UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI STUDI UMANISTICI

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH

DUBLIN

TECHNOLOGICAL
UNIVERSITY DUBLIN

Corso di Laurea Magistrale in Scienze Filosofiche

EXTENTION OF MORAL AGENCY TO ARTIFICIAL AGENTS

Relatore:
Chiar.mo Prof. Carmine DI MARTINO
Relatore Esterno:
Chiar.mo Prof. Brendan TIERNEY
Correlatrice:
Chiar.ma Prof. Rossella FABBRICHESI

Tesi di Laurea:
Sofia QUAGLIA
Matr. 940482

Anno Accademico 2020 – 2021

Table of Contents

Introduction

History books tell of a war that no longer exists. They tell of a call to arms to defend empires and crusades in distant lands, of horses and other animals, of ambitious inventions and discoveries of war including mechanical instruments of destruction, characteristic of the First and Second World Wars. Although the battlefields become wider and the distance between armies increases, the common denominator of each battle is constant: the human being. With it emerges a shared goal of losing as few lives as possible and in turn maximising enemy losses; and war, as it is today, is the realisation of this goal. Today, battles are fought in offices, where soldiers manipulate *consoles* and screens, guiding drones that operate thousands of kilometres away. Operators can take a break, have a chat with a co-worker, and when the mission is over, they go home to their families for dinner. A new form of the art of war, a combat never seen before.[1]

There are many projects that aim to put humans '*out of the loop*' with massive financial backing.[2] Countless ethical questions take shape with the exclusion of humans from the decision-making process and the possibility of monitoring the Artificial Intelligence (A.I.) is brought into question. An autonomous robot decides in a fraction of a second whether to use its destructive potential or not, and the result can be the death of many, soldiers, and civilians alike. Such a scenario is realistic, since it is impossible for even the most advanced A.I. to distinguish enemy infantry from non-combatants, and this inability completely violates the fundamental ethical precepts of a just war according to the *jus in bello,* regulated by the Geneva and Hague Conventions, as well as other international protocols.[3]

Is it possible to build an ethical A.I.? If it is possible, how can one be sure that it does not violate human rights? Who, or what, should be held responsible when these rights are infringed? Are the laws and protocols in place appropriate for this new type

[1] Cfr. N. Sharkey, *Killing Made Easy: From Joysticks to Politics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 111-115.

[2] Cfr. N. Sharkey, *Cassandra or the false prophet of doom: AI robots and war,* in *IEEE Intelligent Systems* 23 (4) (July-August), 2008, pp. 14-17.

[3] Cfr. N. Sharkey, *Killing Made Easy: From Joysticks to Politics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 118-123.

of combat? Can A.I. be held responsible?[4] These questions are central to the goals of *Robot Ethics*, and although the most innovative products are present on the battlefield, their use is not limited to this. Significant research and investment promote projects aimed at taking the '*man' out* of manual, think of autonomous machines and *bots* that manage the stock market. It is legitimate and indispensable to ask all these questions, not only in life-or-death scenarios, but also in the more ordinary and habitual situations where A.I. are used without too much publicity.

The subject of this thesis, which is developed in the field of A.I. and Robot Ethics, is *Robot Rights*. The ultimate goal of this research is to answer the questions: (i) What is the function of rights and who are the rights holders?; (ii) What are the requirements to be considered a moral agent? (i.e., to be held accountable); (iii) Could artificial agents be recommended as moral agents? (ontological requirements); and (iv) Would it be convenient to consider them as such? (ethical implications).

In order to respond, it is necessary to break them down into macro-concepts and investigate each one separately so as to reveal its implications. Since it is essential that there is no confusion in the terminology used, none of these concepts will be taken for granted. Especially since, in the field of AI, there is a tendency to use the same word to mean different entities and categories. Once defined, they will be placed in relation to each other, so that the research questions mentioned above can be answered.

The current bibliography was considered in the construction of the project. Specifically studied were David J. Gunkel, who in *Robot Rights* [5]frames the problem of the rights of robots in a broad and in-depth manner, giving an account of the theories present and above all distinguishing the ontological problem from the ethical problem. Joshua C. Gellers who in *Rights for Robots* [6] places the problem of the rights of robots alongside that of the rights of animals and nature, in order to outline the need for a post-human ecological ethic. Finally, Mark Coeckelbergh, who in *Robot rights? Towards a social-relational justification of moral consideration*[7]*,* focuses on the

[4] Cfr. G. A. Bekey, C*urrent Trends in Robotics: Technology and Ethics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, p. 32.

[5] D. J. Gunkel, *Robot Rights,* MITP, Cambridge Massachusetts, 2018.

[6] J. C. Gellers, *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*, Routledge, Abingdon-New York 2021.

[7] M. Coeckelbergh, *Robot rights? Towards a social-relational justification of moral consideration*, in *Ethics, Information and Technology 12*, Springer, 2010, pp. 209-221.

ethical and relational substratum: robots exist, they are present in the everyday life of human beings, does it really make sense to focus on ontological research that does not lead to useful results for society and its real problems?

These authors were fundamental in learning the state of the art and identifying the most urgent problems, as well as in understanding the methods of investigation used by the other authors. Although fundamental, these authors will not be examined in this thesis work because it was decided to start from different assumptions.

In fact, the decision was made to try to understand whether it might be possible to consider some AIs as artificial moral agents, and therefore responsible, following the indications of *standard* and more conservative philosophical theories and comparing them with the existing judicial system.

The research of this thesis work has the advantage of proposing a rigorous but immediate solution: it takes into consideration existing and everyday technologies, standard and contemporary philosophical theories and above all it refers to actual existing legal systems (in particular the Italian legal system).

The research work will begin with the *First Section*, dedicated to discovering the function of rights. It will be necessary to understand whether rights are natural, inherent in human biology, or positive, i.e., based on consensus and relative to the legal and juridical system that produced them. Moving then on to an analysis of the relationship between rights, the set of legal norms that regulate the relations of the subjects of a given society, and moral laws, the rules of right action, trying to identify which moral theory best represents this relationship. For this, deontology and consequentialism will be considered.

An attempt will be made to show how all rights are positive, relative because they are based on consensus, identifying their historical and conceptual foundation in moral laws. It will also be argued that a consequentialist theory best describes the relationship between positive rights and moral laws, since an action is judged right, rather than wrong, on the basis of its consequences: if the effect of an action is positive then it will be good, and vice versa. This would explain perfectly why moral laws exist; they exist to prevent the emergence of dangerous situations for society and its cohesion. Think, for example, of the *Ten Commandments*, how they protect the social fabric from the endless cycle of revenge. At the same time, positive rights exist to make laws more effective, enforceable, and institutional.

The work will then move on to the analysis of the subject of rights: what kind of category is the 'rights holder'? Is it fixed or does it change over time? On the basis of what criteria are rights assigned? Be careful, this question is not posed in an ethical sense, but ontologically, focusing on the necessary requirements. To answer this question, the theories of available rights will be explored, which make explicit the functions of a right.

There are two main theories: *will theory and interest theory.*[8] For the latter, rights must protect the interest of the holder, i.e., if there is interest there is right and every person has a duty to respect the rights of those who have an interest[9], while the former theory states that the function of a right is to give one person control over another person's duty.[10] This theory recognises freedom as the foundation of all rights and consequently having a right means being free to choose, to do or not to do. Both theories introduce the concepts of right and duty and, since both do not clarify the relationship between them, it will be useful to refer to *the* correlativity thesis, which is essential for conceptualising the relationship between right and duty.

The thesis of correlativity may be defined as follows: subject A has a duty to respect the rights of subject B, so every right claimed by B is related to the corresponding duty of A. Duty, by definition, is a moral or legal obligation, a responsibility. If B claims a right and A has a duty to respect it, the question arises as to who or what A is. Starting with an easy answer, it can be said that subject A is a moral agent, a person, capable of moral responsibility; in short, there are no rights without the corresponding responsibilities.

Treating the relationship between rights and duties in this simplistic way, a legitimate question to ask whether it is only morally responsible agents who are subjects of rights, and consequently whether an agent can behave as he wishes in

[8] Cfr. L. Wenar, *The Nature of Rights*, in *Philosophy and Public Affair vol. 33*, Wiley 2005, pp. 223-252.

[9] Cfr. K. Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, in *Robot Ethics, the Ethical and Social Implications of Robotics*, MIT Press, Cambridge Massachusetts 2012, p. 39; J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford Clarendon Press 1907.

[10] Cfr. L. Wenar, *The Nature of Rights*, in *Philosophy and Public Affair vol. 33*, Wiley 2005, pp. 238-240.

relation to non-agents. The correlativity thesis, now misunderstood, becomes the *converse correlativity thesis* and leads to circumstances in which people with rights mistreat, without disregarding the law, entities that have no rights at all (i.e., Animals, Environment, Robots, etc.).

However, it is when one looks at actual exchanges and practices that one corroborates or falsifies a thesis: the scenarios depicted by *converse correlativity thesis* are not the order of the day, especially in recent times. An agent has many duties that do not correspond to any rights. Take for instance the duty of a society not to pollute the environment: to operate it needs to adhere to specific standards and *guidelines* (i.e., Paris Accords) despite the fact that Nature itself has no rights. Entities such as animals and the environment should be recognised as moral patients. When asking whether an A.I. can or cannot be a moral agent, the question must be broadened to include whether an A.I. can or cannot be a moral patient, given the characteristics it shares with other such entities.

Through the study of the theories of rights, it will be noted that the category of the subject of law and moral agency are open categories whose margins are elastic. On a theoretical level, it would therefore be possible to extend them to A.I. Moreover, why is moral agency so important?

Starting in the *Second Section* this question will be acknowledged. As far as humans are concerned, most ethical paradigms assume that agency is essential to the theory of mind.[11] The theory of mind is the ability to attribute mental states to subjects with whom one interacts. There is, however, a small congenital flaw with this amazing ability: humans apply theory of mind to everything around them. Intentional states are used to explain everything, for example even A.I. It is very common to attribute human emotions to so-called 'socially active robots', despite the fact that there is no trace of emotion in them and the relationship is definitively one-sided. Highlighting the analogies that come from the application of the theory of mind will allow a deeper understanding of the ontological differences, characteristic of the human condition and the 'artificial condition'.

[11] Cfr. G. Verruggio, K. Abney, *Roboethics: The Applied Ethics for a New Science*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, p. 355.

At the end of this investigation, the requirements, and their emergence in the two conditions considered will be reviewed. So that the study can finally move on to which aspects of moral agency are actually required by today's social systems. This thesis will focus on legal systems and it will be interesting to see how this considers only some of the requirements. Indeed, legal systems revolve around the existence of two decision-making systems: the deliberative system and the emotional system. Shifting the focus, from the philosophical narrative of moral agency to its normativity, shows the asymmetrical deviation, the perhaps unnecessary space, between the two.

Therefore, the thesis will continue with an explanation of the deliberative and emotional system, analysing concrete examples where the presence, or absence, of one of the two systems (minor age/mental disability, autistic spectrum disorders and psychosis) shows that only the deliberative system is necessary for the recognition of the moral agency, i.e., responsible before the law. The deliberative system is that system which involves the ability to shape alternative futures in the form of mental representations and to choose what is preferred. Consequently, it will be considered how decisions are made, opening the way to the question "Do the artificial deliberate?", which is essential for taking a position on the existence of artificial moral agency.

The Second Section will conclude by referring to contemporary Western legal systems and the main philosophical theories, the requirements to be considered a moral agent.

Knowing the prerequisites for a human being to be regarded as a moral agent, one can turn to the following part of the research question: Could artificial agents be regarded as moral agents?

In philosophy, there are four main answers to this question.[12] The first considers A.I., as they are known, to be non-agents, but perhaps in the future they will be.[13] For the second, however, they are not and will never become so.[14] There is also an opposing view that A.I. are moral agents, whereas humans cannot be considered as

[12] Cfr. J. P. Sullins, *When Is a Robot a Moral Agent?*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 151-161.

[13] Cfr. D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365.

[14] Cfr. S. Bringsjord, *Ethical Robots: The Future Can Heed Us*, in *AI and Society* 22, 2008, pp. 539-550, (https://doi.org/10.1007/s00146-007-0090-9), 18th November 2020.

such[15] The last response is represented by L. Floridi and J.W. Sanders, of the Information Ethics Group at Oxford University, who argue that in order to go beyond all the paradoxes implicit in the three previous visions, it is necessary to adopt a 'mind-less morality' that avoids controversies such as free will and intentionality, issues that have never been resolved in philosophy. Indeed, it is inappropriate to apply categories to artificial agents that are in any case obscure.[16]

Five requirements will be identified: autonomy, free will, intentionality, conscience, and responsibility.

The work will continue with the *Third Section*, in which the requisites for being a moral agent will be examined, each of which will be observed as it occurs in the human condition and in the "artificial condition". The aim is to construct a symmetrical system that allows a clear analysis of the concepts taken into consideration (requirements) and enables the emergence of ontological differences between these two entities.

This part will be the most technical and interdisciplinary of the thesis. It will include an up-to-date review of the characteristic mechanisms of A.I. decision making and will explore the technical vanguard of quantum computing, arguing that with it comes a new form of *decision making* and the possibility for A.I. to have free will. With regard to the A.I. algorithms to be considered, it is proposed to delimit the analysis to one or more specific application fields in which these algorithms have an impact and a real implementation (i.e., Military Drones, Autonomous Robots, Deep Learning Algorithms etc.).

After this study, conclusions will be drawn. These conclusions will be guided by Professor Brendan Tierney, lecturer in computer science with experience in ethics at the Technical University of Dublin and director of Ethics 4EU, in order to make up for those shortcomings that arise physiologically in the composition of a master's thesis in philosophy.

The goal would be to come up with conclusions with a precise opinion and a clear perspective on "(i) Whether there are A.I. that exhibit autonomy, free will, intentionality, consciousness and responsibility; (ii) If these requirements emerge, how

[15] Cfr. J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248.

[16] Cfr. L. Floridi, *On the Morality of Artificial Agents*, in in *Machine Ethics*, Cambridge University Press, Cambridge, pp. 184-212.

do they emerge, what are the similarities and what are the differences with humans?; and (iv) What would be the consequent implications for the current legal system? With the hope to succeed, the missing part of the research question can be unravelled: If an artificial agent can be a moral agent, ought it to be one?

In order to get to this point, the path has crossed different fields of research. The theories of rights are dwelled upon in order to understand their function, then proceeded to analyse the relationship between rights and duties that introduces the correlativity thesis, with its inevitable alteration in the form of *converse correlativity thesis*. A discussion ensued regarding the concepts of moral responsibility and moral agency and the phrasing of symmetries and differences of the human and 'artificial' condition.

Among the many features that emerged from the discussion, the focus was on decision-making systems and the fundamental requirements (autonomy, free will, intentionality, consciousness, and responsibility) with their theoretical and technical implications. Only now, with all the indispensable tools at hand, the questions "Whether A.I. can be considered as an artificial moral agent” and then “whether or not it should be", can now be addressed.

First section: Rights

1.1 Premise

These first paragraphs may seem precocious and confusing but when the topic of "rights" is discussed there is this innate feeling of dealing with a familiar and well understood concept. Newspapers are constantly reporting debates about recognition, protection, or violation of rights. Every year there are more and more claims for the rights of different people who recognise themselves as part of a discriminated minority. The word "right" is uttered so many times that no one questions anymore, what a right is. Therefore, when asked whether robots can or even should be subjects of the law, it is easy to feel a bit lost: what is an artificial intelligence (A.I.)? Why should it have rights? What exactly are rights?

