1, \ldots, s_n)$, and n is the number of players

- A **finite game** is a game with finite number of players and finite strategy set
- A **mixed strategy** is an assignment of a probability to each pure strategy, defined by means of a given probability distribution [42].

3.3 Nash Equilibrium: Definition, Examples, Existence and Uniqueness

Nash Equilibrium: Definition and Examples Let $G = (S, s)$ be a game with n players, where S_i is the strategy set for player i, $S = S_1 \times S_2 \times \ldots \times S_n$ is the strategy profile and $s(x) = (s_1(x), \ldots, s_n(x))$ is its payoff function evaluated at $x \in S$. Let x_i be a strategy of player i and be x_{-i} be a strategy profile of all players except for player i. When each player $i \in \{1, \ldots, n\}$ chooses strategy x_i resulting in strategy profile $x = (x_1, \ldots, x_n)$ then player i obtains payoff $s_i(x)$. Note that the payoff depends on the strategy profile chosen, i.e., on the strategy chosen by player i as well as the strategies chosen by all the other players. A strategy profile $x^* = \{x_i^*, x_{-i}^*\} \in S$ is a Nash equilibrium if no unilateral deviation in strategy by any single player is profitable for that player, that is [43]:

$$\forall i, x_i \in S_i : s_i(x_i^*, x_{-i}^*) \geq s_i(x_i, x_{-i}^*) \tag{3.5}$$

By applying the above definition, it's possible to identify that the strategy profile $\{B, B\}$ is actually the only Nash equilibrium for the prisoner's dilemma (see 3.1). Indeed, let's assume prisoner Y - the "column player" - makes the choice to use the strategy **B**: in this case, prisoner X - the "row player" - can maximize it's payoff by means of the strategy B, which leads to a payoff -6, which is instead -7 in case of X choosing B. This means that betraying is a so-called **dominant strategy** for player X under the assumption that player Y is betraying as well. Let's consider Y opting for the strategy S: in this case, the best choice for X is again B, leading to a payoff of 0, and therefore betraying is a dominant strategy for X for each possible choice of Y. By exchanging X and Y roles, it can be easily found that B is a dominant strategy for Y as well for each possible choice of X. Considering the definition 3.5, the Nash equilibrium can be reviewed as a strategy profile which is a dominant for all players, and therefore we can conclude that $\{B, B\}$ is the Nash equilibrium for the prisoner's dilemma. It's remarkable that, even if both X and Y could have a better choice in terms of individual payoff - which is -1 for both prisoners - this can be only obtained with the strategy profile $\{S, S\}$, which is not a Nash equilibrium, as starting from this profile each player can gain a better payoff by changing only his own strategy.

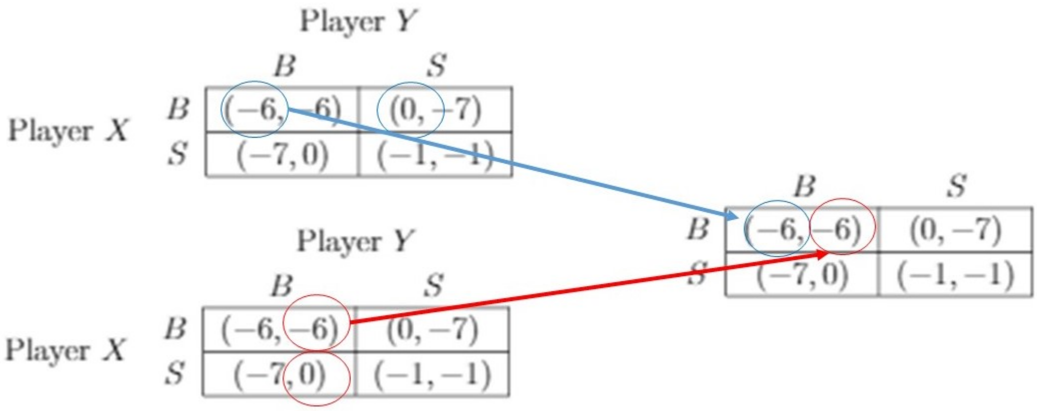


Figure 3.8: Nash Equilibrium for Prisoner's Dilemma as intersection of row player X (blue) and column player Y(red) dominant strategies.

Informally, a Nash equilibrium can be defined [44] as "a solution concept of a non-cooperative game involving two or more players in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only his own strategy."

Stated simply, two players A and B are in Nash equilibrium if A is making the best decision he can, taking into account $B's$ decision while B's decision remains unchanged, and B is making the best decision he can, taking into account $A's$ decision while $A's$ decision remains unchanged. Hence, a group of players are in Nash equilibrium if each one is making the best decision possible, taking into account the decisions of the others in the game as long as the other parties decisions remain unchanged, implying that no player can gain more by unilaterally changing strategy.

Nash Equilibrium: Existence and Uniqueness In his doctoral thesis, John Nash proved that if mixed strategies are allowed, then every finite game has at least one Nash equilibrium, which might be a pure strategy for each player or might be a probability distribution over strategies for each player [45]:

Theorem 1. ***(Existence of Nash Equilibrium)*** *Every finite game has an equilibrium point.*

Regarding uniqueness of Nash equilibrium [46], let's first introduce the following:

Definition 1 (***Diagonally Strictly Concave***)**.** *The function* $\sigma(x, r) := \sum_{i=1}^{n} r_i \phi_i(x), r \in \boldsymbol{R}_{+}^{n}$ *is diagonally strict concave (DSC) if* $(x^1 - x^0)^T g(x^0, r) + (x^1 - x^0)^T g(x^1, r) > 0 \ \forall x^0 \neq x^1 \in C$ *where:* $C \subset \boldsymbol{R}^n$ *is a closed bounded convex set,* $\phi_i(x)$ *is the payoff function of player* i*, continuous in* x *and concave in* x_i*,* $x \in C$*,* $g_i(x, r) := r_i \nabla_i \phi_i(x)$*, and* $(\cdot)^T$ *is the transpose operator.*

The following theorem holds, based on DSC definition:

Theorem 2. ***(Uniqueness of Nash Equilibrium)*** *If* $\exists r > 0$ *s.t.* $\sigma(x, r) := \sum_{i=1}^{n} r_i \phi_i(x)$ *is DSC, there is an unique Nash Equilibrium.*

3.4 Game Theory and Autonomous Machines Decision Making

Game theory and Nash equilibrium can be usefully applied to autonomous machine decision making process. Indeed, as described in [47], given an autonomous vehicle-target assignment problem, where a group of vehicles are expected to optimally assign themselves to a set of targets, a game theoretical formulation of the problem in which the vehicles are viewed as self-interested decision makers can be introduced. Let's considering n_v vehicles or mobile machines assigned to n_i targets, labeled as $V_1, V_2, \dots, V_{n_v}$ and the targets as $T_0, T_1, \dots, T_{n_i}$, where a fictitious target T_0 represents the "null target" or "no target". A vehicle can be assigned to any target in its range, denoted by $A_i \subset T$ for vehicle $V_i \in V$. The assignment of vehicle V_i is denoted by $a_i \in A_i$, and the collection of vehicle assignments $a_1, ..., a_{n_v}$, called the assignment profile, is denoted by a.