This work will begin by understanding why laws exist by outlining a brief phenomenology; then it will observe them in their changing form, following the rhythm of societies. It will arrive at the present, at the age of universal declarations and will ask whether these have a different foundation than any other positive legislation. Then it will give an initial, perhaps early, but important answer to continue the discussion.

It is important to clarify the terminology from the outset: on the following pages the concepts of "law" and "right" will be used. The choice between the two is not random, by law it is meant every principle by which the order of human reality is made known and which is placed as the judge of behaviour. Laws can be norms of ethical, social and legal conduct; for example, they are the rules of discernment between good and evil. Right, on the other hand, is the set of juridical rules that regulate the behaviour of those who are the recipients of it. The word right will be used to refer to specific legal norms with specific subjects of law.

1.2 Why the laws

Fratricide is the beginning of every sacred and profane history. Cain is jealous of the relationship that Abel has with God, so he kills him in the ploughed fields. The account of the fratricide of the firstborn sons of Jewish humanity shows how brotherhood has no moral value in itself. A value implies the knowledge of what good and evil are, and it is unthinkable that rules of conduct are created before the actions to which they refer. Who creates laws? What is their *raison d'être*?

The movement of humanity in avoiding its own annihilation is constant and the tactics used are manifold. In the history of philosophy there are many authors who explain this awareness of the danger that comes from the world and the consequent need to protect oneself. Philosophical anthropology is rich in narratives of the human as a being who, thanks to technique, is able to exonerate himself from the pressure of the world and political philosophy tells of terribly dangerous states of nature that are made liveable and civilized thanks to positive law. One of the most recent and captivating analyses is to be found in Peter Sloterdijk, a German philosopher who from 1998 to 2009 committed himself to building a paradigm that explores the constitution of human subjectivity and communities, which are observed in their highest inclusive expansion and modern deflagration.

Peter Sloterdijk, in the trilogy of *Spheres*[17], describes the human being as creator and inhabitant of spaces.

> The reason why the search for ours *where* it makes more sense than ever lies in the fact that it questions the place that men produce in order to have what they may appear to be. This place bears the name of the *sphere*. […] . Spheres are the creation of spaces endowed with an immunosystemic effect for ecstatic creatures on which the outside world works. [18]

Talking about microspheres, or partaking in *Microspherology*[19], means talking about the human as an inhabitant of the interior structured by an original spatiality, the maternal uterus, which will try to repeat always and everywhere with every means. This is the history of uterotechnics: the attempt, always unfinished but vital, to repropose outside the uterus the intrauterine conditions.

The first uterotechnical event is the horde, the oldest social collective and original place of formation of humanity. This is a total inclusive group in which members are bred and supported in order to perpetuate the existence of the horde itself; it is the

[17] P. Sloterdijk, *Sfere I – Bolle, Microsferologia,* Raffaello Cortina Editore, Milano 2014; *Sfere II – Globi, Macrosferologia,* Raffaello Cortina Editore, Milano 2014; *Sfere III – Schiume, Sferologia plurale,* Raffaello Cortina Editore, Milano 2015.

[18] P. Sloterdijk, *Sfere I - Bolle, Microsferologia,* Raffaello Cortina Editore, Milano 2014, cit., p. 54, transl. mine.

[19] "Microspherology" is a term coined by Sloterdijk. It refers to all analyses on the construction of the individual and subjectivity, which are contrasted with the canonical self-sufficient image of the subject.

original form of paleo-politics. All real societies, the primitive ones, like the complex ones, are spheropoietic projects. The small groups of hunters lead a self-rounding and self-inclusive existence, *similar to a greenhouse without walls*. The interior is more important than the exterior and this warm core is protected not by brick walls but by walls of relationships; the description horde could in fact, be redefined as a relational greenhouse. The horde is the expansion of the uterus, a social incubator where the most upsetting *biological experiences of moulding* have taken place.

The prehistoric tribes meet, individuals move, and the spheres burst and recompose themselves, ever expanding. Increasing the outside and the danger, life begins in an enlarged immunological sphere, within which human beings will be able to live only if they maintain the techniques developed in the social incubator.

This integral force consists in transposing the integral space.

> The openness of the outside, of the extraneous, of the fortuitous and absurd, of what makes the spheres burst forth immediately competes with a process of poetics of the world that works to quarter each exterior in a wider interior [...]. Until the outside is eliminated or reduced to tolerable sizes. The order is above all the effect of a transposition of the interior towards the exterior. [20]

How do the inhabitants of the world protect themselves in a greenhouse that is sufficiently expanded but at the same time resistant? If before the protective layer was relational, now to be able to build complex and omnicomprehensive containers, there is a need to know the world and to have a concrete image of it.

To make the uterus visible, walls and fortified imperial cities are built, which are functioning immune systems. Sloterdijk uses "immunity", a medical-biological term, in an anthropogenetic context; the meaning he gives it is almost the same: he means by immunity the condition of defence which is implemented by natural or acquired mechanisms against foreign or dangerous substances. A foreigner infects the interior, the immune reaction is to incorporate it, reorganising it through the transposition mechanism, which is the re-proposal of corroborated *uteromimetic* behavioural practices that resemantise it. The second book of the trilogy is the phenomenology of *uterotechnology*, on the attempts to construct spheres that could replace the loss of the original *microspherical* unit. Religions and metaphysical implants represent immune

[20] *Ivi*, cit., p. 48, transl. mine.

systems designed to protect the psychic balance of human groups: in general, all that is considered “culture” is to be considered an immune system.

In other words, *macrospherology*[21] consists in the analysis of collective agglomerations and their philosophical, historical, and spatial undertakings that have led to the succession of *different images of the world;* images that have constituted the structure of humanity over the centuries. The *theory of the spheres* is a means of constructing the exodus of human outside the primitive symbiosis and towards the action of human history in global empires and systems: it is a great narration of human adventure, on a philosophical, anthropological, religious, and historical-spatial level that goes from the great greyness to the threshold of the contemporary world.

This very brief reference to Sloterdijk served to show a possible analysis of the sudden movement of the humankind in avoiding death. The human finds himself in a hostile world and is committed to making it more hospitable; this effort takes the form of techniques and tools, but also of rules of conduct: the traffic of culture that is created and modified.

> [...] the rules of conduct are aimed at modifying interindividual relations in order to make possible a peaceful coexistence and the very survival of the group. Tools and rules of conduct form the world of "culture" as opposed to that of "nature". [22]

Overwhelmed by the danger of the state of nature, Hobbes' human practices defence techniques: systems of rules that condemn aggressive actions and promote solidarity and collaborative activities. Cain is condemned because, if in a community everyone could kill his own brother without any consequence, it would ultimately be consumed in an endless cycle of revenge.

1.3 The changing form of laws

The rules follow social evolution, they are the product of history. The *Ten Commandments* are imperatives that aim to achieve desired behaviour and eliminate others, the promised sanctions are divine. The *Hammurabi Code and* the *Laws of the*

[21] Macrospherology" means the investigation, contained in the second volume of the trilogy, on the dynamics of transition from microspheres to macrospheres. The macrospheres are God, the World and all the symbolic figures of both political and metaphysical nature, which have given rise to "immune" devices: they can be interpreted as constructions through which human collectives erect defences against the senselessness and exteriority of the world.

[22] N. Bobbio, *L'Età dei diritti*, Einaudi, Turin 2014, cit., p. 49, transl. mine.

XII Tables make up the portions of a moral world that remedies the evil that a person could do. The first laws, the first moral worlds are mostly obligations and duties; the problem of morality is considered from the point of view of society; therefore, the codes of rules have the function of protecting the group. Cain must not kill Abel, not because the latter must be protected as an individual, but to prevent the group from irremediably disintegrating; in fact, "not to kill" has value only for the members of a particular group, while enemies are better off eliminated.

Philosophical reflection on politics, perfect forms of governing, and the role of laws in cities has focused for centuries on the role of government. The legislator was considered a deity, the laws he/she promulgated were sacred and unquestionable, most of them consisted of obligations for subjugation and were fundamentally positive. Positive law is understood to be the law in force in a given time and space, laid down by the legislator in legal regulations. The individual was a passive subject, he/she had a duty to obey the law. Being a citizen meant first of all to respect the laws. The philosophical doctrine that placed the individual at the centre of morality and law is natural law.

The philosophical-legal current is based on the idea that there is a natural law, intrinsically just because it conforms to the nature of humankind and that it is superior to the positive law that is produced by men. The history of natural laws is ancient, already Sophocles in *Antigone,* tackled for the first time in 442 B.C., the antinomy between written laws and "unwritten laws". Antigone decides to give a dignified burial to Polynices, her brother, in contravention to the decree of the legislator Creon; when asked if she was aware that she had disobeyed the laws of the city, Antigone answers:

> Not Jupiter to me launched such a ban,
> nor Justice, which dwells together
> with the Demons of Avernus, so other laws
> were imposed upon men; and thy proclamations
> I did not believe that they had such strength
> to make the laws of the Heavens,
> unwritten, unshakeable, could
> Overwhelming a mortal: why not now?
> Were sanctioned, or yesterday: eternal live
> they; and no one knows the day they were born.
> And violate them and make them right
> to the *Numi*, I couldn't out of fear
> of any superb. [23]

[23] Sophocles, *Antigone*, Trad. Ettore Romagnoli, transl. mine.

Although ancient, this philosophical doctrine was fully developed in the 17th century, and it was John Locke who was the main inspiration for the first legislators of human rights. The state of nature painted by Locke in the *Second Treaty on Government*[24] is a state of perfect freedom, where every human being can regulate his actions and dispose of his private property within the limits of the law of nature, without having to submit to the positive law of any legislator. The Lockian state of nature is totally opposed to the Hobbesian state: it is neither dangerous nor miserable, but free within the laws of nature. The state, according to Locke, is not a great organism, an "artificial human" where everyone represents an organ, but an individualistic state. In it the individual has value in itself, the state is made for the individual.

The only great reason why humans put themselves into society and abandon themselves to the state of nature is the possibility of a state of war; in fact, where there is authority, the continuation of the state of war is excluded and the dispute is decided by that power. Therefore, the only reason for a person to deprive himself of his natural freedom and accept the constraints of civil society, is the agreement with other people to join and unite in a community, safe and peaceful, and above all has the enjoyment of each their own private property. The preservation of their property is the great and main purpose for which people gather in political communities and submit themselves to a government.

The individualistic perspective is the present. Since the end of the Second World War, more and more importance has been given to the recognition of human rights, which take ever different forms, following the evolution of society. Norberto Bobbio defines *specification as the gradual but increasingly accentuated shift towards a further determination of rights holders.* [25]Since the *Universal Declaration of Human Rights* (1948) many other documents have been approved in the last seventy years: *Declaration of the Rights of the Child* (1959), *Declaration on the Elimination of Discrimination against Women* (1967) ...

This concludes this very brief history of the evolution of rights. The aim was to show how rights were historically relative, changed with society and were only valid

[24] J. Locke, *Two treatises of government, Essay concerning the true original, extent and end of Civil Government*, Awnsham Churchill 1689.

[25] N. Bobbio, *L'Età dei diritti*, Einaudi, Turin 2014, cit., p. 56.

because they were accepted. The current political condition has been defined as individualistic and it has not been clarified whether or not the centrality of the subject derived from the philosophical doctrine of natural law. From the mention of the Universal Declaration of Human Rights, which shows the existence of universal values, it can be deduced that, because they are universal, human rights are natural and that it is precisely in human nature that they merge. Refer to the investigation into the foundation of natural rights in the chapter entitled *1.8 Natural rights do not exist*[26] yet now focus on understanding what rights are at the legal level.

1.4 Wesley Hohfeld incidents

The escalation of the rhetoric of rights is out of control; in Western democracies public debates are phrased in legal language. This proliferation is exemplified in the contemporary debate on abortion: who is "pro" abortion appeals to the right of women to control their reproduction, who is against appeals to the right of the foetus to live. Every public debate has to do with rights, and the confusion about the meaning of rights probably comes from their proliferation, from their inconsistent use. To try to give a definition of "right" and describe its form, it will be used Wesley Newcomb Hohfeld's classification.

Fundamental legal conception[27], the work containing Hohfeld's insights, was published after his death in 1919; although Hohfeld was a jurist and analysed legal rights, his work was transposed, modified, and used to explain moral and political rights. The Hohfeldian system is used because, although it has been criticised over time, it is widely accepted and used as a starting point. Hohfeld composes a *template* to examine the four types of fundamental rights, the "Hohfeld incidents"; they are: privileges, claims, powers and immunities[28]. The discussion will follow Leif Wenar's interpretation[29].

[26] *1.8 Natural rights do not exist* pp. 40-43 of this work.

[27] W. N. Hohfeld, *Fundamental legal conceptions as applied in judicial reasoning and other legal essays,* ed. W. W. Cook, Yale University Press 1919.

[28] In the original language in Hohfeld's work they are correspondingly defined as: *Privileges, Claims, Powers and Immunities*.

[29] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, pp. 223-25

Each right can be broken down into four basic elements, Hohfeld incidents. There are two basic forms to assert a right: "A has a right of φ" and "A has a right for which B φ", where "φ" is a verb, and when these basic forms are combined, they form the four incidents.

If a suspect locks a door behind him, a policeman has the right to break it down, he/she has no obligation not to break down the door: "A has a right Y of φ" implies that "A has no obligation Y of φ" (A's right to break down the door implies that A has no obligation to respect the private property in question). The policeman has a privilege, or freedom, which *exempts* him/her from an obligation that should generally be respected. In this case it is a single privilege, but there are also double privileges or *paired privileges*: a person has a double privilege when he/she has to decide whether or not to perform some kind of action, at his/her discretion. Double privileges allow a person to choose whether to do or not to do; A has the right to speak but can decide not to make a sound. So, privileges have the function of exonerating when they are single and they give the possibility of discretion, hence choice, to the subject.

When not only "A has a right of φ", but "A has a right that B φ" you have a claim; A can claim the obligation of B of φ, so an employee has the right to receive the salary and the company has the obligation to pay it. Complaints have the functions of *protection*, *provision,* and *performance.* Privileges and complaints are often combined, as Wenar indicates, in the United States a prisoner has the privilege of not speaking, so he claims that the police do not force him to speak.

The two remaining Hohfeld incidents are powers and immunities, the former give the subject the right to alter his or her privileges and claims, give him or her authority; while the latter give him or her the right to ensure that they are not altered, the subject is protected. Let us observe the Hohfeldian scheme at work; the four incidents combined together in the complex right that everyone has to dispose of their own body. A has the *privilege*, the freedom to move his body and has the right to *complain* against those who touch his body; moreover, A has the *power* to renounce this right or not and is *immune* from those who would renounce his complaint, against those who touch his body.

Wenar believes that *all rights are accidents of Hohfeld*[30] *,* even if the proof of this statement is inductive:

> Our confidence in this inductive step will increase as we successfully explicate more and more rights with the Hohfeldian diagrams, and as we fail to find counterexamples. The reader may want to satisfy himself or herself that confidence in this inductive step is justified and may wish to test the framework with more sample rights. [31]

However, for the moment it is sufficient to know that the structure of most legal rights is a combination of privileges, claims, powers and immunities.

1.5 Theories of rights

All rights are included in the template created by Hohfeld, yet not every combination of incidents corresponds to a right. In *Essay on Bentham*[32] , H. L. A. Hart specifies that the concept of Hohfeldian right is very broad; Hohfeld in fact uses it to refer to "immunity": defining it as the disabling of a certain power, as already said, immunity protects A from the power that B would have to alter A's legal position. The problem arises when the concept of law should be used to refer to an immunity relating to an advantageous change; Hart imagines the scenario that his neighbour does not have the power to exempt him from paying taxes; this immunity does not constitute any legal right for Hart, and indeed in the realm of legal language, it is not defined as such. On the contrary, an individual's immunity is a right when the implementation of the power of some other individual would deprive him of one of his rights or benefits protected by law. Not all combinations of Hohfeldian incidents are rights, only those that have a specific function are rights[33]; Wenar identifies six of them: *exemption*, *discretion*, *authorization, protection*, *provision,* and *performance*[34].