Let $V = \{V_1, V_2, \dots, V_{n_v}\}$ and $T = \{T_0, T_1, \dots, T_{n_i}\}$ and let $A = A_1 \times A_2 \times \ldots \times A_{n_v}$, then the assignment of vehicle V_i is denoted by $a_i \in A_i$, and the collection of vehicle assignments $\{a_1, \dots, a_n\}$, called the assignment profile, is denoted by a. Each assignment profile, $a \in A$, corresponds to a global utility function $U(a)$, that can be interpreted as the objective of a global planner. The vehicles can be viewed as autonomous decision makers, and, accordingly, each vehicle V_i is assumed to select its own target assignment $a_i \in A_i$, to maximize its own utility function, $U_i(a)$. Hence, the vehicles are facing a multiplayer game, and as said in previous section a well-known equilibrium concept for multiplayer games is the notion of Nash equilibrium. In the context of an autonomous target assignment problem, a Nash equilibrium is an assignment profile $a^* = (a_1^*, a_2^*, \dots, a_n^*)$ such that no vehicle could improve its utility U_i by unilaterally deviating from a^*. In Chapter 4 the identification of suitable utility functions to be embedded in a game-theoretical decision making system will be discussed, with the main goal of allowing autonomous machines to reach the assigned targets, while ensuring at the same time their safe behaviour towards surrounding human beings.

Chapter 4

Game Theoretical Approach to Safe Decision Making System Development for Autonomous Machines: Mathematical Modelling

In this Chapter a possible translation of machine ethics in algorithmic form will be introduced, with main reference to non-maleficence principle and safety ethical feature.

4.1 Machine Safety Risk Index

As discussed in previous Chapters, in order to translate machine ethical principles into machine-readable code programs, some guidelines can be introduced. Let's consider the specific case of non-maleficence principle and safety ethical feature in order to introduce a breakdown of practical steps to be executed to reach this goal.

1. Select the ethical principle
2. For each selected ethical principle, select the ethical feature
3. For each selected ethical feature, introduce a risk index, expressed in quantitative form.

Therefore, as we are selecting the non-maleficence principle and its related safety ethical feature, the so-called **Machine Safety Risk Index (MSRI)** must be introduced.

Let's consider the specific case of a totally autonomous robotic system (TARS) (see 3.1, 3.4), which can travel in a defined region A of the euclidean two dimensional space with a certain velocity $v_r(t)$.

Let $P_r(t) \in A$ be TARS position as a function of the time, $vmax_r$ its maximum velocity and T its assigned position target. This means that TARS, starting from an initial position $P_{r0} = P_r(t = 0)$, is required to reach the final position T at a certain time t^*, ideally with $v_r(t^*) = 0$.

Let's assume that, in the same region $A \subset \mathbf{R}^2$, there are N humans $H_i,\ i = \{1, ..., N\}$, free to move with velocities $v_i(t)$, and let $P_i(t) \in A$ and $vmax_i$ be their respective positions and max velocities. According to the methodologies provided by ISO standards for machinery safety (see 2.1), all possible hazards must be identified for each of the human beings H_i, and the related risk must be then evaluated. In this scenario, a collision between TARS and any of the humans constitutes the main hazard which can occur: that's why an autonomous mobile robot is normally equipped with a safety rated personnel detection field (e.g. based on laser technology), able to stop it over a short time in case a human is detected on its path. Let's assume that TARS is equipped with the above mentioned safety field, so that there is the need to evaluate the residual risk of collision, as a function of TARS and humans positions and velocities. This is particularly needed in case of large number of humans moving randomly in A, as due to the possible sudden change of velocity and direction, and to the dynamics of the safety detection field, it could be not always possible for the autonomous mobile robot to avoid any incidents. Furthermore, a sudden stop also leads to a reduction of robot productivity, and therefore it can be reasonable to identify other strategies, e.g. a speed gradual reduction based on the position and velocities of humans as a function of the time. With this goal, let's identify a possible risk index of the collision between TARS and the generic i^{th} human H_i. As $P_r(t)$ and $P_i(t)$ are the respective positions of TARS and H_i at the time t, the safety risk of a potential collision can be reviewed for instance as a decreasing function ϕ^s of their euclidean distance $d_i(t) = \|P_r(t) - P_i(t)\|$; e.g, in case of N humans, the overall risk index could be defined as the sum of all the N safety risks $\sum_{i=1}^{N} \phi^s(d_i(t)) = \sum_{i=1}^{N} \phi^s(\|P_r(t) - P_i(t))\|$.

In general, ϕ^s can be a function of P_r and the NP_i human positions, and therefore:

$$MSRI := \phi^s(P_r(t), P_1(t), ..., P_n(t)) \tag{4.1}$$

As an important remark, MSRI is not defined as dependent on TARS and H_i velocities $v_r(t)$ and $v_i(t)$, because any possible interaction is considered as potentially dangerous, regardless the actual value of such velocities.

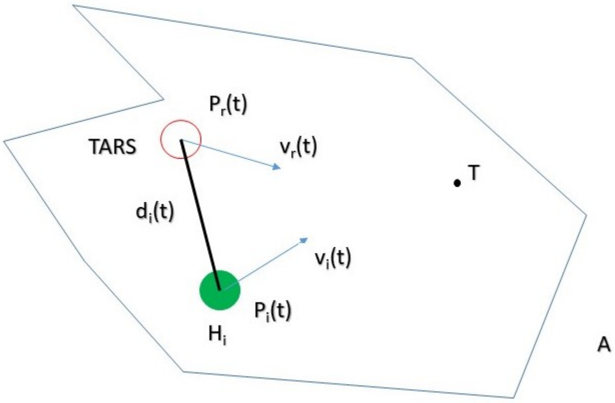


Figure 4.1: TARS and H_i: position, velocity and euclidean distance in A

4.2 Game Theoretical Safe Decision Making: Mathematical Formulation

As described in Section 4.1, TARS is provided with a specific task to be executed, that is to reach a final position T starting from its initial position $P_r(t = 0)$. Let's assume that TARS decision making system can make the decision to change its position to execute its task every Δt, and let $t_k = t_{k-1} + \Delta t$, $k = 1, ..., M \in \mathbf{N}$, the k^{th} instant of time, with $t_0 = 0$.

At each time t_k, TARS can move from its current position $P_r(t_k)$ in a finite number of directions, with a finite set of possible velocities. Let's define the values of TARS velocity scalar components v_{rx} and v_{ry} as follows:

$$\begin{cases} v_{rx}(l) = -vmax_r + 2(l-1)\dfrac{vmax_r}{Ns}, l = 1, .., N_s + 1 \\ v_{ry}(m) = -vmax_r + 2(m-1)\dfrac{vmax_r}{Ns}, m = 1, .., N_s + 1 \end{cases} \tag{4.2}$$

with N_s even integer higher or equal than 2. Therefore, at each time t_k, there will be $(N_s + 1)^2$ possible strategies for TARS decision making system, corresponding to the selection of a given position displacement with scalar components:

$$\begin{cases} \Delta P_r(l, m)_x = v_{rx}(l)\Delta t \\ \Delta P_r(l, m)_y = v_{ry}(m)\Delta t \end{cases} \tag{4.3}$$

Let $\theta_{rj}, j = 1, ..., (N_s + 1)^2$ denote these possible strategies, and let's define a one-to-one correspondence between the values of j with index pairs (l, m), $l, m = 1, .., N_s + 1$, as follows:

$$\begin{cases} j = 1 \leftrightarrow l = 1, m = 1 \\ j = 2 \leftrightarrow l = 1, m = 2 \\ ... \\ j = N_s + 1 \leftrightarrow l = 1, m = N_s + 1 \\ ... \\ j = (N_s + 1)^2 \leftrightarrow l = (N_s + 1), m = (N_s + 1) \end{cases} \tag{4.4}$$

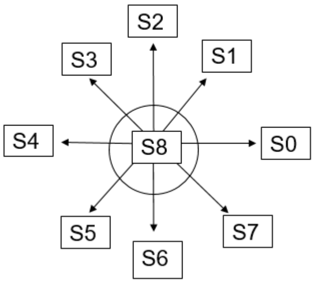


Figure 4.2: TARS strategy set for $N_s = 2$: robot can move in 8 different directions (S0 to S7) or keep standing still in the current position (S8)

At each time t_k there will be at least one strategy $\theta_{rj^*}(t_k) = \theta_r^*(t_k)$, such that $\|P_r^* - T\| = \|(P_r(t_k) + \Delta P_r(l^*, m^*)) - T\| = \min_{\theta_{rj}} \|(P_r(t_k) + \Delta P_r(l, m)) - T\|$, where (l^*, m^*) is the pair of l and m indexes corresponding to j^* as per 4.4. In other words, by selecting the strategy $\theta_r^*(t_k)$, TARS decision making system will minimize the distance with respect to its final target T.