Although Hohfeld's scheme and analysis is very useful to understand what rights are and what kind of assertions can be made, it does not explain who has rights or why. What is the function of rights? In order to answer it is necessary to study the theories of rights, they aim at typifying an agent, with specific characteristics and abilities to which the rights are attributed. The predominant theories of rights dealt with are *Will*

[30] *Ivi*, cit., p. 237.

[31] *Ivi*, cit., p. 236.

[32] H. L. A. Hart, *Essays on Bentham*, Clarendon Press Oxford 1982.

[33] L. Wenar in *Rights. Stanford Encyclopedia of Philosophy*, 2020, pp. 6-7 mentions some of the criteria by which theorists discriminated which of the Hohfeld incidents were rights.

[34] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, p. 244.

Theory and *Interest Theory*. It is interesting to note, just at the beginning of the discussion, how the distance between them is already traced by the criterion adopted to discriminate which assertions are to be recognized as rights and which are not. The first recognizes as rights only those combinations of incidents that offer the subjects of right choices, the second considers that the implementation of the well-being of the subjects is the criterion for settling the dispute. The debate on which of the two theories offers a better explanation of the function and recipients of rights has been ongoing for centuries.

1.5.1 Will theory

The individual for Will Theory, or theory of choice, is the ruler of a small-scale kingdom[35]; he is free and able to exercise his power. In general, supporters of choice theory believe that the individual who holds one right, has the capacity for discretion over the obligation of another, the function of the right is to offer the exercise of this capacity. For example, the owner of a house has a right because he has the power to waive, or not to waive, the obligation that other individuals have not to violate his private property. Thus, the owner, the holder of the right, has the ability to exercise his power in altering the obligations of other individuals. The right, for the theory of choice, is closely linked to freedom and authority, one reads authority to control the actions of others, and it is very easy to see how the class of possible holders is extremely restricted: the subjects of law are beings specifically endowed with the volitional capacity. This is the category that is taken into consideration, no other.

The limits of the theory of choice are very narrow and, as Wenar points out, this limitation is evident first of all in the set of recognised rights. Some of the most important rights are considered because they are structured on the "power to alter a complaint" paradigm, but many others are not: *nobody, for example, has the legal power to renounce the complaint of someone not to be enslaved, or to cancel the complaint of not being tortured or killed*[36]. Yet these are considered as rights, among the most important rights an individual can have.

The theory of choice has been harshly judged for its narrowness and the theorists who support it have responded to these criticisms in two ways. The first possibility is

[35] H. L. A. Hart, *Essays on Bentham*, Oxford Clarendon Press 1982, cit., p. 183.

[36] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, p. 239.

to narrow the scope of the theory (H. L. A. Hart uses this strategy); while the second possibility is to search among the subjects to whom the law applies, someone who is the holder of that specific right. Following the example of the former: no one has the legal power to overturn someone's claim not to be tortured except the government; the officer in charge has this power, this right. The conclusion is clear, as much as it is disturbing, in fact no one, except the elected few, has a right against the possibility of being tortured.

The restrictions of the theory of choice can be deduced not only from the rights considered, but also from the definition of a category that deletes from the legal game anyone who does not have those certain characteristics, which are different depending on the author and historical context. Children and incompetent adults are not considered as subjects of law and the possibility of incorporating entities such as environment and animals as subjects is not even contemplated.

1.5.2 Immanuel Kant and the freedom of the person of reason

Kant understands the human being as a free and rational individual, he is capable of making deliberate decisions respecting a defined standard: the categorical imperative; the latter is a guiding paradigm of practical action. Freedom is not being forced by the choice of another individual, in the individual it is innate and constitutes the first right of birth. Thus, the philosophy of practical action is to be understood as a set of rules governing the behaviour of rational beings; this system of rules includes all human actions, both in a pure and empirical sense. This is precisely why Kantian political philosophy is founded and derived from its philosophy of practical action.

Kant makes explicit the importance of the study of human beings as particular agents in particular contexts: his is the description of a pragmatic government, not some logically correct and theoretically sensible political form, but a lack of any contact with reality. The above practical philosophy, together with the categorical imperative that governs it, are placed by Kant as the foundation for any system built on the deliberative rationality of the individual.

The work that presents most of the themes of Kantian justice theory is the *Metaphysics of Customs*[37]*;* it consists of two parts: The Doctrine of Law and the Doctrine of Virtue. This distinction is due to a very important theoretical separation;

[37] I. Kant, *La Metafisica dei costumi*, Laterza Bari 1970.

in fact, for Kant, political rights and duties are unrelated to morality, they are something different. A right exists only if three conditions are met: first, a right concerns only those actions that have a direct or indirect influence on another person; the second characteristic is that a right must have to do with the decisions of others, not their desires, but actions that imply a choice; lastly, the object of another person's action does not concern the law, what belongs to him is the typical form of deliberate action. If the rights are those of a free individual, of a subject who chooses freely, then they will have a very precise form, tailored to his needs. When B's action interferes with A's freedom of choice, B's behaviour is to be considered wrong, whatever it may be, even if it is benevolent.

The difference between a political right and a virtue lies in their relationship with the freedom of the individual. As has been shown, rights determine the rights and duties that free individuals have towards one another; the aim is to protect the freedom and free choice of each. Rights outline the appropriate behaviour in respect of a given anthropology, while virtues are the personal motivations that push an individual to act in accordance with a right and the consequent duty.

Kant's anthropological description is the starting point of the orthogonal projections that mark the characteristics of the just man, of the being who can be a holder of rights. The boundaries and characteristics of the rational person have a precise foundation: Kant indicates this in the *Metaphysical foundation of customs*[38]. He bases morality on the principle of autonomy in reason; that is why he always speaks of 'rational beings' and not of 'men': rational beings are simply defined by the fact that they participate in reason[39].

> The absolutely good will, the principle of which can only be a categorical imperative, will therefore contain, indeterminate with regard to each object, only the *form of will in* general, and precisely as autonomy; that is to say: the capacity of the maxim of every good will to become universal law is itself the only law that the will of all rational beings imposes on itself, without any motive or interest being laid as a basis for this capacity. [40]

The innate right, freedom is the foundation of every state: power cannot be rooted in the welfare and interest of citizens, a position which is supported by the Interest Theory. Kant maintains that only freedom has the universal character to be the

[38] I. Kant, *Fondazione metafisica dei costumi*, Laterza Bari 1997.

[39] *Ivi*, cit., p. x.

[40] Ivi, cit., p. 125, transl. mine.

foundation of a system of rights; any other term is subjective and iridescent, think of happiness. Therefore, rights are based on the form of free choice.

It has been said that the human being is a rational and autonomous individual, his congenital right is freedom, and every right is built around this paradigm, the possibility of free choice. How then is it possible to justify the existence of the state? A state by definition is that entity that exercises its power over a given territory, in full independence from other entities, and therefore has the power to control the freedom of citizens. The existence of the state would seem to be a contradiction, Robert Paul Wolff in *The Conflict between Authority and Autonomy,* the first part of *In Defense of Anarchism* [41]precisely uses this argument to support anarchy as the only logical solution.

Conversely, Kant believes that the state is not an impediment to freedom, but the means to be free and to dispose of one's own freedom: if a subject A by his actions violates the freedom of a subject B, the state has the power to intervene to protect subject B and his freedom from the behaviour of subject A. The coercion of the state is justified because it is compatible with the highest degree of freedom, the freedom on which the paradigms of rights are built; such coercion does not reduce freedom, on the contrary it creates the necessary conditions to shape a context in which freedom is assured. In the essay *on the common saying: this may be right for the theory, but it does not apply to the practice*[42] Kant writes:

> *Right* is the limitation of the freedom of each person to the condition of his or her agreement with the freedom of all others, as this is possible under universal law; and *public law* is the set of *external laws* that make such an all-inclusive agreement possible. Now, since any limitation of freedom through the will of another is called *coercion*, it follows that the civil constitution is a relationship of *free* men but (without prejudice to their freedom in their relations with others as a whole) is under coercive laws: since reason itself wants it, and precisely pure a *priori* legislating reason, which does not take into account any empirical purpose. [43]

In this essay Kant describes the *a priori* principles on which civil status is based: the first, already mentioned, is the freedom of every member of society as a result of being a rational being. The second is the equality of every member as a subject; it is not substantial, but formal; in fact every member of the state is equal to every other

[41] R. P. Wolff, In *Defense of Anarchism*, Harper and Row 1970.

[42] I. Kant, Sul detto comune: questo può essere giusto in teoria, ma non vale per la prassi [1793], in *Scritti di storia, politica e diritto*, Laterza Bari 1995, transl. mine.

[43] *Ivi*, cit., p. 137, transl. mine.

before the law; and finally the independence of every member of a common body as a citizen. Although the use of the word citizens could appear as a reference to a democratic universe, in which every citizen contributes to the decision-making process, this is not the case. In fact, Kant does not extend the principle of citizenship to everyone: women and children are excluded by nature.

Kant explains the beginning of civil society in a rational beginning, the original contract. The latter is to be considered as an idea of reason, not a historical event, which pushes the sovereign to promulgate laws in the way that the whole society would have posed them if it had been able to choose them and to which it would certainly have consented.

> Here, then, is an original contract the only one on which a civil constitution, i.e., a constitution absolutely according to law, and a common body can be founded between men. [...] It is, on the other hand, a simple idea of reason, which however has undoubted (practical) reality: to oblige every legislator to enact his laws as they could have been born from the united will of an entire people, and to consider every subject, inasmuch as he wants to be a citizen, as if he had given his assent to such a will. [44]

1.5.3 H.L. A. Hart and modern choice theory

According to Hart the law and the legal system do not derive from nature, they are an artifice: the authority of the law is social. The foundation of a legal system can be neither a legal norm nor a presupposed norm but must be a social rule that exists only because it is practised. Society directs its movements according to a certain custom and the foundation of the legal system is to be found in it. One could therefore say that a legal system is based on a rule, but it is important to point out that this rule has the same normative force as custom. Every aspect of the law exists because someone, no matter if group or individual, has prepared it. It has a history and can be changed and is knowable and understandable in all its aspects. The operation of a law is given by its use, application, custom and not only by its more or less sensible justification. As Leslie Green explains in the introduction to *The Concept of Law* [45]:

> The ultimate basis of law is neither a justification nor a presupposition but a social construction that arises from people thinking and doing certain things. Jurisprudence explains what this construction is and how it is build up from more mundane social facts. [46]

[44] Ivi, cit., pp. 145-144, transl. mine.

[45] H.L.A. Hart, *The Concept of Law*, Oxford University Press, 1961 [2012].

[46] *Ivi*, cit., p. xix.

If by custom it is meant the shared practice of a certain behaviour or thought, deriving from education according to certain principles shared by a community, Hart specifies: the community that founds the law with its own behaviour and attitudes is that of public authority, of public officials. Thus, when the law is distant from life in society, which means it is only present in the bureaucracy exercised by technicians, they have specific duties and responsibilities, and the contributions of the normal population quickly becomes passive consent.

> There are two minimum conditions necessary and sufficient for the existence of a legal system. On the one hand those rules of behaviour which are valid according to the system's ultimate criteria of validity must be generally obeyed and, on the other hand, its rules of recognition specifying the criteria of legal validity and its rules of change and adjudication must be effectively accepted as common public standards and of official behaviour by its officials. [47]

Private citizens are required to obey the law, the reason for this obedience is not important; they must trust that their obedience to the rules will allow the health of society. While the second condition must be respected by public officials holding authority; of course, they must also respect the rules of behaviour valid for private citizens.

In a famous passage in *Essays on Bentham*, already partially mentioned previously, Hart identifies the fundamental structure of rights. Since the existence of an obligation is an indisputable requirement for the existence of rights, Hart suggests that each structurally necessary part of a right should be considered as the different "ingredients" which, combined together, constitute the control system that the subject of law has over the obligation-duty-responsibility of another subject.

> The idea is that of one individual being given by the law exclusive control, mor or less extensive, over another person's duty so that in the area of conduct covered by that duty the individual who has the right is a small-scale sovereign to whom the duty is owed. The fullest measure of control comprises three distinguishable elements: (i) the right-older may waive or extinguish the duty or leave it in existence; (ii) after breach or threatened breach of duty he may leave it 'unforced' or may 'enforce' it by suing form compensation or, in certain cases, for an injunction or mandatory order to restrain the continued or further breach of duty; and (iii) he may waive or extinguish the obligation to pay compensation to which the breach gives rise. [48]

In the Hohfeld system, the "control ingredients" correspond to the powers and following the Hohfeldian method, Hart enucleates their basic forms, which are valid

[47] H.L.A. Hart, *The Concept of Law*, Oxford University Press, 1961 [2012], cit., p.116

[48] H. L. A. Hart, *Essays on Bentham*, Oxford Clarendon Press 1982, cit., pp. 283-184.

despite the different forms they take in different legal systems and contexts. The result is a binary system in which every action required is deontologically necessary, while its transfer or renunciation makes the action deontologically unnecessary.

1.5.5 Interest theory

The theory of interest is based precisely where the theory of choice fails: the criterion that defends allows both to enlarge the category of subjects of law and to consider as rights assertions with different functions.

The function of rights is to advocate the interests of those who hold them; in other words, rights are those combinations of incidents that promote the well-being of the subject. It follows, therefore, that the necessary and sufficient requirement to belong to the category of subject in law is to have an interest. The possibility of being a legal entity is potentially open to any entity, provided that it demonstrates that the entity has an interest to protect, a welfare to pursue. Through the theory of interest, it becomes possible to protect the interests of children and incompetents and, as will be shown below, protection can also be extended to animals and the environment.

When rights no longer have the function of organising a normative decision-making context, where the subject's discretion can be exercised, but they fulfil the promotion of a certain subject's interest, then those incidents that the theory of choice could not consider legal rights, in this theoretical context, are. For example, think of the immunity that protects freedom of speech. Moreover, in a theory of interest, there is also room for all those rights that had already been recognised by the opposing theory.

As Wenar points out, the supporters of the theory of interest support the thesis that the function of a right is to promote the interest of the subject in general, as a paradigm to which one must refer in normal circumstances, they do not argue the absurd theory that the advancement of each individual right is in the interest of the holder. Despite this clarification, it is precisely on this point that criticism of the theory of interest is established; in fact, there are many rights which, even in the most normalised and general context, do not have the function of promoting any interest. Think of the right, power, that an official of justice has to dictate the sentence for an accused, this right benefits the members of the community, certainly not the official. Very often the rights of a legal system are intended to protect the welfare of the community as a whole, not individuals, in this would lie the limitations of a theory of interest. [49]

Yet it could be argued, as Jeremy Bentham does, that the pursuit of the welfare of the community is closely linked to the welfare of the individual. For this reason, and in order to be the first modern to support the theory of interest, the discussion will continue with him.

1.5.6 Bentham and the utility principle

Jeremy Bentham in *An Introduction on the Principles of Morals and Legislation*[50] outlines the contours of a human being like a black box, of which nothing can be known except that he is ruled by two sovereigns: pain and pleasure. This is the human being, and this is his natural position. Pain and pleasure indicate what morally would be better to do and the direction that actions will take. This is why what is pleasurable is both right and good, while pain identifies what is wrong and evil. Bentham describes a certain type of human being and from that precise anthropology derives certain moral values.

The principle of utility is the one in which the Benthamian subject is recognised and the one to be considered for the foundation of the legal system. This principle serves to discriminate against any type of action. With it any behaviour can be approved or disapproved of, simply by answering the question: does it promote or diminish the happiness of the person concerned?

The principle must be applied not only to the action of each individual but also to every measure the government takes.

> VI. An action then may be said to be conformable to the principle of utility, [...] when the tendency it has to augment the happiness of the community is grater than any it has to diminish it.
>
> VII. A measure of government (which is but a particular kind of action, performed by a particular person or persons) may be said to be conformable to or dictated by the principle of utility, when in like manner the tendency which it has to augment the happiness of the community is greater than any which it has to diminish it. [51]

This is because the community must be considered as a virtual body, whose members are the individuals. Considering society as a pseudo-organism allows Bentham to conclude that if the community is the sum of the individuals, then the result

[49] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, p. 241

[50] J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford Clarendon Press 1907.

[51] *Ivi*, cit., 13.

of the sum of their interests will be the interest of the community. The second consequence is theoretical: if the hinge of the community is the individual, then it is necessary and perhaps sufficient to understand who the individual is and what his interest is and a function of direct proportionality will indicate the characteristics of the community.

The legislator has two tools, pleasure and pain, and the principle of utility to refer to; Bentham goes on to make clear what the value of the former is. The importance an individual attaches to pleasure and pain depends on their intensity and duration, but also on the certainty or uncertainty of their occurrence and their proximity or distance. Two other circumstances that must be taken into consideration are their fecundity, which is the possibility of experiencing this pain/pleasure many more times and purity, that is the possibility that one of the two sensations is not immediately followed by the other. Now pain and pleasure can be associated with a value. In order to understand whether an action, or a measure, is to be pursued or not, let's continue by adding up the values of all pleasures and compare them with that of pain: if the value of pleasures is higher, the action is positive, just and good; otherwise it should not be sought.

The task of every government is to promote the happiness of the community and therefore of every individual. This aim can be achieved by outlining a legal system that promotes well-being, identified through the principle of utility and the pain-pleasing instruments, through rewards and punishments. In such a system there will be a criminal law to define punishments, the latter must be decided in proportion to the act that is useless to the well-being of the society that has been carried out; in fact, to be happy and to experience pleasures it is necessary to insure oneself against pain.

The rightful subject of Bentham has a well-being to pursue and preserve; the value and meaning of each action is determined by its effects; regardless, nothing is bad or good, but it becomes so depending on the result: does it bring pleasure or suffering? The boundaries of this subject are blurred, they are an open area and it is possible to extend the non-human animal margins. In a footnote Bentham writes:

> The day may come, when the rest of the animal creation may acquire those rights which never could have been withholden from them but by the hand of tyranny. The French have already discovered that the blackness of the skin is no reason why a human being should be abandoned without redress to the caprice of a tormentor (see Lewis XIV's Code Noir). It may come one day to be recognized, that the number of legs, the villosity of the skin, or the termination of the *os sacrum*, are reasons equally insufficient for abandoning a sensitive being to the same date. What else is that should trace the insuperable line? Is it the faculty of reason, or, perhaps, the faculty of discourse? But a full-grown horse or dog is beyond comparison a

> mor rational, as well as a more conversable animal, than an infant of a day, or a week or even a month, old. But suppose the case were otherwise, what would it avail? The question is not, Can they *reason*?, nor Can they *talk*? but, Can they *suffer*? [52]

1.5.6.1 H.L.A Hart's study of Jeremy Bentham's doctrine

This paragraph presents part of the analysis of the *Benthamian* theory proposed by Hart in the seventh chapter *Legal Rights*[53] of the *Essays on Bentham*. In it, Hart outlines a comparison between the Hohfeldian and the Benthamian system and also describes very clearly the principles of the latter's legal conception. In fact, in addition to anticipating Hohfeld's work, Bentham has investigated aspects of the subject of law that have no place in *Foundamental Legal Conceptions*[54].

The notion of legal law is elusive; examples of this are the contrasts between the theory of choice and the theory of interest, but also all the different definitions and observations on rights that dot the history of jurisprudence. Hart notes that neither Plato or Aristotle use names or periphrases that can be considered as an equivalent of the concept of "law", as something distinct from the "right to act" or the "right to do"; there is no trace of this in all Greek philosophy. Even in the detailed Roman Law, Hart continues[55], a concept of legal law has never been achieved: the concept of legal law belongs to the modern world. Bentham was the first to give a clear meaning to the concept of legal law and his insights are extremely modern.

Bentham starts by distinguishing two types of rights, their difference is mainly in the different role they have with obligations, or duties. The first category of rights owes its existence to the absence of a legal obligation and the type of system of laws governing these rights is to be defined as non-coercive or permissive. Permissive laws may include: an active permit, where a certain action previously prohibited is now permitted; an inactive permit, whereby a law states that a certain action may not be taken; and cases where the law is silent. The second, on the other hand, is the result of an obligation imposed by law, a service, and the relevant system of laws is of a coercive nature. Three main categories can be determined from this binary classification: privileges, claims and powers. Although approximate to the Hohfeld

[52] *Ivi*, cit., 245.

[53] H. L. A. Hart, *Essays on Bentham*, Oxford Clarendon Press 1982, cit., pp. 162-193.

[54] Work previously mentioned in chapter *1.4 Wesley Hohfeld's Accidents* pp. 12-14 of this work.

[55] He is referring to famous jurists like Main*, Early Law and Costumes*, 1891; Buckland, *Text Book of Roman Law*, 1950; Villey, *Lecons d'histoire de la philosophie de droit*, 1957.

incidents, the Benthamian system does not include a concept that corresponds to immunity.

1.5.7 Singer and animal rights

In paragraph *1.5.5 Interest theory* it has been said that, for the theory of interest, a right has the function of pleading the interest of the person who holds it and that when a certain entity was shown to have an interest in protecting its own well-being, it could be considered a moral agent, within that theory.

Peter Singer in *Animal Liberation*[56] shows how all animals should be considered subjects of law. The philosopher starts by examining the struggle for women's rights and wonders why they should be recognised. One possible answer would be to believe that women have the right to vote because they are able to think rationally, just like men. Women are able to deliberate and make decisions about the future, but this argument does not hold in the context of non-human animal rights, a cat does not understand what it means to vote, so they cannot hold this right. Perhaps women and men should have the same rights because they are similar, while the abysmal difference between non-human animals means that they cannot be considered equal.

Singer goes on to argue that recognising the difference between an animal and a humankind is very simple, it is a clear fact that they are different, yet this is not a good reason not to apply the principle of equality to non-human animals. It would be like supporting the right to abortion for men, which is meaningless, because only women can remain pregnant. In the same way a cat cannot vote and therefore it is a waste of time discussing the right to vote of felines. [57]

The principle of equality does not imply that every member of the community is treated in exactly the same way; rights and duties are held by different people according to their different needs and abilities. Equality means, first of all, giving equal importance before the law. The principle of equality is already applied and already considered in the construction of a legal system, no human being has the same cognitive abilities, there are different types of intelligence, moral capacity. Imagine the scenario in which two young women are subjected to the Intellectual Quotient test, the result of one is much higher than the other; there is no logical reason to believe that

[56] P. Singer, *Animal Liberation*, Ecco Harper Collins Publishers 1975.

[57] *Ivi*, cit., 2.

this quantitative, or qualitative difference justifies any kind of discrimination in the consideration of the needs and interests of the two women.

> Equality is a moral idea, not an assertion of fact. [...] The principle of the equality of human beings is not a description of an alleged actual equality among human: it is a prescription of how we should trat human beings. [58]

Singer's argument continues with references to Bentham and then becomes explicit in the fundamental thesis: the only necessary and sufficient characteristic to be able to consider an entity as a subject in law, is the capacity to suffer. It is not comparable to the capacities that are usually taken into consideration, i.e., the theory of choice. To define moral agency is necessary to be able to suffer and rejoice. Feeling pain and pleasure is a required capacity for agency since there can be no interest without pain and joy.

It is therefore essential to show that non-human animals are capable of suffering, but also of feeling satisfied, they have an interest and welfare that must be safeguarded. The following chapters of *Animal Liberation* are pages full of crude descriptions that aim to manifest the real state of suffering of animals and the role that humans play in it. The subjects of the survey are mainly the food industry and the culture of animal experimentation. The book ends with a detailed exposition on why vegetarianism or veganism are the consequent ethical choices, to be taken in the first step towards anti-speciesism[59].

1.5.8 Raz and the interests of others

At the end of the general description of the theory of interest, it was pointed out that the limited nature of the theory of interest can be seen in the fact that many rights do not protect the interest of the person who holds them, but the interest of a third party. Raz proposes to balance the relationship between right and interest by reinforcing it with the interest of others.

Joseph Raz, in the first essay *in the* collection *Ethics in the Public Domain[60]* , explores the notion of well-being, and asks how an individual's interest can or cannot be pursued by other individuals. The exploration of the possible obligations that one

[58] *Ivi*, cit. 5.

[59] By anti-speciesism is meant thought, movement, attitude which, in opposition to specism, opposes the conviction, considered prejudicial, that the human species is superior to other animal species and maintains that the human being cannot dispose of the life and freedom of beings belonging to another species.

[60] J. Raz, *Ethics in the Public Domain*, Oxford University Press 1994.

individual has for another arises from the assumption that the well-being of the latter is considered central to the realization of human life. If this assumption is made, a part of morality and its theorizing must investigate these obligations in order to promote and protect people's well-being.

Well-being and personality are the fundamental dimensions for a person's basic understanding. A person's goodness and success depend on his or her character, but also on the quality of his or her well-being. Raz meaning of well-being is:

> The whole-hearted and successful pursuit of valuable activities. [61]

By defining well-being in this way, Raz thinks of life as active and *kicking*. The success that can be achieved in it is relative: the success of a doctor depends on the treatment of patients and qualitative: only activities with a certain value lead to the realisation of well-being.

An obvious consequence of the fact that well-being is made up of a considerable range of successful activities is that nobody can do them for another person. The obligation to protect and promote the right of all is expressed in the shaping and implementation of the standards of living of others; on the other hand, it is unthinkable to claim that everyone is solely responsible for their own well-being. In order to achieve the well-being of each person, it requires that the conditions be put in place for the practice of those activities that will lead to relative success.

To what extent do these obligations extend? There can be no obligation of mutual love. As has been said, these obligations must facilitate the achievement of well-being, not seek to achieve it through an exchange of roles. For example, you can only become a doctor if you go to a medical school and practise your profession in a hospital. When the conditions for the future doctor's success are shaped, he can succeed with less difficulty.

Basic-Capacities Principle is one of two principles that Raz outlines to define obligations to others. It promotes the basic conditions for pursuing objectives and building relationships, which are fundamental for a full and satisfying life. Obviously not all abilities are required.

> We should help everyone to acquire the (nearly) universal capacities, i.e. Those necessary for all or almost all valuable pursuits. This includes the basic physical and mental abilities of controlled movement and, where disability deprives one of them, appropriate substitutes. They

[61] *Ivi*, cit., 6.

> also include the mental abilities to form, pursue, and judge goals and relationships. We should also help people to acquire and retain enough other basic capacities, or the opportunities to acquire them, to enable them to pursue an adequate range or projects and goals. [62]

This does not mean completely eliminating the possibility of failure, but simply putting in place the possibility of success. The aim of the principle is to equip people with the necessary skills to take advantage of opportunities that arise during the course of life. Having finished the description of the Basic-Capacities Principle, Raz wonders whether it fulfils the obligations included in promoting the well-being of others. As well as protecting and promoting people's basic skills, what could be done?

The Basic-Capacities Principle does not describe either qualitatively or quantitatively how one's basic skills should be exercised. There are various activities that are denied to people; it may be for relevant reasons, such as knowledge of programming languages if you want to be a programmer, but also for reasons unrelated to this activity, such as gender or religion. Raz solves this problem by outlining the *Principle of Adequate* Access:

> The principle we should uphold is simply that every person should have access to an adequate range of options to enable him to have a successful life. [...] the Principle od Adequate Access is not independent of but is inseparable from an argument about which valuable options should be available in societies. [63]

Think of marital union, this is guaranteed to heterosexual couples: a man and a woman are able to love each other and build a future together with shared goals. They are responsible for each other and are committed, virtually forever, to living that union. Marriages of love are a widespread practice, especially in Western culture, and it is likely to enrich the lives of those who decide to get married. In some countries, on the other hand, a homosexual couple cannot be married. The struggle for homosexual marriage is the demand for the transformation of the requirements for access to a practice envisaged by the state. It is the call for a change of value: why get married? Who can do it?

However, in conclusion, Wenar points out[64] that even Raz's interpretation of the theory of interest fails to justify the foundation of the law in the interest of its holder. Indeed, what should be the interest of a judge in exercising his right to punish a

[62] *Ivi*, cit., p. 19

[63] *Ivi*, cit., p. 26.

[64] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, cit., p. 242.

defendant? Raz's answer is that promoting a judge's interest in exercising his right protects not only the judge but also the public interest[65]. However, the attempt to base a right on the interest of the community is not enough, so Wenar, quoting Frances Kamm, writes that: if a journalist's right of free expression is in function of the interest of his readers, in other words that the satisfaction of the readers' interest is the reason why the journalist has the right to have his interest protected, then the journalist’s interest is not enough in order to hold the obligation not to interfere with the journalist’s freedom of thought[66].

1.5.9 Alternative Theory: Several Function Theory

As it was seen, the debate between Will Theory and Interest Theory is confusing and complicated. It would be impossible in this work to try to outline the arguments, for and against, for every single introduced author. Although very brief, the analysis is useful to understand the common ground and the function of the arguments put forward; it also makes it possible to continue without implying any concept or reference.

The discussion of the theories of rights concludes by referring again to Wenar's *The Nature of Rights.* In the second part of the essay, he tries to get away from the narrowness of the two main theories:

> The will theory and the interest theory are both inadequate to our understanding of rights, the weakness of each being the strength of the other. The will theory captures rights that give discretion to the rightsholder without conferring benefits but fails to capture rights that confer benefits without giving discretion. The interest theory accepts rights that confer benefits but rejects rights whose holders do not benefit from holding them. [67]

However, please remember that not every combination of Hohfeld incidents is to be considered a right, but only those with specific functions (*exemption*, *discretion*, *authorization, protection*, *provision,* and *performance*). All rights are Hohfeld incidents and have at least one of the six functions, but some incidents have none. So, trying to resolve the limitations of Will Theory and Interest Theory by arguing that any combination of incidents should be considered a right is not a valid option.

[65] J. Raz, *Ethics in the Public Domain*, Oxford University Press 1994, cit., pp. 149-51, 274-75.

[66] F. Kamm, Rights, in *Oxford Handbook of Jurisprudence and Philosophy of Law*, Oxford University Press 2002, pp. 476-513, cit., p. 485; cf. L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, cit., p. 242.

[67] L. Wenar, The Nature of Rights, in *Philosophy and Public Affair vol. 33*, Wiley 2005, cit., p. 243.

Wenar offers the *Several Functions Theory* as a solution: every incident or combination of incidents is a right, but only if it performs one or more of the six functions. The theory is inclusive; not only does it accept all the rights recognised by the two majority theories, but it also goes beyond the latter by admitting rights that the former excluded. It includes rights that do not give A the power of discretion over B's liability and rights that do not promote any interest. The rights have different roles and the rights have all the functions they have [68].

Unfortunately, Wenar does not specify how the rights should be assigned, who are the subjects of the law of the Several Functions Theory? As far as the agency is concerned, it is impossible to argue that this theory resolves the contrasts of Will theory and Interest theory.