Therefore, for each time t_k it's possible to define a task performance index (TPI) for TARS as follows:

$$TPI := \phi^p(l, m, t_k) = \phi^p(j, t_k) = -\|(P_r(t_k) + \Delta P_r(l, m)) - P_r^*\| \tag{4.5}$$

reaching its maximum value 0 when TARS decision making system selects at the time t_k the strategy $\theta_r^*(t_k)$ corresponding to (l^*, m^*) or, equivalently, to j^*. Considering both 4.1 and 4.5, in case of single human located in A, an overall payoff function for TARS can be defined as follows:

$$\begin{aligned} \phi(l,m,t_k) = TPI + MSRI = \phi^p(l,m,t_k) + \phi^s(l,m,t_k) = \\ = -\|(P_r(t_k) + \Delta P_r(l,m)) - P_r^*\| + \phi^s(P_r(t), P_1(t)) \end{aligned} \tag{4.6}$$

which takes into account both MSRI and TPI. Thus, a game G can be defined, consisting, from TARS perspective, in reaching the target position T taking into account the presence of the human H_1 in such a way to reduce the risk of a collision. For this game G the complete strategy set S_1 and the payoff function s_1 for TARS are completely specified by 4.4 and 4.6. In a game theoretical approach, at each time t_k TARS decision making system must select the best strategy taking into account its own payoff function (with both performance and safety contributions) and the behaviour of the other player, and therefore it must have available a model of the behaviour of the human H_1 in terms of strategy set and payoff function. This H_1 model can be defined by using some assumptions:

1. H_1 has an assigned task as well, which consists in moving from an initial position $P_h(0)$ to a final target position T_h - same as TARS, in general with different initial and final position.

2. H_1 strategy set S_2 has similar structure as TARS (see 4.4), with maximum velocity on both x and y axes $vmax_h < vmax_r$

3. H_1 payoff function s_2 must be defined to adequately model the human behaviour, taking into account the specific application cases and scenarios.

Here below the complete game definition which will be implemented for the simulation (see Chapter 5):

Player1 (TARS)

Strategy profile: $\theta_{rj}, j = 1, ..., (N_s + 1)^2$ (see 4.4)

Payoff function: $s_1 = \|(p_r(t_k) + \Delta p_r(j)) - T\| + K * \langle p_r(t_k) + \Delta p_r(j), p_h(t_k) + \Delta p_h(j')\rangle = TPI + MSRI$ (see 4.6)

with: $j, j' = 1, ..., (N_s + 1)^2, K \in \mathbf{R}^+$

p_r, p_h: TARS and H_1 position vectors

$\Delta p_r(j), \Delta p_h(j')$: TARS and H_1 position vector increment corresponding to the strategy with index j for TARS and to the strategy with index j' for H_1 (see Section 5.3 for further details on K)

Player2 (H_1 model)

Strategy profile: $\theta_{rj'}, j = 1, ..., (N_s + 1)^2$

Payoff function: $s_2 = \langle p_r(t_k) + \Delta p_r(j), p_h(t_k) + \Delta p_h(j')\rangle - \frac{1}{2} * \|\Delta p_h(j')\|^2$

with: $j, j' = 1, ..., (N_s + 1)^2$

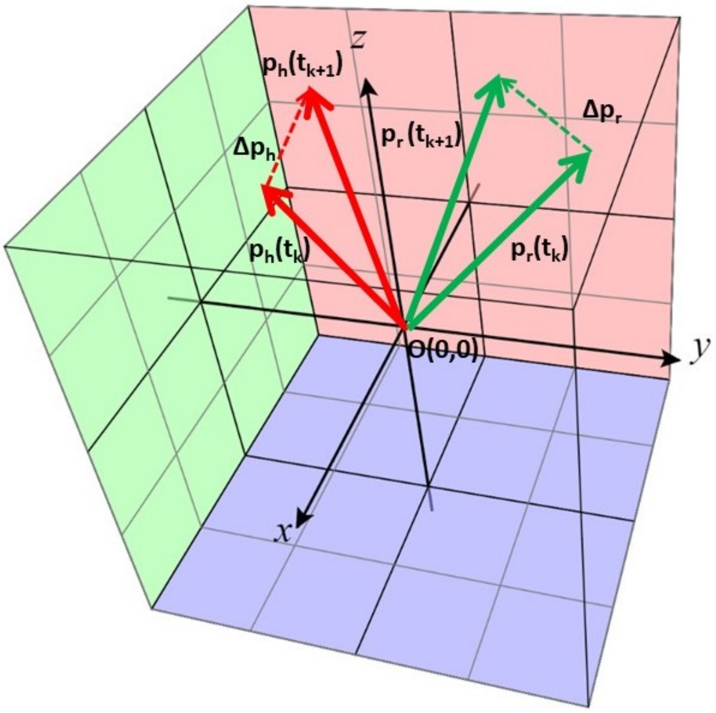


Figure 4.3: Position vectors for TARS (green) and human H_i (red) in 3D space at time t_k and t_{k+1}, depending on the selected strategies at time t_k (green and red dotted line)

Therefore, TARS decision making system will adopt a model of human behaviour aimed to optimize its payoff by selecting a strategy which minimizes the kinetic energy and maximizes the scalar product between TARS and H_1 positions, i.e. considering the human as pursuing a minimum effort action and the worst case of human not attentive to TARS trajectory.

4.3 Application Case Study: Definition, Constraints and Goals

As mentioned in 3.3, Nash equilibrium can be defined as a solution concept of a non-cooperative game involving two or more players in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only his own strategy. As mentioned above, in order to allow this type of formulation, a model of the behaviour of the human being must be used and supposed to be known by machine control system (see 4.2). Hence, at each time t_k, a Nash equilibrium search can be performed in a game G between the machine (player 1) with its strategy profile and the model of the human being (player 2), embedded in the machine control algorithm. This means that, while at each time t_k, TARS decision-making system will apply the strategy resulting from the Nash Equilibrium search as setpoint for the action to be taken at time t_{k+1}, human will keep making decisions independently. Therefore, game G is actually played by TARS only, which will act as first player and, as mentioned above, will use a model of H_1 as second player.

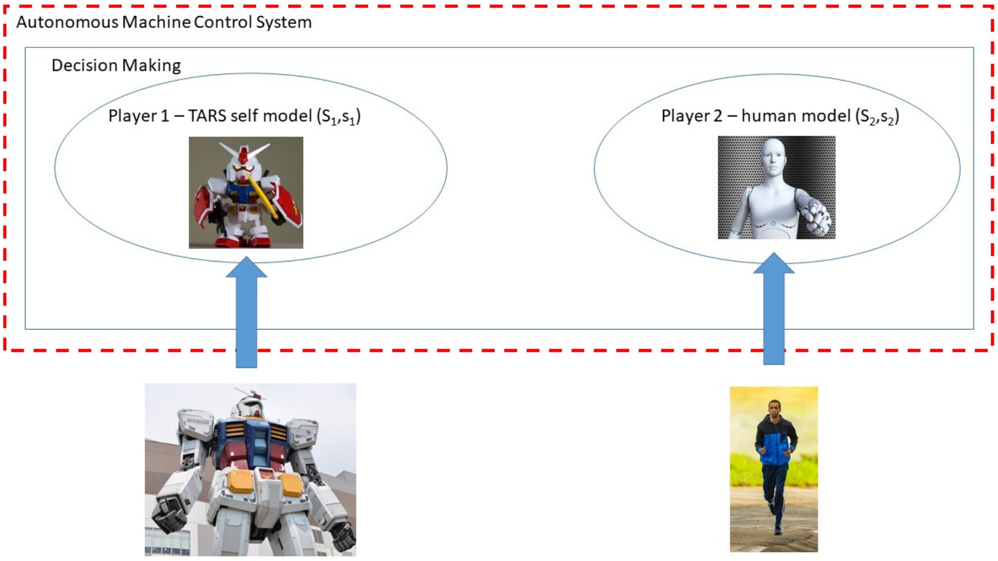

Autonomous Machine Control System
Decision Making
Player 1 – TARS self model (S_1,s_1)
Player 2 – human model (S_2,s_2)

Figure 4.4: Schematic representation of TARS decision making system while playing the game G between itself (strategy set S_1 and payoff function s_1) and human H_1 model (strategy set S_2 and payoff function s_2).

Therefore, the game theoretical formulation will be:

$$\max_{\theta_{rj},\theta_{rj'}} s_I \tag{4.7}$$

where:

$I = 1$ (TARS), 2 (H_1 model)

s_1 is the payoff function for TARS (see 4.6)

s_2 is the payoff function for H_1 model (see Chapter 5)

$\theta_{rj}, \theta_{rj'}$ are the strategy set for TARS and H_1 model.

Based on this framework, the simulation of an application case study has been setup in Matlab© R2020b: TARS and a single human being can move in a region of the euclidean two-dimensional space, more precisely:

- Robot and human trajectories must stay within a circumference C_e with radius R_e
- Robot should not access a forbidden region delimited by a circle C_i with radius $R_i < R_e$
- Both robot and human have assigned with a specific task, e.g. to reach a final position starting from the initial position, standing the above mentioned constraints.

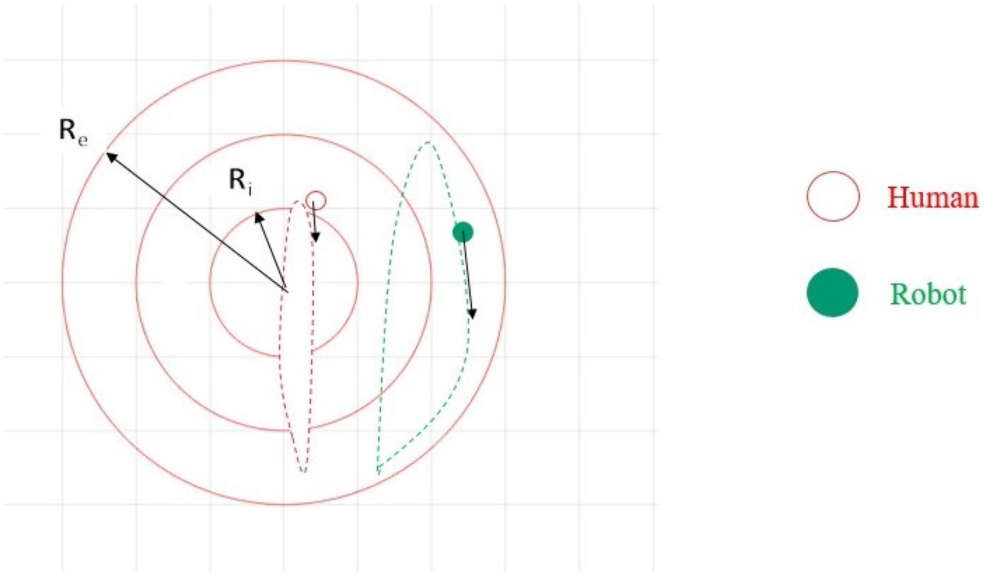


Figure 4.5: TARS and H_1 trajectories in the 2-D Euclidean space subset with constraints

Regarding simulation, the following algorithms have been developed:

- A trajectory planner for the robot
- A trajectory planner for the the human

- A Nash Equilibrium finder
- A game theoretical decision maker

Another important aspect is related to the trajectory constraints - e.g. TARS to move within the circular crown with major radius R_e and minor radius R_i - which will be implemented in the task planners of TARS and H_1. In a similar way, instead of including as a non-linear constraint the minimum distance to be kept between TARS and H_1, that will be included in the TARS payoff function, as defined in 4.1. This will allow to avoid the introduction of algebrical constraints in the game theoretical formulation, which could lead to the existence of multiple Nash Equilibria [48]. Main goal of the simulation is to show the improvement in terms of reduction of potential incidents number, in case of adoption of the game theoretical decision making algorithm developed in this Ph.D. research activity, with respect to traditional optimization algorithms (e.g. exhaustive maximum search). In order to make effective this comparison, a Montecarlo-type simulation will be used, by testing the algorithms on a large, defined set of possible TARS and H_1 trajectories in the 2-D euclidean space. The results of this simulation will be described in details in the Chapter 5.

Chapter 5

Algorithm Design and Simulation Results

In this Chapter the algorithms developed in Matlab© R2020b will be described in details, by specifying their implementation in terms of input, outputs and logic. Furthermore, application case study will be introduced, along with the results of the executed simulation

5.1 Algorithm Implementation

Nash_Equilibrium_Finder

This algorithm takes as inputs the payoff matrices of row and column player and provides as output the position of the Nash Equilibrium in the game bi-matrix ($S1|S2$) (row and column indexes)

Input			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
S1	Player1 Payoff Matrix	$(N_s + 1) \times (N_s + 1)$ matrix (float) [adimensional]	N_s: number of Player1 strategies
S2	Player2 Payoff Matrix	$(N_s + 1) \times (N_s + 1)$ matrix (float) [adimensional]	N_s: number of Player2 strategies

Table 5.1: Nash_Equilibrium_Finder: Input description

Output			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
Nash_indexes	Nash equilibrium row and column indexes	pair of integers between 1 and $N_s + 1$ [adimensional]	row index identifies the robot strategy corresponding to Nash Equilibrium

Table 5.2: Nash_Equilibrium_Finder: Output description

Human_Trajectory_Planner This algorithm takes as inputs the time step of the discretized time vector, the current and the target position T_h of the human player, and identifies which of the possible $N_s + 1$ strategies $\theta_{rj'}$ allows to reach an updated position with the minimum distance with respect to T_h, then provides as output the position corresponding to the identified strategy.

Input			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
dt	time step	float [sec]	dt=$t_{k+1} - t_k$
Xk	Human player Current position	2-D position vector [m]	$P_h(t_k)$ (see 4.2)
Xk+1	Human player updated position	2-D position vector [m]	$P_h(t_{k+1})$ (see 4.2)

Table 5.3: Human_Trajectory_Planner: Input description

Output			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
Xf	Human player target position	2-D position vector [m]	T_h (see 4.2)

Table 5.4: Human_Trajectory_Planner: Output description

TARS_Trajectory_Planner This algorithm takes as inputs the time step of the discretized time vector, the current and the target position T of the robot player (TARS), and identifies which of the possible $N_s + 1$ strategies θ_{rj} allows to reach an updated position with the minimum distance with respect to T which also satisfies the geometrical constraints (that is, TARS within the circular crown with radii R_i, R_e, then provides as output the position corresponding to the identified strategy.