1.6 Summaries and projections

Is the moment to summarize the route taken up to now. When the *raison d'être* of the laws was asked, at the very beginning of this work, the position unconsciously assumed was a positivist and consequentialist one.[69] In fact, laws have been defined as shared norms of ethical conduct, moral laws defined by people in society, and have been described as part of a movement that the human being has always repeated in avoiding his own death. It had been mentioned the state of nature of Hobbes and said that positive law allows the exemption from the danger involved in the deadly context, the one where the coercive power is not concentrated in the figure of the legislator and each *man is wolf for the other*. It was also recalled that the existence of values, the foundation of laws, is linked to the need not to propose dangerous situations, which could lead to a social crisis.

Thus, it has been observed, albeit very briefly, the evolution of the laws in history and the narrative required mention of the existence of two positions: positivism and natural law. If in the past the analysis was that of the type of perfect government and considered the subjects of law as passive, with John Locke a new phase begins, in which the individual is placed at the centre. The paragraph ends with a reference to the

[68] *Ivi*, cit., p.248.

[69] Positivism' and 'Consequentialism' will be analysed in depth in paragraphs *1.7 Naturalism and legal positivism* and *1.9 Consequentialism: the good before the just*. In this paragraph we refer to the two terms in a generic way and determinedly to the criteria explained.

present, the era of universal declarations and realised individualism. Although the various sections of the narrative were dictated by the sequentially of the historical past, the conclusion seemed inconsistent with the positivist position claimed until then. If, as is true for the advocates of natural law, natural rights are the foundation of positive law, then the latter is to be considered as an explication, a systematic organisation of those natural rights, in rights, in force, in a given place and time. If positive rights are established on natural law, what role do moral laws play? If the natural law hypothesis were correct, moral laws would also be based on natural rights and their relationship with legal rights would be questionable.

Before finding an answer to this complicated question it had to be understood what legal rights were and what their form was. To do this a scheme was offered by Wesley Hohfeld and immediately afterwards it was necessary to analyse its function; this was the meaning of the long presentation of rights theories. They did not help to clarify any of the elements of the controversial relationship; the exponents of Will Theory and Interest Theory assume a different and inconsistent origin of rights than the theory to which they belong. Think, for example, of Kant and Hart, both "choice theorists": Kant bases his theory of justice on the imperative[70] of the categorical imperative and on the naturalness of freedom; while Hart, starting from Kant himself, outlines a radically positivist theory in which the authority of law is social. Again, Raz is a pupil of Hart's, just as he is a positivist and not only supports the theory of interest but he is committed to trying to make it less limited by building his system not on the interest of the subject of law, but on that of others.

Also, from the point of view of the origin of natural rights the situation is controversial. In fact, for Bentham, to sustain the existence of natural rights is *logically absurd and morally pernicious;* even the simple association of these two terms is a contradiction: there can be no rights without laws. It follows that there can be no morally natural laws and concludes that there can be no natural rights[71]. On the contrary, Kant believes that natural law is the set of indications according to the right reason: natural law defines the content of the rule, while positive law defines the sanction.

[70] Read "imperativism" as the Theory of Law according to which the legal norm, unlike other norms, is an imperative, i.e., a command to do or not to do something.

[71] L. W. Sumner, *The Moral Foundation of Rights*, Clarendon Oxford Press 1987, cit., p 112.

> The intrinsic value of man is based on freedom, on the fact that he possesses his own will. Since he must be the ultimate goal, his will must not depend on anything else. [...] Man's freedom is the condition under which he can be an end in himself. [...] Law is the limitation of freedom, according to which freedom can coexist with that of anyone else according to a universal rule. Someone may like a place, on which another already sits, and wish to move him from there. Now: I can sit where I want, and the other person can sit where he wants; but [where] he sits, I cannot sit at the same time. There must therefore be a universal rule, according to which the freedoms of both can coexist. [...] Freedom must be limited, but through natural laws this is not possible; otherwise man would not be free; he must therefore limit himself. The right is therefore based on the limitation of freedom. [72]

What are natural rights and what is their origin? What is the relationship between natural rights, moral laws and positive rights? If it can be proved that natural and moral rights are positive, it can be argued that the recipient identified by the various authors is also positive. Each author outlines the identity of a subject of different rights: on the one hand there is a subject whose welfare and interest must be protected, on the other hand a free, reasoned, and deliberative subject. Is it possible that even the definition of moral agency is arbitrary and instrumentally cut out according to the type of normativity one wishes to advocate? Since the aim of this work is to understand whether it is possible to consider a robot as a moral agent, but also to ask whether the extension of the agency is an ethical and desirable philosophical move. If the answer is affirmative, it will have to be understood how to behave.

1.7 Naturalism and legal positivism

What is the relationship between natural rights and positive rights? Are natural rights to be considered as analogous to moral values and as the foundation of the positive legal system that is their explanation and execution; or are all rights positive and therefore natural rights are not natural at all? This chapter will have as its object the controversy between naturalism and legal positivism; they will be presented in their historical and philosophical evolution. Norberto Bobbio's insights will be followed, referring to his entire work and in particular to *Juusnaturalism and legal positivism*[73], and will conclude with an argument in support of the just-positivist approach. This analysis will clarify and define the relationship between natural rights

[72] I. Kant, *Lezioni sul Diritto Naturale,* edited by N. Hinske and G. S. Bordoni, Bompiani Editore Milano 2016, cit. pp. 71-72, transl. mine.

[73] N. Bobbio, *Giusnaturalismo e positivismo giuridico*, Laterza, Roma-Bari 2011.

and positive rights and will also prepare the conceptual ground for determining the link they have with moral laws.

By natural law it is meant the current that promotes the superiority of natural law over positive law; the two types of rights are distinct both from an axiological and normative point of view. Bobbio indicates the beginning of modern-day natural law in 1625, coinciding with the work of Hugo Grotius *De Iure Belli Ac Pacis*[74] ; and its end in the early nineteenth century with the publication of Hegel's juvenile essay, *The scientific ways of dealing with natural right*[75].

Natural law is characterised by the use of a theoretical model, of the Hobbesian type, and a geometric, rational method. In Hobbes' work, the civil status replaces and overlaps the state of nature, as in a change of state, caused by the stipulation of the social contract between individuals: if in the state of nature individuals acted according to their own passions and instincts, the pursuit of the purpose of one was a threat to the life of the other; on the contrary, in the civil status one can live according to reason.

> Precisely because the state of nature is a state of perpetual insecurity, men aspire to change it, to move from the state of nature to civil status. In order to establish in the civil state that security that can only make bonds effective by transforming them from internal to external, individuals agree to renounce all the rights they had in the state of nature (with the exception of the right to life) and transfer them to the sovereign. [76]

It is thanks to the rational method that it is possible to reduce law and morality to a demonstrative science; the object of law is studied and analysed with geometric reason. This method is ahistorical and detached from any previous intuition, it pursues an exclusively rational reconstruction of the origin and foundation of the state. [77]

Bobbio identifies multiple forms of natural law including scholastic, modern rationalistic and Hobbesian. For scholastic law-naturalism, an example of which can be found in Tommaso D'Acquino's *Summa Theologica,* positive law is determined by natural law and the legislator must govern according to the latter. Natural law is composed of a set of generic principles and positive law, the human right, specifies them. The universe of school naturalism divides the world between legislators, who

[74] H. Grotius, *The Rights of War and Peace*, Liberty Found, Indianapolis 2005.

[75] G. W. F. Hegel, *Le maniere scientifiche di trattamento il diritto naturale,* edited by C. Sabbatini, Bompiani Editore, Milano 2016, transl. mine.

[76] N. Bobbio, *Locke e il diritto naturale*, Giapichelli, Turin 1963, p. 44.

[77] N. Bobbio, M. Bovero, *Società e stato nella storia della filosofia politica moderna,* Il Saggiatore, Milan 1979, p. 36.

have the task of determining positive law by following the principles of natural law and the subjects who must respect the laws imposed, even if unjust.

Modern rational naturalism can be recognised in Kant:

> Natural law is the set of *dictamina rectae rationis* that provide the subject matter of regulation, while positive law is the set of practical-political expedients (such as the institution and organisation of a coercive power) that determine its form or, in other words, the first constitutes the preceptive part of the rule, the one that attributes the normative qualification to a given behaviour, the second the punitive part, the one that makes the rule effective in a world, such as the human one, dominated by passions that prevent most from following the dictates of reason. [78]

As has already been pointed out, the Kantian analysis describes the transition from provisional to peremptory civil status. The transition to civil status is a necessity of reason, it is a moral duty; because freedom, the first right of every man, is guaranteed through the exercise of the right. The combination of natural law and positive law is congruent with that between provisional and peremptory law, in all similarity and dissimilarity. If natural law is the result of the relationship of coexistence between individuals in the state of nature, positive law does not differ in content, but in the different procedures used to enforce it. For Kant, the right is also addressed to individuals, with their freedom, as well as to the legislator. Positive law ensures the effectiveness of natural law, determines its effectiveness.

What Bobbio calls Hobbesian is the third form of natural law and this differs from the previous theory in the origin of the content of regulation: it is determined by the sovereign, the human legislator. Natural law has the function of anchoring the sovereign's right and prescribing the obedience of his subjects. One rule alone is sufficient and if in a society of equals it sounds like "one must keep promises", in a society of inequalities it is "one must obey the commands of the superior". The natural norm allows the positive right, chosen and promulgated by the legislator, to be activated, and in this system, is a return to conceiving individuals as subjects. Natural law for Hobbes founds positive law as a legitimate right. In the Hobbesian doctrine, the law is all positive, except in the legitimation process. This conception represents, also historically, the transition from natural law to legal positivism. [79]

[78] N. Bobbio, *Giusnaturalismo e positivismo giuridico*, Laterza, Roma-Bari 2011, p. 110, transl. mine.

[79] *Ivi*, cit., p. 111.

One can glimpse the beginning of legal positivism in Germany between the end of the eighteenth century and the beginning of the nineteenth century, in the philosophical-cultural movement of historicism that led to the birth of the Historical School of Law. Historicism radically criticised natural law: think of Gustav Hugo's *Treaty of Natural Law as a philosophy of positive*[80] scientifically systematized the common law in force; in France the idea of codification developed during the French Revolution and became reality with Napoleon's Code of 1804. In England Bentham outlined the broader theory of codification in opposition to *common law*. He considers it uncertain and retroactive; the authority of the judges is uncontrollable and, above all, *common law,* does not follow the principle of utility. [81]

On the philosophical level, on the other hand, the naturalistic legal model is completely in crisis with Hegel. He shows that the distinction between the ahistorical natural law and the positive law irremediably contextualised is problematic. Hegel does not believe in the stipulation of a pact or contract:

> Naturalists have imagined civil society as a voluntary association of individuals, while the state is the organic unity of a people. They have placed at the foundation of this association, mistakenly confusing it with the state, a contract, i.e. an institute of private law, which can give rise to forms of partial societies in the state of nature, but certainly does not serve to explain and justify the leap from nature to history. [82]

Legal positivism is a certain approach to the study of law, a certain theory of law and a certain ideology of law. The approach of legal positivism is based on the clear dichotomy between law as fact and law as value. A judgement of fact costs the objective existence of a certain behaviour, rather than context or object; on the contrary a judgement of value evaluates and is subjective, the statement is addressed according to some axiological sense and the latter is also communicated to others, inviting them to behave in the same way. It is necessary to keep the investigation of the value that includes the problem of justice separate from the investigation of the law as a set of rules that aims to develop a legal system. The legal positivist is a scientist, he observes and describes, he does not evaluate. In the definition of law, no distinction is made between good or fair law and unjust or bad law. Positivism divorces law from its value and thus is antithetical to natural law, so a rule must be right to be valid.

[80] G. Hugo, *Lehrbuch des Naturrechts als einer Philosophie des Positives Rechts,* Berlin 1798.

[81] For a more detailed analysis of Bentham see paragraph *1.5.6 Bentham and the utility principle*, p. 22-24 of this work.

[82] N. Bobbio, M. Bovero, *Società e stato nella filosofia politica moderna,* cit., p. 94, transl. mine.

Positivism as theory refers to the state conception of law, where the legal phenomenon is connected to the formation of the modern state, of a sovereign power exercising coercion. In a positivist conception of law, the sovereign is the holder of force and enforces the norms that he identifies and promulgates by force. When considering the law as a fact, its coercive dimension must be recognised. It has been said that for positivism, the authority of the law is superior to any other source. For this to happen it is necessary that the system is complex, ordered and hierarchically structured. In it the norms cannot be antinomic (theory of coherence) and there must be no gaps (theory of completeness). Moreover, lawyers must adopt a precise method for the interpretation of norms; it is based on the mechanistic theory of interpretation:

> Legal positivism conceives the activity of jurisprudence as not producing, but reproducing the law, i.e., making explicit by purely logical-rational means the content of legal norms already given. [83]

If by ideology it is meant the complex of beliefs, opinions, representations, and values that guide a particular social group, legal positivism has a specific legal ideology: it gives the law a positive value, outside of any relationship from ideal law. In general, there are two ways to support this thesis, one more extreme than the other. First of all, one can say that positive law is just, because it is valid, thus flattening the criterion of justice on the criterion of validity; or more moderately, one can convince oneself about the usefulness of law. Regardless of a moral investigation, it is considered that the law is useful because it allows a society to achieve certain objectives. However, for both, obedience is a moral duty. To affirm an ideology means to slide from a simple description of a real and historical situation, typical of a law-positivist theory, to a moral evaluation. With ideology one communicates one's preference for a certain legal system in force.

1.8 Natural rights do not exist

Bobbio continues his analysis with the positivistic criticism of each of the three naturalistic legal positions described at the beginning of the previous chapter. According to the scholastic position, natural law is the set of essential ethical principles, but for historicism there are no self-evident principles with universal value.

[83] N. Bobbio, *Il positivismo giuridico*, cit., p. 131, transl. mine.

Each principle can be reinterpreted according to the context in which it is presented. The laws that modulate, regulate and direct life in society are necessarily historical, they change with time and this means that good and evil also take on different meanings. In a modern state the interpretation in force is one that has the approval of the majority or success at the political level.

For the Kantian position that can be recalled from before, it included the relationship between natural and positive law in the binomial of peremptory and provisional law, positivistic criticism is very simple: it is not the content that determines the law, but the form; any kind of statement can become peremptory if it follows the production process of the norm and takes the form of the legal rule. Peremptory law is the one that contains the legal rules and is therefore the only one that has value. The provisional right has no value.

Finally, Hobbes placed natural law as the cornerstone of positive law, the first grafted the second, which then continued its process and execution autonomously. The foundation of the law on one law has the defect of being an endless process: what is the first law based on, another law? For positivism, the only real principle is that of effectiveness.

> What makes a set of rules of conduct in a given society a legal system is no longer the existence of a duty of obedience on the part of its affiliates, derived from an extra positive law, but the fact, the naked fact, historically ascertainable, that that system is usually obeyed by the majority of the people to whom it is addressed. [84]

Natural law, lacking the principle of effectiveness, does not protect people in society. Over the centuries, totally conflicting natural rights have been considered as natural rights; the paradigm of justification was the same: the inscription into human nature. It changes and the requirements change, but the legitimacy of natural rights remains the same. The problem is that from such circumstantial considerations, natural law derives a judgement of normative value. From the observation of reality, descriptions can be drawn, common characteristics can be noted, but to give an axiological evaluation is an absolutely unnecessary movement, nature is like this, it is neither good nor bad.