Input			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
dt	time step	float [sec]	$dt = t_{k+1} - t_k$
Xk	Robot player Current position	2-D position vector [m]	$P_r(t_k)$ (see 4.2)
Xk+1	Robot player updated position	2-D position vector [m]	$P_r(t_{k+1})$ (see 4.2)

Table 5.5: TARS_Trajectory_Planner: Input description

Output			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
Xf	Robot player target position	2-D position vector [m]	T (see 4.2)

Table 5.6: TARS_Trajectory_Planner: Output description

TARS_GT_Decision_Maker This algorithm takes as inputs the time duration of the simulated experiment, the time step, TARS and H_1 initial and target positions and max velocities, the number of possible strategies $M = (Ns + 1)^2$ for both players, and, by calling the other algorithms previously defined (Nash_Equilibrium_Finder, TARS_Trajectory_Planner, Human_Trajectory_Planner) identifies at each time frame t_k the TARS position at the next time frame $P_r(t_{k+1})$, as a result of a Nash Equilibrium search for the game G defined as follows (see Section 4.2):

Player1 (TARS)

Strategy profile: $\theta_{rj}, j = 1, ..., (N_s + 1)^2$ (see 4.4)

Payoff function: $s_1 = \|(p_r(t_k) + \Delta p_r(j)) - T\| + K * \langle p_r(t_k) + \Delta p_r(j), p_h(t_k) + \Delta p_h(j')\rangle = TPI + MSRI$ (see 4.6)

with: $j, j' = 1, ..., (N_s + 1)^2, K \in \mathbf{R}^+$

p_r, p_h: TARS and H_1 position vectors

$\Delta p_r(j), \Delta p_h(j')$: TARS and H_1 position vector increment corresponding to the strategy with index j for TARS and to the strategy with index j' for H_1 (see Section 5.3 for further details on K)

Player2 (H_1 model)

Strategy profile: $\theta_{rj'}, j = 1, ..., (N_s + 1)^2$

Payoff function: $s_2 = \langle p_r(t_k) + \Delta p_r(j), p_h(t_k) + \Delta p_h(j')\rangle - \frac{1}{2} * \|\Delta p_h(j')\|^2$

with: $j, j' = 1, ..., (N_s + 1)^2$

This algorithm provides as outputs TARS and H_1 complete trajectories during the simulated experiment, and two binary flags: flag_collision, indicating if there has been a collision between TARS and H_1, and flag_hit, indicating if TARS has been able to reach its target position T.

Input			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
Td	Time duration of simulated experiment	float [sec]	Maximum value of discretized time vector $t = t_0, t_0 + dt...., T$
dt	time step	float [sec]	$dt = t_{k+1} - t_k$
M	Number of total strategies for TARS and H_1	integer [adimensional]	$M = (N_s + 1)^2$
Pr0	Robot player initial position	2-D position vector [m]	$P_r(t_k = 0)$ (see 4.2)
PrT	Robot player target/final position	2-D position vector [m]	T (see 4.2)
Ph0	Human player initial position	2-D position vector [m]	$P_h(t_k = 0)$ (see 4.2)
PhT	Human player target/final position	2-D position vector [m]	T_h (see 4.2)
vmaxr	Robot player maximum scalar velocity on x and y axes	2-D velocity vector [m/sec]	$vmax_r$ (see 4.2)
vmaxh	Human player maximum scalar velocity on x and y axes	2-D velocity vector [m/sec]	$vmax_h$ (see 4.2)

Table 5.7: TARS_GT_Decision_Maker: Input description

Output			
Variable Name	Description	Type and Unit of Measurement	Notes/Comments
xr	Robot player trajectory	TARS trajectory in 2-D space as as an array of float position vectors [m]	$\left(\frac{T}{dt}+1\right)\times 2$ matrix
xh	Human player trajectory	H_1 trajectory in 2-D space as as an array of float position vectors [m]	$\left(\frac{T}{dt}+1\right)\times 2$ matrix
flag_collision	Binary flag equal to 1 when a collision between TARS and H_1 occurs during the simulated experiment	binary (0 or 1) [adimensional]	This flag can be used to identify the number of collisions in a Monte Carlo simulation
flag_hit	Binary flag equal to 1 when TARS reaches T within the simulated experiment duration	binary (0 or 1) [adimensional]	This flag can be used to identify the number of TARS hits in a Monte Carlo simulation

Table 5.8: TARS_GT_Decision_Maker: Output description

5.2 Simulation Results

In order to execute the simulation, a plane region with different concentric circles has been considered. Robot position on the plane is represented by the green point, human operator position by the red circle. The robot and the human must stay within the circumference with maximum radius R_e, robot should not enter within the circumference with minimum radius R_i. As an important remark, in this simulation the human position at each time frame t_k is assumed to be known by the robot (e.g. by using an indoor GPS detection system). Based on the algorithm described in the previous section, a Monte Carlo simulation has been run with the following parameters:

- Niteration (number of simulated experiments): 100
- vxmax (max robot scalar velocity on x and y axes) = 3 m/sec
- vhmax (max human scalar velocity on x and y axes) = 1 m/sec
- Td (duration of each simulated experiment/single run) = 10 sec
- dt (time step) = 1 sec (Note: as this parameter must include the time required for sensor acquisition, on-board algorithm execution, setpoint generation and action execution, it will be always assumed to be higher or equal than 500 ms = 0.5 sec)
- M (number of possible strategy for both TARS and H_1) =9
- K (coefficient of MSRI term, see player1 payoff function) = 1.5.

In order to make more effective the evaluation of the performance of the game theoretical decision making algorithm, for each single run TARS and H_1 initial and final points have been randomly generated by forcing $T' = P_r(0), T = P_h(0)$, (target position of human has been set equal to TARS initial position and viceversa) in order to maximize the probability of collision.

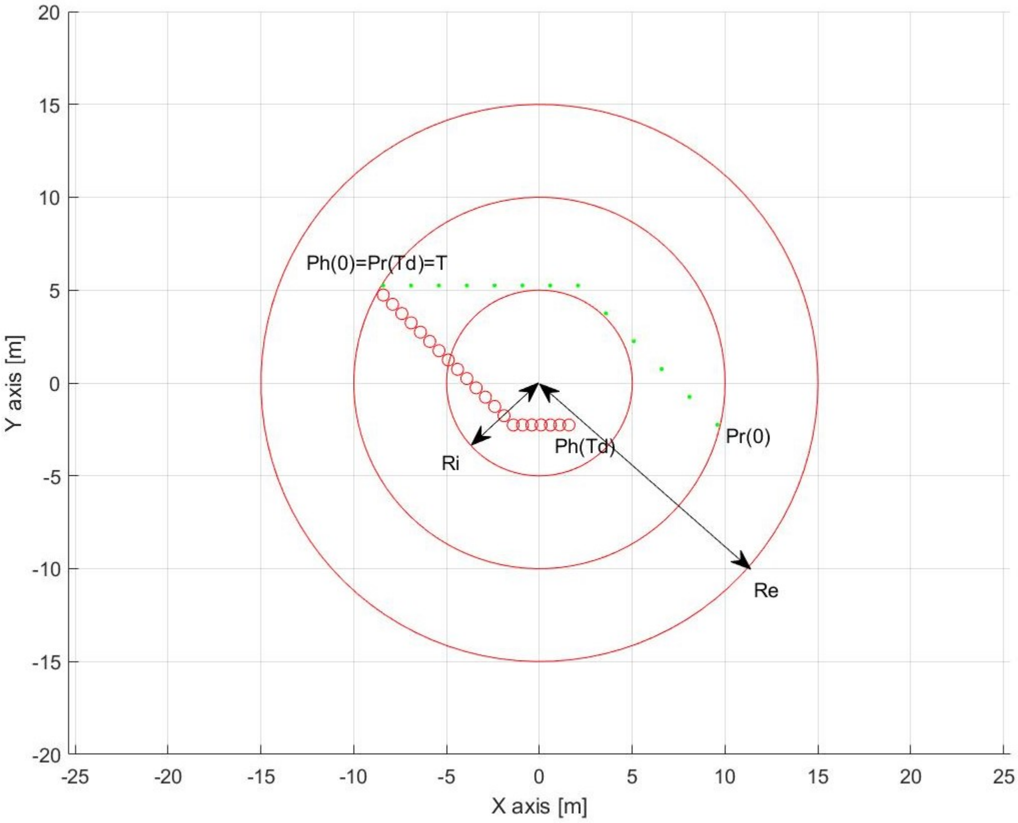


Figure 5.1: Robot (green dot) and human (red circle) trajectories in the 2-D plane (single run, Td=10 sec, dt=0.5 sec, vmaxr=3 m/sec, vmaxh=1 m/sec))