[84] N. Bobbio, *Giusnaturalismo e positivismo giuridico*, Laterza, Roma-Bari 2011, p. 112, transl. mine.

Naturalism is back again and again and this is nothing new[85]. Most of the authors who support it and hope for its rebirth, invoke it in the name of a moral crisis, of which natural law would be the remedy; well, according to Bobbio, natural law *is not a moral but a theory of morality*[86]. A morality is a certain prescribed content that generally has as its object, the behaviour of humans, while a theory of morality is the justification of this content, a convincing rational foundation, not the content itself. This can be seen from a series of historical facts: the naturalist theory has supported and defended the most distant morals and the most opposite descriptions of nature. In natural law, the meaning of "nature" is multiform, because it is the root of law, nature is the atlas of a content that changes according to necessity, and therefore natural law is an ethical system, a theory of morals.

> In the shadow of natural law, as has been noted several times, different moral maxims have been upheld, sometimes opposite, in defence of both slavery and colonial conquest as well as freedom and wars of national and colonial liberation; both private and collective property; as much of obedience to the law of the sovereign even when it is unjust, as of civil disobedience; as much of the feudal regime and its hierarchical order, as of the bourgeois regime and its merely formal order, or of the socialist regime (in the most ancient socialists and utopians) and its community order.[87]

Another proof in support of the thesis that natural law is a theory of morality and not a moral is that the same objectives have been obtained by non-just-naturalist theories and have achieved their maximum expansion precisely when they have been unhinged by the natural law system.

The next step is to prove the unsustainability of natural law as a moral theory. There are things that are unavoidably natural: the sky, the rocks and the mountains, but also the food chain, the lions that hunt gazelles, the seagulls that teach the young to fly by diving off a cliff, the cycle of life. "Natural", however, is neither good nor bad, it is a fact and does not establish any criteria of comparison. If "natural" does not mean good, then it is not implicit that "artificial" is negative and vice versa. It happens that the side of a stone is sharp and rubbing against it causes a cut and it happens that flowers are born in the most unthinkable places; in both cases nature is only nature and does not involve any axiology. Therefore, there is no point in founding an institution

[85] *Ivi*, cit., pp. 155-156.
[86] *Ivi*, cit. p. 157.
[87] *Ivi,* pp.159-160, transl. mine.

on nature, except that of camouflage, of an absolute and transcendental foundation of an absolute, objective morality.

Demystifying natural law not as a moral but as the theory of morality does not mean denying the importance of its historical function. Natural law doctrine has argued in favour of the limitation of the sovereign, deriving the existence of such limits from the reality of laws that are superior to any deliberation of the legislator. The sovereign was forbidden to cross certain boundaries because there was a natural right that had to be respected, certain explanations of freedom are inherent in the nature of every individual and are protected by natural law.

> Natural law, as an objectivistic theory of morality, has served very well as the foundation of any theory favourable to the limits of state power. The need for a limited state of natural law has given rise to modern constitutionalism against Machiavellianism, against the theories of the reason of state and divine law of kings, against paternalistic absolutism and Hobbesian absolutism; the liberal conception of the state against the various forms of more or less enlightened despotism; the rule of law of the last century against the police state and the ethical state; and finally the theories of the international guarantee of human rights against the perennial danger of the totalitarian state. [88]

Since the second post-war period with the *Universal Declaration of Human Rights* (1948), the function of natural law has been fulfilled by international conventions and other currents of thought which, although they share the principles of the defence of the individual, theoretically have absolutely nothing in common with the doctrine of natural law.

So returns the first of the questions that required clarification: if the present is described as the age of rights and protection of the individual, is it consistent to support a positivist position? Now, after having shown the groundlessness of natural law as a theory of morality and having stressed its very important historical function, it can be argued that: natural rights do not exist, they are human artifices, techniques produced by social history. Even human rights are not natural rights and do not derive from naturalistic legal doctrine, they are positive rights, belonging to universal conventions that it is hoped will be respected by most states.

The same arguments against natural rights can be used to prove the groundlessness of the illusion of an absolute foundation of human rights. First of all, what are human rights? The definitions are all very vague: human rights are those inherent in the nature of all humans and their recognition is necessary, but the problem is that associating a

[88] *Ivi*, cit. p. 167, transl. mine.

certain characteristic with a value connotation is controversial; values change according to the context, the ideology pursued. When the historical evolution of human rights it is observed, it comes the realization that they are a variable class, they are transformed by progressive liberations.

Perhaps it is not a question of justifying human rights, but of protecting them; the *Universal Declaration of Human Rights* has founded them and proved their validity:

> The problem before us, in fact, is not philosophical but legal, it is in a broader political sense. It is not so much a question of knowing what and how many are these rights, what their nature and foundation is, whether they are natural or historical, absolute or relative, but what is the surest way to guarantee them, to prevent them from being continually violated despite solemn declarations. [89]

The proof that certain rights are well-founded is the consent collected in a given historical period, the only type of foundation proven by the facts. Fundamental rights are relative, not absolute.

1.9 Consequentialism: good before right

After having clarified the relationship between natural rights and positive rights and decreed the unnaturalness of the former, it remains to investigate the relationship that positive rights have with morals and moral laws. The two main ethical theories, deontological and consequentialist, will be introduced and will conclude with an argument in favour of the second.

A consequence is a conclusion logically deduced from a premise, but also what may or may not derive from a particular cause or condition. Consequentialist doctrine maintains that regulatory properties are strictly dependent on consequences, so the axiological value of a given entity can be inferred from the effects it has in the future. This *modus operandi* applies to the correctness of moral acts: an act is morally good and just, if the consequences of that act, or what is related to it, have a positive impact on reality. If this principle sounds similar to the principle of utility and Benthamian hedonism one would be correct in thinking so, as classical utilitarianism is to be considered a paradigmatic case of consequentialism. As mentioned above, for the principle of utility an act is morally just when it maximises the well-being of an individual and the collective well-being is the sum of the individual's wellbeing.

[89] N. Bobbio, *L'Età dei diritti*, Einaudi, Turin 2014, cit., p.25, transl. mine.

> I shall take act-utilitarianism to be the view that an act is right if its consequences are at least as good as those of any alternative. As given this view is consequentialist, welfarist, aggregative, maximizing and impersonal, and the principle of utility that it endorses sets up what I shall call the utilitarian goal. The view is consequentialist, in that it holds that acts are right or wrong solely in virtue of the goodness or badness of their actual consequences. [90]

Classic utilitarianism is a very complex theory, because not only does it postulate the maximisation of pleasure, hedonism, as an objective and the principle of utility as a method to achieve it, but it is composed of a long series of logically independent principles. This means that it is theoretically possible to support one principle and deny another; classical utilitarianism accepts them all, exposing itself to the *mercy of* many criticisms, from which the rest of the consequentialist theories that differ from each other in the principles they promote and those they deny.

Apart from the consequentialist principle that the morality of an action depends only on its consequences, the other principles accepted by classical utilitarianism are that the morality of an act depends only on the *actual consequences* of that *specific act*; therefore, predictable or expected consequences are not considered and the intention, the motive behind the action, does not matter. The moral goodness of consequences must be *assessed* and *evaluated* with specific objective criteria, and the criterion for deciding whether a consequence is good or not is *hedonistic*: the value of consequences depends only on the degree of *pleasure and the* degree of *pain*. Morally correct consequences are the *best possible* and they must be morally correct *net of total* and *universal well-being*; moral correctness depends on the consequences for the well-being of all sentient beings. On the whole, the value of individual individuals is *equal,* so the consequences are not observed from a subjective perspective, but are neutral, *agent-neutral*.

Therefore, in a consequentialist theory, an ultimate principle must be specified, "the good" to be pursued, the guide for every action that one wants to correct, because what is good, is also right. The principle is identified differently according to the authors and in general there are two types of approaches: when only one principle is chosen it is hedonistic, as for Bentham, but there are also theories that define several principles in a pluralism of values, which are then organized and classified; John Stuart Mill pursues this approach. He argues that the mere pursuit of pleasure lowers humans

[90] R. G. Frey. Act-Utilitatianism, in *The Blackwell Guide to Ethical Theory*, Second Edition. Edited by LaFollette and Ingmar Persson, Blackwell Publishing Ltd 2013, cit. p. 221.

to the level of animals, he distinguishes between high and low levels of pleasure, based on the preference of those who have experienced both[91]. Attention should be paid to the notion that the application of the principle of utility does not imply a precise calculation of consequences before action, each time; if this were the case it would be understood as a meticulous decision-making process to be followed consciously. The utility principle is a criterion, a standard that indicates what is valuable to do.

The neutrality and objectivity of the evaluation for the maximization of the common good is one of the problematic characteristics of the consequentialist approach. It has been said that for consequentialism the consequences are evaluated net of total universal well-being and each individual has the same value, so that the consequences can be weighed from a neutral perspective. But what can the normative consequences be? Whose rights are they? Who do they protect? Suppose there are five patients in a hospital waiting for a transplant; all five will die if they do not get an organ transplant and each of them needs a different organ. Now imagine that a sixth patient arrives who is healthy, and his tissues are compatible with all five patients. A surgeon is willing to operate and transfer the organs from him to the other five patients; the outcome is already known, that all transplants will be successful and there will be no organ rejection. Furthermore, the doctor and the assistants who will perform the operation will not be prosecuted. According to the principle of utility, killing the sixth donor patient would maximise the general utility, because one life is worth less than five. For classical utilitarianism the sacrifice of the donor is a morally just and correct act, because it is good according to an evaluation of its consequences: five will live[92]. A conclusion that in some ways would seem counterintuitive, even children know that killing another person is wrong.

When a neutral, agency-free theory is supported, it is very difficult to defend the rights of an individual; in the case of the example, the rights of the sixth patient who would become a donor: how can one defend his right to life? This is why some consequentialists introduce the agent, the relative subject, into the theory of values. In this way, when comparing a reality in which transplants have been completed, with one in which the sixth patient is still alive, a consequentialist can argue both the former,

[91] J. S. Mill, *Utilitarianism*, Oxford University Press, New York 1861, cit. p. 56.

[92] For other examples P. Foot, *Abortion and the doctrine of double effect*, Oxford Review 1966, pp. 28-41.

based on the principle of utility, and the latter because killing the sixth patient would not be useful in relation to his or her perspective. Also from the doctor's point of view: a consequentialist theory with agency can promote the argument that the transplant should not be done, because it would be morally wrong to kill the sixth patient and the doctor would perform an act that is not morally correct.

On the one hand, one could say that a consequentialist theory does too little, that it is inadmissible to opt for the death of innocent people, even if this would maximise the general utility. On the other hand, it is as if there were no distinction between moral dress, obligations and duties and a space in which actions, decisions, circumstances are not moral, but conventions and customs, cultural modes. For example, deciding to write a thesis on the varieties of teas present in the world is much less useful than writing one on normative ethics, but is writing a thesis a moral act? One could answer by recalling that the principle of utility is a criterion, not a decision-making practice, but the problem remains. Is it necessary to comply with the principle of utility in order to make this decision? Where is the limit?

1.10 Deontology: the right before the good

If for consequentialism the correctness of a certain action derives from its positive effects in the world, for a deontological approach it is exactly the opposite: certain actions are categorically forbitten, even if their consequences are positive and in compliance with the above-mentioned rules, everyone can pursue his or her own projects, whatever they may be, free from the constant demand to be inspired by a moral criterion that requires to shape his or her own objectives and actions in the pursuit of the common good.

A choice cannot be justified by its effects, some choices are morally forbidden even when their effects are morally good and vice versa. An action can be morally good even if its effects are reprehensible. A decision is right when it conforms to a pre-existing rule.

Deontological theories can focus on agents, *agent-centred,* or victims, *victim-centred*; both can be considered Kantian. The former establishes a regulatory framework in which obligations and permissions are assigned to an agent, these are the reasons for action. The reasons are objective, i.e. they do not depend on subjective psychological experience and are relative to the agent, in the sense that they depend on his reason and should not necessarily be shared and understood by anyone else. Therefore, it is inferred that both obligations and permissions are relative to the agent, they are instructions on how to behave in one's relationships, with one's own things. The epicentre of morality is the person, the individual who must take care to keep himself or herself bound by categorical moral laws and keep away from possible violations.

The agency can define itself through the intentions and true purposes of individuals, understood as states of mind. For the authors who build the agency in this way, forecasts, risks and motives are conceived as mental states. The double effect theory clearly makes this position explicit: it is permissible to take a moral action although at least two consequences may derive from it, one positive and the other negative. For it to be lawful it is necessary that the action is good, or at least morally indifferent; that the intention of the acting subject is good; that the good effect does not depend on the occurrence of the negative effect; and to conclude that there are no actions capable of preventing the negative effect. Therefore, it is strictly forbidden to want to have the intention to perform evil acts, such as killing or torturing an innocent person, even if this evil act would minimise the evil acts that would be performed in the future. On the contrary, if the reason for an evil act is the expectation that it will diminish negative actions in the future, the action is lawful.

Otherwise, the agency can only be determined by actions: sticking a dagger into a person's chest is different from wanting to stick a dagger into a person's chest. Therefore, if the agency is defined in this way, the obligations and permissions will describe the behaviour to be taken with regard to specific actions. The focus is on the cause: the desire to want to stab that person must cause that person to be killed; so, causing is different from omitting, allowing, facilitating, and accelerating a certain action. One wonders what one should think in a situation where A accidentally runs over B and kills him; A did not want and wasn't doing anything to facilitate the eventuality of the running over of B, B was crossing the road on a pedestrian crossing, green lit traffic light. Who is responsible? A had no intention of making what happened happen, and yet it happened.

The third way of understanding the agency could have a coherent answer for the circumstance of the accident between A and B just decrypted; the agency derives from the combination of mental states and actions, agency requires that there are intentions and causes: intentional causes. However, the agency is understood, *agent-centred theories* present very similar problems. First of all, the obsession with the self is cumbersome, it has the bad taste of egocentricity. The foundation of ethics on this agent alone, concentrated on himself, following categorical empty imperatives, encounters the problem of compliance with the law: why should this individual respect it? What value do imperatives have for the self. It should also be added that the double effect doctrine is morally unattractive and conceptually inconsistent; the differences between cause-omission, cause-permission, cause-acceleration, and cause-facilitation are morally insignificant. To conclude, in a theory that revolves around the agent, it is quite simple to legally manipulate obligations and permissions so as to turn them in one's favour.

If *agent-centred* theories establish a system of obligations and duties, *victim-centred* or *patient-centred* theories unfold a panorama consisting of rights. The pivotal norm of these theories is the right that protects an individual from being used as a means to achieve a certain objective, even a good one, without his or her consent; in other words, the right against exploitation of someone for the benefit of another. More specifically, it is understood as exploitation not only to dispose of someone else's body without permission, but also to take advantage of their work and talents without consent. To return to the scenario of the five patients waiting for a transplant and the sixth person who without his knowledge could become a donor, killing him to save the other five would be morally wrong and a violation.

The problem is that although *patient-centred theories* are better constructed than *agent-centred theories* and give a more valid reason for obeying the rules and are not self-centred, they stop at the same deontological paradox. For an agent-centred theory, the agents' reasons are conceived as neutral, i.e., if A has the exclusive right to dispose of his body, it means that every other individual has the same reason to respect this right. If morally, respect for A's right is of equal value to respect for B's right, and protecting C's rights is equally important, why is violating C's rights not allowed when doing so protects A and B's rights, which would otherwise be violated? One could solve this problem by making the duty related to C's rights a duty related to the agent; if this duty were associated with an agent, the above theories would have more theoretical material to get out of the paradox and the result would be: one cannot violate C's rights even when this violation would eliminate the possibility of similar violations in the future. The result is exactly the same as for *agent-centred* theories because if exploitation of a subject is an evil to be avoided, can it not be argued that more exploitation is worse? Isn't it a bit strange to condemn an act that conveys the realisation of a better reality? It would mean that when the broken wagon slides onto the tracks, one should not pull the lever to save the five people at the end of the track, at the expense of the driver's life[93] .

1.11 The moral foundation of rights

Having posed the question about the status of natural rights and their origin, their relationship with positive rights was investigated. The analysis of legal positivism and natural law led to the conclusion that all rights are positive, based on consensus and related to a certain jurisdiction and legal system. Natural rights are identified because they are important and fundamental rights, the origin of natural rights, however, has nothing natural about it, they are positive and must be protected because they are at the heart of the cohesive functioning of a society, state and community and it is desirable that they are protected by as many legal systems as possible.

Natural rights in the current legal-political landscape are called human rights, because their subject is the human being and humanity as a whole. Yet this makes no difference: human rights are positive and relative to a given historical moment, perhaps the naturalness of human rights should be considered more as a direction, it would sound like 'respecting human rights and protecting them via positive law, is natural and due'.

[93] The trolley problem introduced by P. Foot in 1967 sounds like this: here is a runaway trolley barrelling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person on the side-track. You have two options: i) do nothing and allow the trolley to kill the five people on the main track ii) pull the lever, diverting the trolley onto the side-track where it will kill one person. Which is the more ethical option? Or, more simply: What is the right thing to do?

Having clarified the relationship between natural rights and positive rights, it was necessary to understand their relationship to moral rights and for this purpose consequentialism and deontology were analysed. Consequentialism was described as the theory where the goodness and correctness of an act depends on the consequences of that act. In a consequentialist theory a principle to be pursued is identified and a functional system for achieving it is constructed. The problematic nature of the neutrality and objectivity of this theory was also shown: if there is no subject to protect or the interests of an individual to pursue, whose rights, are they? A consequentialist theory does too little; recall the thought experiment of the transplant in which the innocent person is killed to maximise general utility and at the same time demands too much, for it does not draw a defining line between what is moral and what is not.

A deontological theory is antipodal: a decision is right and proper when it conforms to a pre-existing norm and generally conceives of the individual as the epicentre of morality. He or she is the recipient of obligations and duties, and the individual's agency is defined by his or her intentions, actions, or a combination of both. Where deontological theories do not revolve around the obligations and duties of the individual, *agent-centred* theories, they construct a system of rights to protect the victim, the latter being defined as *victim-centred* and/or *patient-centred* theories.

It has been observed that all deontological theories encounter the same difficulties: what is the reason why one should respect one's obligations and duties? If in a consequentialist theory the reason and justification of a legal norm is obvious because it is based on the consequences of a certain action, in a deontological theory it is not so clear and immediate: correctness and justness depend on being in conformity with a norm, a movement that irrevocably recalls a further norm. The detachment between norms and morality in a deontological theory is so marked and obvious that one encounters nonsensical paradoxes, such as the condemnation of an act that conveys the realisation of a better reality. Deontology does not explain the reason for positive laws, nor for moral laws; it does not answer the question "why do they exist?", why is a certain action X wrong and, on the contrary, Y right? How is it possible to find an answer to this question? How the relationship between positive rights and moral rights can be defined?

Leonard Wayne Sumner, Canadian legal philosopher, Fellow of the *Royal Society of Canada,* and Professor Emeritus at the University of Toronto, in *The Moral Foundation of Rights*[94], asks whether moral rights exist and how they can be founded. He does so by devising a system that implements a standard of authenticity to decree the conditions for the existence of rights. This consists first of a conceptual analysis to understand what rights are and continues with a substantive analysis to understand what kind of rights there are.

An individual possesses a legal right within a legal system when the rules valid in that system confer and dispose of freedoms, claims, powers and immunities; the system is effective when private individuals observe it and when public offices, state officials, accept it. Rights are legal, they exist because they are officially recognised in a legal system and the latter is determined by specific social practices; rights are social facts: they are artifices, social creations, they depend on the context and are conventional.

> The existence conditions for legal rights are supplied by two different sorts of social practice. The content and the scope of particular rules of the system, including those which (individually or collectively) confer rights, are determined by the decisions of legislative and adjudicative institutions. Thus legal rights exist only where they have been accorded official recognition within a legal system. The existence of a legal system is in turn determined by the general social practices of complying with accepting its rules. [...] legal rights are themselves social facts. [...] Legal rights are the products of social practices, both institutional and non-institutional. [...] their artificiality. To say that legal rights are artefacts is simply to point to the fact that they are created, sustained, and extinguished by the decisions of human agents. [95]

For Sumner, legal rights are conventional and are a particular type of institutional rights. Conventional rights derive from the agreement of two or more persons and are assigned and conferred by a conventional system, in the same way a legal right must be recognised in a legal system and an institutional right in an institutional legal system. Institutional and legal rights are conventional; but can conventional, non-institutional rights confer rights? It is essential to find an answer to this question because moral rights belong to this category. What is a conventional system of rights? Can a moral system confer moral rights?

The differences between a legal system and a social moral system are manifold: first of all, a legal system is institutional, whereas a moral system is not; moreover, it is not composed of complex and determined rules, organised, and validated according

[94] L. W. Sumner, *The Moral Foundation of Rights*, Clarendon Oxford Press 1987

[95] *Ivi,* cit. pp. 67-68.

to a precise validation system. In a legal system there are sanctions, whereas moral rights, on the contrary, have no coercive power, they have moral force. Therefore, when a moral right is enforceable, its existence counts as a moral consideration in favour of a certain course of action rather than another, which means that moral rights also have normative force, since they function as reasons and motives within a context of shared morality, the origin of the normative force residing entirely in the background of shared practices and customs.

Sumner, thus realising that the normative force of moral rights is different and unconventional:

> That force does not depend on recognition or acknowledgement of the rights within any conventional rule system or by the members any institution or association. Or so we are assuming. It is this independence of conventional recognition which enables moral rights to set standards of justice for the design of conventional rule systems. If a particular system meets this standard then the rights which it confers may have moral force. However, because many systems conspicuously fail to meet these standards many conventional rights lack such force. [96]

The moral force of a conventional right is always given by reference to a right that is not, which means that if pure moral rights exist, their conditions of existence cannot be given by conventional rights. So, the search must continue: if moral rights are not conventional, what are they?

They could be natural rights. Not finding the objective foundation of moral rights in legal rights, Sumner looks for it in natural rights and, after recounting the history of natural rights and natural law theories, decrees that a moral theory is a theory of natural rights when: (i) it contains moral rights, (ii) if it refers to the possession of natural properties, and (iii) if it treats these properties as natural and objective. However, as has been shown, there is nothing natural about natural rights, in fact they exist as positives; Sumner realises that a theory of rights as natural is completely arbitrary, the arguments based on nature are inconclusive and circular.

> The problem is that too many such arguments seem to be valid, and that there seems no way to arbitrate among them by further appeals to the facts. But if there is so then nature, or even our nature, underdetermines selection of a set of basic rights and thus provides no effective control over the proliferation of basic rights principles. There is a general problem about arguments from nature: they always threaten to be either inconclusive or circular. [97]

[96] *Ivi,* cit. p. 90.

[97] *Ivi*, cit. p. 126.

Therefore, the analysis continues with the search for the foundation of moral rights in contractarianism, a form of deontological theory, as well as in consequentialism.

In a contractarianism a subject has a moral right if he possesses a corresponding morally justified conventional right. Moral principles are a collective choice, rights are artefacts that are invented at the earliest conjectural moment to be understood as the beginning of civil society, think of Thomas Hobbes' theory. Moral rights are inherently relative, existing in some circumstances and not in others. The problem is that a contractarianism is always unprepared to solve the dilemma that arises from the postulation of that original collective choice: why should certain rights be selected over others? Why should certain conventional rights be selected to be the justification, the correlates, of certain moral rights.

So, Sumner tries consequentialism. Again, a consequentialist theory of good, holds that there are certain states of the world desirable to all, or at least beneficial to the majority. A consequentialist theory is neutral, agent-neutral, whereby there are fundamental, ultimate, neutral goals; they must be combined, fused into a single overarching value and then it must be specified how to promote it. Thus, moral rights exist and are based on the consequences of actions and a conventional legal right is only justified if the legal system recognises it in its promotion of the defined objectives. A consequentialist theory does not stop at the same dilemma as a contractarian theory because the consequentialist theory of value consists of a set of neutral goals whose justificatory system is unbiased in origin.

> Existence condition for moral rights require a substantive theory of rights. A consequentialist moral framework appears to be capable, in principle at least, of supporting such a theory. [...] A consequentialist theory of rights tells us that a right is genuine just in case the social policy of recognising it in the appropriate rule system is the best means of promoting some favoured goal. [98]

When Sumner tries to ground moral rights by following a contractarianism procedure, he encounters the dilemma of the first collective choice, it is impossible to choose some laws over others, especially when that decision is made in an amoral state. Then comes the consequentialist perspective and Sumner understands that if moral rights exist, they can only and are only found in a theory where axiology, the meanings of right and wrong, depend on the desirability or otherwise of certain effects.

[98] *Ivi,* cit. p. 199.

Sumner tries to understand the relationship between legal rights, natural rights and moral rights with the ultimate aim of founding moral rights: he does this on the basis of legal rights, as if they were already there. Legal rights are taken for granted and the fact that they too must somehow be founded is not considered, just as a justification is required for moral rights. It follows that after the analysis of legal rights, differences with moral rights are drawn, without really defining their relationship. Sumner's analysis does not clarify what relationship moral rights have with legal rights, it is as if it takes the 'backwards' route: the failure to specify the relationship means that the analysis of moral rights is completely independent and autonomous from that of legal rights, they are not related from the outset.

It was fruitful to use his analysis because it made it possible to prove the existence of moral rights and made it possible to justify them in a consequentialist theory. To believe that one can decide between the two theories would be presumptuous, but for the purposes of this thesis it is important to define a coherent system that allows the questions posed to be answered, a system that is rooted in reality and refers back to real world examples. So, how to answer the question of the relationship between positive rights and moral rights?

The moral value of a certain action is given by its consequences and effects: in fact, moral laws exist to avoid deadly situations. Imagine a conjectural, fictitious beginning, a mental experiment of the beginnings in a society of minimal dimensions, a household without morals: the first moral system is built up by careful observation of actions and reactions, the chemical reactions of the first human forms that take the measures for living together and working together, determine priorities and compose scales of values. A moral system cannot be determined anywhere other than in the place of reality, where life pulsates and the need to survive guides decisions and social institutions.

Since moral values have moral force, not coercive force, it is useful to construct an ordered and organised system of effective positive laws; a legal system that is objective and totally detached from the sentimental and historical background of moral laws, but at the same time conveys the same values, proposes the same categories and priorities as the moral system to which it refers. Legal norms define an objective system that applies to the inhabitants of a certain place; it changes over time because the moral values of communities change and needs are not absolute, but relative. Legal norms ensure the effectiveness of moral values, which in themselves are not coercive and do not determine obligations, duties, and responsibilities.

It is much easier to work with legal norms than with morality because morality is always fresh, irrevocably linked to pulsating life and painful death. Legal norms are emotionless, peremptory and in a sense sterile; it is familiar to forget that they work with the same vibrant matter. This, however, is their role: to order, to clean, to make comprehensible and objective. Interestingly, while a moral law is based on concrete and natural consequences (in the sense of actual causality), legal norms establish and institute a relationship between certain realities and specific consequences that have the legal form of obligations duties and responsibilities.

As has been shown, natural rights and human rights are positive and could be considered as molecules of moral laws that emerge from the legal system and are distinguished by their importance and groundedness. Human rights go beyond the relativity and circumstantiality of local legal systems, yet they are also relative and circumstantial: they change over time, because morality does the same. The principles to be pursued are many and varied, they are historical and contextual. To think of a consequentialist theory as a system that pursues a definite and inescapable good would perhaps be to make it deontological; the principles to be pursued change, as do the normative imperatives.

1.12 Correlativity and moral capacity

In the concluding lines of the preceding chapter, it was pointed out that legal rights have stable and fixed consequences, like controlled chemical relations, which are called obligations, duties and responsibilities. The theory of correlativity places rights and duties in a very close relationship:

> The correlativity thesis makes clear that rights claims do entail duties, not for the rights holder, but for all other persons-if I have a right, then you have (and everyone else has) a correlative duty. The correlativity thesis is essential to rights theory, in conceptualising the relationship between rights and duties. It has a slogan form: "no rights without responsibilities"-rights do not exist unless others have duties. Rights are guaranteed freedoms, which then guaranteed duties for everyone else. [99]

[99] K. Abney, Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed, in *Robot Ethics, the Ethical and Social Implications of Robotics*, MIT Press Cambridge Massachusetts 2012, cit. p.39.

Subject A has a duty to respect the rights of subject B, so each right claimed by B is related to the corresponding duty of A. A duty, by definition, is a moral or legal obligation, a responsibility. If B claims a right, then A has a duty to respect it. The problem is that to interpret the reality of rights exclusively in this way is not only reductive and simplistic, but also prejudices a whole series of other relations involved.

First of all, stereotyping the relationship to a correlation means considering rights only as freedom, an indication of adopting a perspective similar to that of *Will Theory* and, as has been shown, choice theory is only one of the theories present. Accepting such a view means encountering the same obstacles regarding agency: subjects of rights thus become an exclusive category, only lucky individuals, endowed with a certain package of characteristics, qualities and capacities can be subjects of rights, any other entity is cut off, excluded.

We need to understand who these 'all' people are who must respect the rights of others. An easy answer is to define A, the subject who must respect B's rights according to the previous example, as a moral agent, a person capable of moral responsibility; in short, there are no rights without the corresponding responsibilities. Thus, being a subject of rights implies being responsible for the rights of others, fulfilling one's obligations and duties, but also being capable of responsibility; it will therefore be necessary to understand what it means to be responsible.

> It makes no sense to claim that threes or dogs or the environment have a moral responsibility to respect my freedom of speech; given that "ought implies can" they are incapable of it. If a tree falls on my head and silences me, we cannot hold t morally responsible! So, "no rights without responsibilities" carried an additional implication: on the will theory only morally responsible agent can have moral rights. If I am incapable of agency, of the exercise of liberty, of rational free will, them I am incapable of being a rights holder. If there were no mora agents, there would be no moral rights-because there are no rights without responsibilities. [100]

Thus, any entity deemed incapable of moral responsibility for its own behaviour is not subject to rights, which is why many entities do not have, and according to this kind of system should not have, any rights at all. Animals, the environment, some cases of people with differing abilities, those suffering from certain mental illnesses, and robots are not moral agents, so they should not have rights. This is where another major misunderstanding by the name of '*converse correlativity thesis'* comes in: if moral agents are the only holders of rights, and rights and responsibilities are correlated, it

[100] *Ibidem*.

means that obligations, duties and responsibilities only arise when referring to a moral agent; hence the opportunity to act in complete freedom towards those entities that are not moral agents.

> Such reasoning usually commits the fallacy of assuming a statement and its converse are equivalent- in particular, the correlativity thesis and its converse. And it mistakes the true nature of the relationship between rights and duties. The correlativity thesis: if I have a right, then all other agents have a correlative duty. The converse correlativity thesis: if I have a duty, then someone else have a correlative right. Upon a moment's reflection, the latter is absurd. [101]

It is from observation of actual exchanges and in actual practice that one corroborates or falsifies a thesis: the scenarios depicted by *converse correlativity thesis* are not the order of the day, especially in recent times. An agent has many duties that do not correspond to any rights. Take, for example, the duty of a society not to pollute the environment: in order to operate, it needs to adhere to specific standards and *guidelines* (i.e. Paris Accords) despite the fact that Nature itself has no rights. If entities such as animals and the environment cannot be regarded as subjects of rights, moral agents, because they do not have the necessary capacities, they are recognised as *moral* patients, towards whom moral agents have rights, duties and responsibilities.