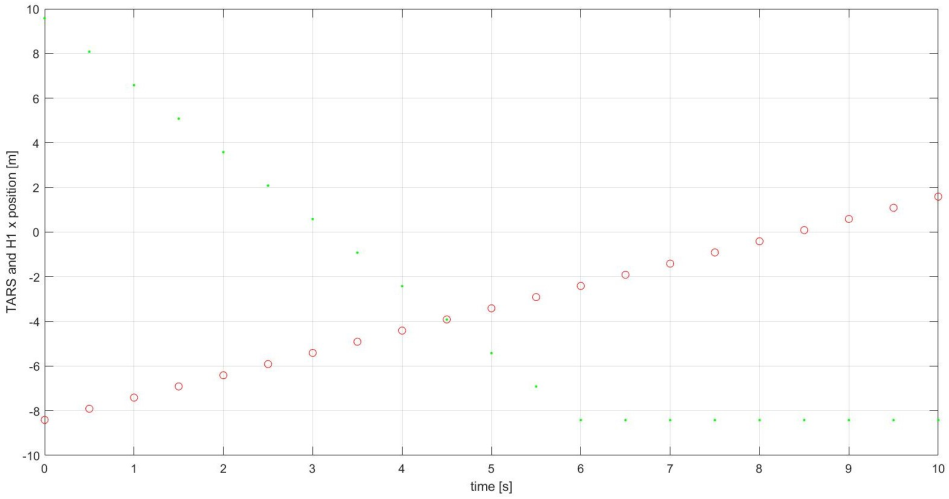


Figure 5.2: Robot (green dot) and human (red circle) x axis position vs single run simulation time (single run, Td=10 sec, dt=0.5 sec, vmaxr=3 m/sec, vmaxh=1 m/sec))

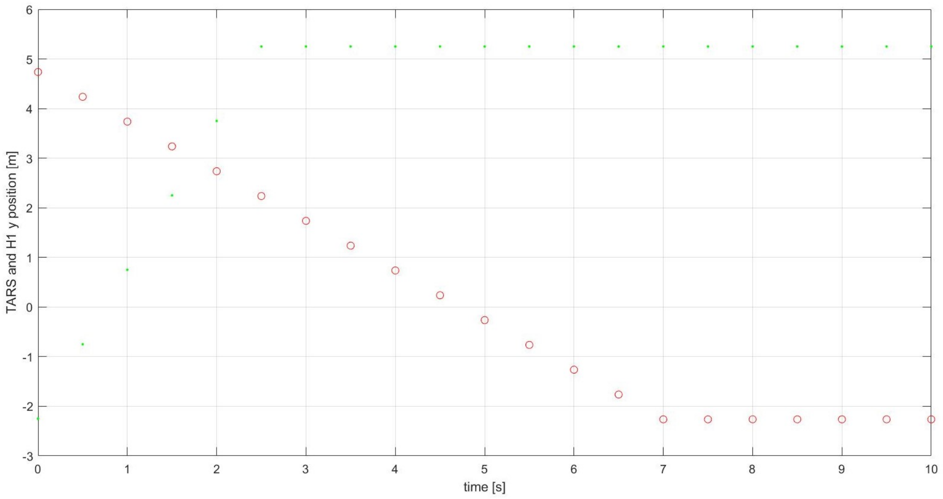


Figure 5.3: Robot (green dot) and human (red circle) x axis position vs single run simulation time (single run, Td=10 sec, dt=0.5 sec, vmaxr=3 m/sec, vmaxh=1 m/sec))

At each single run, two different algorithms have been executed:

1. Exhaustive search optimization (ESO): it only makes use of the s_1 payoff function, by identifying the strategy leading to its maximum value calculating s_1 for all possible strategies. In this case, as s_1 includes the MSRI term, it is expected a better performance of this algorithm with respect to the pure built-in collision avoidance/personnel detection feature available to the robot (see 4.1)

2. The algorithm TARS_GT_Decision_Maker developed in this Ph.D. research work.

The goal of the simulation is to compare the traditional exhaustive search algorithm with the game theoretical optimization algorithm, based on the

Nash Equilibria search for the game between TARS and H_1 model, in terms of number of avoided collision (i.e., number of times each algorithm leads TARS to avoid a collision with the human during a single run). Every time in a single run TARS_GT_Decision_Maker is able to avoid collision, while ESO is not, the run is considered as a win for the game theoretical decision maker, whereas opposite circumstance is considered as a loss. If both algorithms are able to avoid the collision, or both lead to a collision in the same single run, a draw occurs. In the tables below some main results of the simulation have been reported for different values of vxmax, dt, M, and K.

Results table 1	
Parameter	**Value**
Niteration	100
M	9
vxmax	3 m/sec
Td	10 sec
dt	0.5 sec
Percentage of wins	11%
Percentage of losses	0%
Percentage of draws	89%

Table 5.9: Simulation results - first set

Results table 2	
Parameter	**Value**
Niteration	100
M	25
vxmax	3 m/sec
Td	10 sec
dt	1 sec
Percentage of wins	12%
Percentage of losses	3%
Percentage of draws	85%

Table 5.10: Simulation results - second set

Results table 3	
Parameter	**Value**
Niteration	100
M	49
vxmax	3 m/sec
Td	10 sec
dt	1 sec
Percentage of wins	13%
Percentage of losses	2%
Percentage of draws	85%

Table 5.11: Simulation results - third set

Summarizing: for each set of parameter values, TARS_GT_Decision_Maker algorithm allow to obtain a higher number of wins with respect to exhaustive search. By increasing the number of possible strategies M (or, equivalently, Ns s.t. M=Ns+1) the percentage of TARS_GT_Decision_Maker wins increases.

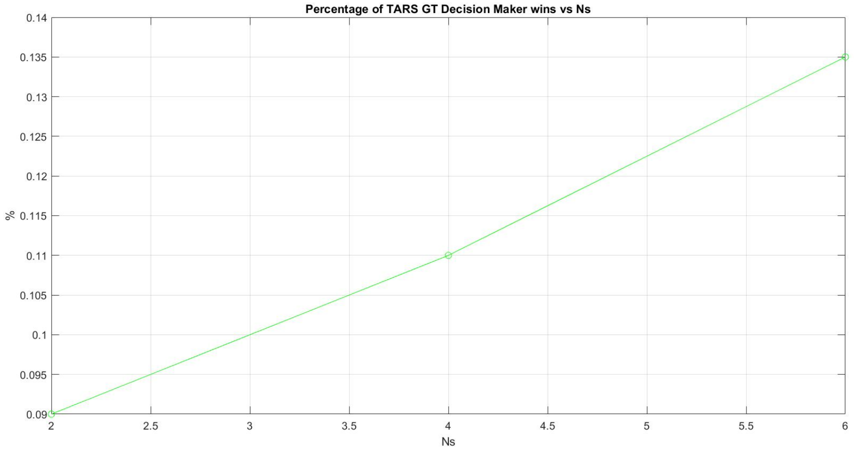


Figure 5.4: Percentage of TARS_GT_Decision_Maker wins vs Ns

5.3 The Ethical Knob

The parameter K introduced in the equation 4.6 plays a crucial role, as it determines the weight of the MSRI contribution to the payoff function, which is used in both ESO and game theoretical decision maker algorithms. In other words, K is acting as an "ethical knob", as defined in [26] (see Section 2.2). By setting $K = 0$, the payoff function s_1 will only take into account the TPI term, and therefore the safe behaviour of TARS will be only determined by the on-board personnel detection field (see Section 4.1. By increasing K the safe behaviour of the optimization algorithms Therefore, in terms of collision avoidance is expected to increase. This trend has been confirmed by means of a dedicated set of Monte Carlo Simulations, each one executed by setting $Niteration = 100, Td = 10sec, dt = 1sec, vmax_r = 3m/sec, M = 49$, and using different values for $K \in [0, 2.5]$: The results of this simulation are summarized by the following figure:

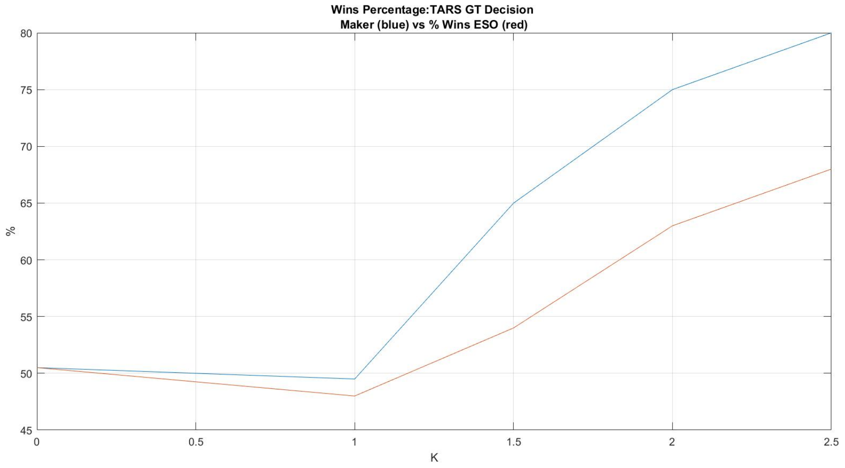


Figure 5.5: Simulation results - percentage of avoided collisions by means of the game theoretical algorithm as a function of K

Chapter 6

Conclusion

In this Ph.D. thesis a novel approach to safe decision making system development for autonomous machines has been introduced and discussed, in order to identify a methodology allowing to improve the safety of the human beings exposed to the autonomous task or operation, by reducing the probability of incidents or dangerous events such as collisions during machine travel. This approach has been based on the definition of some key ethical principles and related features, with main reference to safety aspects, translated in mathematical models in order to define some machine-readable procedures and algorithms. These principles concern the need to avoid or minimize the occurrence of harm for humanity, during the execution of the task the machine has been designed for. Within this scope, four fundamental problems can be introduced:

1. First Problem: Machine Ethics Principles or Laws Identification

2. Second Problem: Incorporating Ethics in the Machine

3. Third Problem: Human-Machine Interaction Degree Definition
4. Fourth Problem: Machine Misdirection Avoidance.

This Ph.D. research activity has been mainly focused on First and Second Problems, with specific reference to safety aspects. With this purpose, a game theoretical formulation of the human-machine interaction has been proposed, by using some main concepts as game, strategy sets and payoff functions as a mathematical model of the interaction dynamics. Machine payoff function has been defined in order to include two different contribution, expressed in terms of a task performance index (TPI) and a machine safety risk index (MSRI). Based on these concepts and mathematical models, the interaction process has been reviewed as a game, consisting from the machine perspective in reaching a given position in a plane region, ensuring at the same time to reduce the risk of collision with surrounding human beings. Therefore, using Matlab© R2020b, a decision making algorithm has been implemented , capable to select the most suitable strategy as the solution of a Nash Equilibrium search problem. The implemented procedure has been compared with a exhaustive search optimization algorithm in a defined scenario, by executing a Monte Carlo simulation obtained by randomly generating impact trajectories between the robot and the human and by changing relevant parameters such as total number of possible strategies, time step, autonomous maximum machine velocity, in order to identify the percentage of collisions avoided as key overall performance indicator, Main results can be summarized as follows: for each set of parameters value, the developed game theoretical decision making algorithm has allowed to obtain a higher percentage of wins (e.g. collisions avoided) with respect to exhaustive search. By increasing the number ofpossible strategies M the percentage of game theoretical algorithm wins increases. Furthermore, another comparison has been performed by means of the same simulation environment, focused on the parameter K introduced in the equation 4.6, which determines the weight of the MSRI contribution to the payoff function, acting as an "ethical knob", as defined in [26]. By increasing K the safe behaviour of the optimization algorithms is expected to improve, and, therefore, their collision avoidance capability is expected to increase. This trend has been confirmed by means of a dedicated set of Monte Carlo Simulations, each one executed by setting $Niteration = 100, Td = 10sec, dt = 1sec, vmax_r = 3m/sec, M = 7$, and using different values for K (see Figure 5.5).

The research work described in this Ph.D. thesis can be extended in a number of different directions:

- Regarding ethical principles and risk indexes, as this work is focused on non-maleficence principle, with specific reference to safety, e.g. just one of the possible ethical features introduced in Chapter 2 (see Figure 2.1), other principles and related features could be identified, in such a way to broaden the scope of the mathematical modelling to further aspects. As mentioned in the Chapter 2, for instance ISO 12100 provides a methodology to calculate the machine safety risk index, as a function of: the severity and The probability of occurrence of the harm to the people issued by a machine. This formulation allows to model and quantify the safety in 4.6. Similarly, other indexes could be defined, e.g. to evaluate privacy violation risks in cybersecurity systems [49].

- Regarding definition of payoff functions, both s_1 and s_2 could be modified by introducing other definition of safety (or privacy, politeness, dignity and so on) risk index, and my modeling in a different way the human behaviour.

- Furthermore, recent breakthroughs in machine learning and deep learning [50] could inspire novel approaches by integrating game theoretical concepts and methods. Indeed, as machine learning deals with algorithms that can learn from the data, by creating models of the environment or physical phenomena, using such models to make predictions and take decisions, most of the issues could be translated to optimization problems with conflicting objectives, which is the scope of game theory itself [51].

Aknowlegments

During the writing of this Ph.D. thesis I have received a great assistance and help from many persons. I would first like to thank my supervisors, Professors Michele Ciarletta and Vincenzo Tibullo, whose expertise and guidance has been invaluable in supporting my research work. I would also like to acknowledge my colleagues and my manager, always showing the greatest understanding and comprehension for my committment on this Ph.D. activity, and providing precious contributions in high level and informal discussions as well as in technical deep dive on the matter of this research. I also thank my friends, providing support and friendship that I needed. Lastly, but not least, I would like to thank my father, my mother and my girlfriend for their suggestions, moral support and for all the insights provided throughout these three years of intense research and challenging - but always highly emotional and exciting - enterprise.

Bibliography

[1] M. Anderson and S. L. Anderson, “Machine Ethics: Creating an Ethical Intelligent Agent”, AIMag, vol. 28, no. 4, p. 15, Dec. 2007.

[2] L. Muehlhauser, and L. Helm, "The Singularity and Machine Ethics", In: Eden A., Moor J., Søraker J., Steinhart E. (eds) Singularity Hypotheses. The Frontiers Collection. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32560-1_6, 2012.

[3] I. Asimov, "I, Robot", Greenwich, Conn: Fawcett Publications, 1950.

[4] K. Shahriari and M. Shahriari, "IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems" 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, 2017, pp. 197-201, doi:10.1109/IHTC.2017.8058187.

[5] A. Rizzo, "Ethically Aligned Design", Version 2, IEEE, December 2017,

[6] L. Floridi and T. Clement-Jones, "The five principles key to any ethical framework for AI" New Statesman https://tech.newstatesman.com/policy/ai-ethics-framework, 2019.

[7] H. Liu, M. Maas, J. Danaher, L. Scarcella, M. Lexer, and L. Van Rompaey, "Artificial Intelligence and Legal Disruption: a New Model for Analysis, Law, Innovation and Technology", 12:2, 205-258, DOI: 10.1080/17579961.2020.1815402, 2020.

[8] Dictionary, o., 2021. Machine Meaning In The Cambridge English Dictionary. [online] Dictionary.cambridge.org. Available at: <http://dictionary.cambridge.org/dictionary/english/machine> [Accessed 17 January 2021].

[9] European Commission, Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery and amending Directive 95/16/EC (recast) (Text with EEA relevance) OJ L 157, 9.6.2006, p. 24–86, 2006.

[10] S. Cave, R. Nyrup, K. Vold, and A. Weller, "Motivations and Risks of Machine Ethics", Proceedings of the IEEE. PP. 1-13. 10.1109/JPROC.2018.2865996, 2018.

[11] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines", Minds & Machines 30, 99–120, 2020.

[12] V. Dignum, "Responsible Artificial Intelligence: Designing AI for Human Values", 2017.

[13] V. Dignum, "Responsible Autonomy", IJCAI, 2017.

[14] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence", The Cambridge handbook of artificial intelligence, 1, 316-334, 2014.

[15] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society", Harvard Data Science Review, 1(1), 2019.

[16] The Future of Life Institute, "Asilomar AI Principles" `https://futureoflife.org/ai-principles/` accessed 1 December 2017

[17] European Commission's European Group on Ethics in Science and New Technologies, "Statement on artificial intelligence, robotics and 'autonomous' systems", Brussels, doi:`10.2777/786515`, 2018.

[18] UK House of Lords, "Artificial Intelligence Committee's report", 2018.

[19] J. Kivistö-Rahnastoi, "Machine Safety Design: An Approach Fulfilling European Safety Requirements", VTT Publications, 2000.

[20] N. Wiener, "Some Moral and Technical Consequences of Automation." Science, vol. 131, no. 3410, pp. 1355–1358. JSTOR,`www.jstor.org/stable/1705998`. 1960.

[21] V. Dignum, "Responsible Artificial Intelligence", "Artificial Intelligence: Foundations, Theory, and Algorithms", `doi:10.1007/978-3-030-30371-6`, 2019.

[22] International Organization for Standardization, "ISO 26262-1:2011 - Road vehicles — Functional safety — Part 1: Vocabulary", 2011.

[23] International Organization for Standardization, "ISO 13849-1:2015 - Safety of machinery — Safety-related parts of control systems — Part 1: General principles for design", 2015.

[24] L. Dennis, and M. Fisher, "Practical Challenges in Explicit Ethical Machine Reasoning", 2018

[25] U. Furbach, C. Schon, and F. Stolzenburg, "Automated Reasoning for Robot Ethics", 2015.

[26] G. Contissa, F. Lagioia, and G. Sartor, "The Ethical Knob: ethically-customisable automated vehicles and the law", 2017.

[27] S. Morimoto, F. Wang, R. Zhang and J. Zhu, "Cybersecurity in Autonomous Vehicles", 10.13140/RG.2.2.31503.23207, 2017.

[28] Z. Rappaport, "Robotics and artificial intelligence: Jewish ethical perspectives", Acta neurochirurgica. Supplement. 98. 9-12. 10.1007/978-3-211-33303-7_2, 2006.

[29] T. Sorell, "Asimov's Laws of Robotics aren't the moral guidelines they appear to be", 2017.

[30] R. R. Murphy and D.D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics", IEEE Intelligent Systems, 24, 2009.

[31] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety", 2016.

[32] International Organization for Standardization, "ISO 12100:2010 - Safety of machinery — General principles for design — Risk assessment and risk reduction", 2015.

[33] R. Myerson, "Game Theory: Analysis of Conflict", Cambridge, Massachusetts; London, England: Harvard University Press, Retrieved December 27, 2020, from `http://www.jstor.org/stable/j.ctvjsf522`, 1991.

[34] J. Von Neumann, "On the Theory of Parlor Games" (in German), Mathematische Annalen, Vol. 100, pp. 295–320, 1928.

[35] M. Dresher, "Theory and Applications of Games of Strategy", Santa Monica, Calif.: RAND Corporation, R-216, 1951. As of January 27, 2021: `https://www.rand.org/pubs/reports/R216.html`

[36] B. Von Stengel and T. Turocy, "Game Theory. Encyclopedia of Information Systems", 2. 10.1016/B0-12-227240-4/00076-9, 2003.

[37] M. Czubenko, Z. Kowalczuk, and A. Ordys, "Autonomous Driver Based on an Intelligent System of Decision-Making", Cogn Comput 7, 569–581, `https://doi.org/10.1007/s12559-015-9320-5`, 2015.

[38] A. Vinogradov, A. Terentev, V. Petrov, and O. Petrov, "Development of Mathematical Model of Moving Wheeled Robot Using Visual Programming Platform Labview," 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, 2017, pp. 1056-1059, `doi:10.1109/EIConRus.2017.7910738`.

[39] S. Wandile,"A Review of Game Theory", 2013.

[40] I. Aguirre, "Notes on Non-Cooperative Game Theory-Microeconomic Theory IV", 2009.

[41] N. Milovsky, "The Basics of Game Theory and Associated Games", 2014.

[42] MIT OpenCourseWare. 6.254: "Theory with Engineering Applications", Lecture 6: Continuous and Discontinuous Games, Spring 2010.

[43] D. Fudenberg, and J. Tirole, "Game Theory", The MIT Press, Cambridge, MA, 1 edition, 1991.

[44] M. J. Osborne and A. Rubinstein, "A Course in Game Theory", Cambridge, MA: MIT, 1994.

[45] J. Nash, "Non-Cooperative Games", Annals of Mathematics, 54(2), second series, 286-295. `doi:10.2307/1969529`, 1951.

[46] J. Walrand, "Concave Games, Learning in Games, Cooperative Games", EE228a - Lecture 20, 2006, Retrieved from `https://people.eecs.berkeley.edu/~wlr/228S06/L20.pdf`

[47] G. Arslan, J. R. Marden, and J. S. Shamma, "Autonomous Vehicle-target Assignment: A Game-theoretical Formulation", Journal of Dynamic Systems, Measurement, and Control 129 (5), 584-596, 2007.

[48] R Spica, D. Falanga, E. Cristofalo, E. Montijano, D. Scaramuzza, and M. Schwager, “A Real-Time Game Theoretic Planner for Autonomous Two-Player Drone Racing”, <i>arXiv e-prints</i>, 2018.

[49] S. Mascetti, N. Metoui, A, Lanzi, and C. Bettini, "EPIC: A Methodology for Evaluating Privacy Violation Risk in Cybersecurity Systems", Transactions on Data Privacy. 11. 239-277, 2018.

[50] J. Schrittwieser, I. Antonoglou, T. Hubert, et al. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model", Nature 588, 604–609, 2020. https://doi.org/10.1038/s41586-020-03051-4

[51] A. Agrawal, and D. Jaiswal. "When Machine Learning Meets AI and Game Theory.", 2012.