Finally, the work done so far makes it possible to know that rights are not only freedoms corresponding to obligations and responsibilities of others. In certain cases, this is true, for example when Paul owes Mark five euros: Paul's right to receive the five euros corresponds to Mark's duty to return them. The correlation is not only circumstantial, but necessary, because Paul's right could not exist without Mark's duty and vice versa: they are implicit in each other. These kinds of close correlations are frequent, they are above all the result of contracts, established and regulated relationships. But there are other cases in which the correlation between rights and duties is not of this kind:

> The doctrine of correlativity sometimes assumes a particularly strong form, when it is held that rights and duties do not merely imply one another but do so because they are conceptual correlatives. This idea is that "there can be no right without a corresponding duty, or duty without a corresponding right, any more than there can be a husband without a wife, or a father without a child". [...] The relation here is like that between "right" and "left". [...] But this cannot be all there is to it, for the propositional functions, so stated, are incomplete. Rights and duties do not only connect ordered pairs (or set) of persons; they also have contents. [...] There can be an independent relation of rights and duties between the same to persons. [102]

[101] *Ivi,* cit. p. 40.

[102] D. Lyons, The Correlativity of Rights and Duties, in *Noûs*, Vol. 4, No. 1, Wiley 1970, cit. pp. 47-48.

Consider, for example, the notion of a right as an immunity, illustrated earlier in this work: the right of free speech provides that an individual is free to express himself freely, all others must respect it and must not deprive him of this right; yet would it be exhaustive to describe the respect of the right of free speech as an obligation and the right of free speech as a freedom, a privilege? It is not. Indeed, imagine a situation where Paul is on a stage giving a speech against the army's involvement in the Iraqi conflict; as he speaks, he is attacked by Mark, who disagrees and does not want Paul to convince others. Paul was exercising his right and Mark violated it, preventing him from exercising it. The right of free speech, although the name may be misleading, cannot be understood as a freedom, a privilege, protected by the corresponding obligation of another individual, in the example Mark; the right of free speech is an immunity and as such it is correlated with the disabling of someone else's right. Rights have many forms, functional to work with actual circumstances, in general they can be privileges, claims, powers and immunities; not all their combinations are rights, only those that have a specific function are. Each author identifies different combinations that lead to carve out rights with specific functions; as do different theories, think for example of Will Theory and Interest Theory.

Subjects and objects of rights have changed and are changing over time. As has been shown, the category of human being has broadened and gradually lost rigidity; this is because the attribution of agency is arbitrary, and the characteristics required to be recognised as human beings and recipients of rights change according to the moral and legal system in place. This shows that 'agency' is a concept with blurred edges and that it is important to investigate it, stretching its elastic boundaries to see where they break, if they break at all.

Therefore, the next sections will focus on the analysis of the individual as a subject of law to whom rights, duties and responsibilities are attributed. The study of deontological theories has introduced themes such as "agency", "victim" and "patient" that will be used in the analyses that follow, once the human agency and its characteristics are understood; it will be compared with artificial intelligence: the same characteristics will be sought, described, and analysed in robots. The objective will be to understand whether A.I. can be considered as agents of law, and when the question is asked whether a A.I. can, or cannot, be a moral agent, the question must be broadened, to also ask whether a A.I. can, or cannot, be a moral patient, given certain characteristics it shares with such entities.

Second Section: Moral Agency, ontological requirements, and use

2.1 Design of the section

The objectives of the *Second Section: Moral Agency, ontological requirements and us*e are to understand (i) what is an agent, (ii) what is to be a moral agent, (iii) is personhood a requirement to be a moral agent and (iv) can A.I. be artificial agents and artificial moral agents? In order to try to answer this question it will be necessary to start from the bottom: to understand what a robot is. For this reason, an analysis will be dedicated to the discipline of Artificial Intelligence, it will be made an attempt to define it and tell its story, from the first ideas born of science fiction and imagination, to the latest, real-world, cutting edge projects.

To define a boundary around the concept of A.I. and the entity of "robot", it will be explicated the reason of this work by the analysis of those situations in which A.I. act autonomously and the consequences of their actions have a moral value. Since very often A.I. are find in circumstances easily considerable as ethical, if the protagonists where humans, is it possible to consider those A.I. as responsible moral agents?

To attempt to answer this huge question, the research will start with the explanation of what agency and moral agency is, an evaluation of the needed characteristics to be considered moral agents by the legal jurisprudence will follow. The *Second Section* will be concluded with the presentation of the main four philosophical accounts on moral agents and the possibility to consider some A.I. as morally responsible.

2.2 Humanity and artifacts: an ancient bond

Identifying the fundamentals of A.I. is complicated because it contains some of the most typical characteristic and fundamental expressions of humanity. It is a compilation of humanity's greatest hits, its boldest desires, deepest fears, primal needs, and impertinent interests all rolled into one discipline. One could start with the story of technology, the evolution and motives of the relationship between the human being and it's tools. The human being is the animal that lacks any evolutionary biological specification, but it is also that animal that builds tools, Arnold Gehlen, considers the human being as a deficient animal that creates tools because it is not equipped with biological means to defend itself[103]. What relationship does humanity have with the artifacts it builds? What is the relationship between humans and technology? At the end of the nineteenth century, in precisely 1877, Ernest Knapp published *Grundlinien einer Philosophie der Technik*[104], translated *Baselines of the philosophy of technology,* where the term "philosophy of technology" was introduced for the first time.

In the twentieth century many authors theorized about this relationship: Ernst Junger[105], describes the totality of the world as subdued in the use of technology and the inevitable domination of technology. For Martin Heidegger[106] humanity's relationship with technology is alienated, not because technology is dangerous in itself, but because there is no inquisition into the essence of technology, no investigation into the effects of humanity's usage of the technology it produces. Therefore, the more humanity believes it has control over technology the more it finds itself dominated by it. The destiny of mankind is to be governed by technology and it is impossible to modify its direction by normalizing it, limiting it through a morality.

Also, for Gunter Anders the human being is bound to technology, humans cannot be separated from their technological tools, since these tools are connected to each other across an immense network and refusing to use one results in being unable to access the rest in the network. In the same vein, to exclude oneself from this network, refraining from using technological tools, would mean not being able to participate in the common collective that takes place in real spaces, with real people in both physical and virtual forms.

> The lovable mention of "human freedom" is not enough to eliminate the fact that there is a "compulsion to consume"; and the fact that in that very country where the freedom of the individual is exalted (United States of America) certain goods are called *musts*, i.e. "obligatory" goods, is not really an index of freedom. And the talk of *musts* is certainly justified: because the lack of just one of these must-equipment jeopardises the entire equipment of life, which is determined and governed by the other devices and products; whoever takes the "freedom" to

[103] A. Gehlen, *L'uomo. La sua natura e il suo posto nel mondo,* Mimesis Edizioni, Sesto San Giovanni 2010; A. Gehlen, *L'uomo nell'era della tecnica, Problemi socio-psicologici della civiltà industriale*, Armando Editore, Roma 2003.

[104] E. Knapp, *Grundlinien einer Philosophie der Technik,* Brunswick, 1877.

[105] E. Jünger, *Scritti politici e di guerra 1919-1933*, LEG Edizioni, Gorizia 2005.

[106] M. Heidegger, *La questione della tecnica*, in *Saggi e Discorsi*, a cura di G. Vattimo, Ugo Mursia Editore, Milano 2014.

> renounce one, thereby renounces all and therefore his life. Is "yes" able to do so? Who is this "yes"?[107]

This is not the only reason why humanity shares this extremely close bond with technology. Humanity is ashamed, like Prometheus. For Anders "Promethean shame" is the daily difference in level between humans and the instruments they produce, they are constantly out of sync: humans create better instruments, more functional than themselves, and can never reach the level of these tools because of their mortality.

Paul Alsberg in *The Enigma of the Human*[108] outlines an alternative analysis. For Alsberg it is unthinkable that the human species was completely devoid of biological and natural devices to respond to the challenges posed by the environment around it. Assuming that humans, as well as animals, must respond to the demand for environmental adaptation, the hypothesis is that humans were, at one stage, completely adapted to the environment and then when they no longer needed these biological adaptations, they got rid of them, replacing them with exosomatic tools. With every technological progress made these is a corresponding regression of the human body and one of its functions. What differs between human and animal is the evolutionary principle: *for the animal the body is everything* because it is the key to adaptation (endosomatic adaptation), on the contrary the human body is deactivated because the principle of adaptation for the human species is exosomatic adaptation. The instrument is not built because there is a biological deficiency, in fact there is a moment in which "the human of origin" had both body adaptation and instruments at hand and the sudden and constant use of artifacts caused a gradual deactivation of the body. The biological deficiency is produced by the instruments, which in turn proves their effectiveness and ability to replace adaptations of the human.

Tools can be weapons, utensils, clothes, but also language and concepts. Language tells stories of experiences that can be passed to others, even if they were not lived by those listening. The words used by the storyteller are signs that create a mental representation of the scenario that is taking place. In fact, language is not a biological function, but an exosomatic and instrumental one. It is precisely language that allows

[107] G. Anders, *L' uomo è antiquato, I. considerazioni sull'anima nell'epoca della seconda rivoluzione industriale,* Universale Bollati Boringhieri, Turin 2003-2007, cit. p. 12, transl. mine.

[108] P. Alsberg, *L'enigma dell'umano, Per una soluzione biologica*, Schibboleth Edizioni, Rome 2020.

the emergence of reason, the conceptual faculty that Alsberg defines as the "third faculty of the instrument".

The complete analysis of the relationship between language and reason, between words and concepts, in the full Alsbergian theory would take this introduction too far and for now this description is sufficient. Alsberg's theory has been mentioned because it is useful for this work, as it explains the reason for technology without tearing the human from its biology and allows the drawing of a continuous line from prehistory to the contemporaneity of artificial intelligence:

> With the most different types of tools we deactivate the hand, with cars we deactivate the legs, with calculators we deactivate the brain. [...] In every artificial operation, however, the guiding principle remains the same, i.e. the liberation of the body through the tools of self-adaptation. [...] Even if modern cultural life from a technical point of view seems to influence us in an impressive way, in it and behind it there is only the inconspicuous principle of liberation from the body limited by nature, the principle of adaptation to nature through the greater performativity of the instrument. [...] Thus the principle of liberation from the body celebrates its greatest triumph in modern technology. [...] But technique is not a privilege of our time; instead, the civilized human being has simply continued to build on what previous and older generations had begun to prepare. The instrument has always dominated the life of the human being. [109]

2.3 Theoretical Prodromes

The origins of technology and the ancient relationship that humanity has with it have been briefly investigated, the intuitions that can be drawn from this narration are many: (i) technology is the human figure; (ii) language and reason are inscribed in the principle of body deactivation, as are artifacts; (iii) alphabetic writing is the fixation of concepts, of those words that no longer need to be supported by the biological memory of finite beings. Alphabetic writing opens the space to the linearization of time, that is, to the history and potentially infinite preservation of social memory. When thought, language and alphabetic writing meet, logic and verbal thought are born.

> Why is what happened on the memory side with alphabetic writing so decisive? The great leap in the field of A.I. happened with the statistical version of artificial intelligence, the neural networks, this happens with the widening of the data repository. Alphabetic writing is the inauguration of *big data*, an endless archiving in principle. The limit of data storage is only empirical. The alphabet contains the possibility of digitisation, the transformation through a code of all knowledge. The greatest capitalization is the alphabet. Alphabet and *big data* are

[109] *Ivi*, cit. pp. 95-98, transl. mine.

> the same possibility, the same passage. The digital revolution is the implication of the alphabet revolution. [110]

What connects alphabetic writing to logic? Semitic, Indo-European and ideogrammatic languages are dependent on oral memory, knowledge must already be known mnemonically, this is the only way to interpret what is written. The alphabet is revolutionary: first of all, it represents exhaustively and exactly all the sounds present in language; consequently no sign has a double use, there are no ambiguities; it is cheap and it is very easy to use. Imagine now to write a period, connect it with a conjunction to another, then add a subordinate one and continue in this way, at a distance of days you can reread what you have written, you can rethink your thoughts and correct them, then think again. The exhaustiveness, clarity and cheapness of the alphabet make what is written autonomous and self-sufficient. Connecting one period with another is a logical act, it is the establishment of a relationship according to logic.

> The discourse begins to be treated differently. The written discourse describes subjects with certain properties, the copula is born, the logical connection between a subject and predicates. One can start talking about stable conditions, about subjects. Philosophy is born, in Greece not in China. Then science appears which is based only on the philosophical analytical process. The defining logical proceeding, the discursive proceeding that is philosophy, is a scheme that has its matrix in the eye and in the hand, of a mind that writes and has before its eyes what it produces. Thought is restructured starting from writing and writing is the device of argumentation. An oral mind is an acoustic mind, an alphabetic mind is a logical visual mind. [111]

Now, in order to continue the investigation of the fundamentals of the discipline that is A.I., it is necessary to understand the connection between philosophy and artificial intelligence.

The first clue can be found in the Aristotelian *Organon,* which in Greek means instrument. In it are grouped all the essays that Aristotle writes about techniques and ways to conduct investigations and arguments. Aristotle was the first to define laws governing rational thought, rational mind and established a system of syllogisms to be used to reason correctly. Is it possible to apply formal rules to draw valid conclusions? Syllogisms can be seen as the *first attempt at formal codification of "correct thinking", that is irrefutable processes of reasoning*. [112]A syllogism is a scheme in which one has

[110] C. Di Martino, *Lesson 15 Gnoseologia LM A.A. 19-20*, pp. 62-63, notes transcription and transl. mine.

[111] C. Di Martino, *Lesson 16 Gnoseologia LM A.A. 19-20*, p. 65, notes transcription and transl. mine.

[112] S. Russel, P. Norvig, *Artificial Intelligence, A modern approach ed. III vol. I*, Pearson Prentice Hall, Milan-Turin 2010, cit. p. 6.

correct premises, and equally correct conclusions are obtained, according to a deductive process: "Socrates is a human; all humans are mortal; therefore, Socrates is mortal".

At the end of the sixteenth century Thomas Hobbes compared rational human thought to numerical calculation and in the seventeenth century Blaise Pascal built the Pascal calculator to help his father, a cloth merchant, add and subtract. This device is implemented by Gottfried Wilhelm Leibniz, who builds a calculator suitable to perform all four operations. The observation of the machines at work leads to the hypothesis that they can actually think and act autonomously and this hypothesis becomes the prelude to an enormous philosophical debate: in fact, to maintain that the mind, in certain aspects, operates according to logical rules is quite different from saying that it is a formal logical physical system. What is the relationship between mind and physical brain? How does the mind flow from the brain?

René Descartes clearly distinguishes the mind, *res cogitans,* from the body, *res extensa*. Although it is true that they interact, when you decide to pick up a pencil the arm moves[113], it does not mean that the mind exists in the body, the latter cannot be considered as part of the body. Descartes defends dualism, considering a part of the human mind, the soul, not subject to the laws of nature, because it must defend and leave room for free will; in fact, if the mind is completely governed by the laws of nature and if it is not separated in some way from the biological brain, it is much more complex to argue that human action is different from any phenomenon, like lightning striking a tree. On the opposite side of the debate are the supporters of materialism, for whom the brain as part of matter and with matter proceeds according to the laws of physics; so "free will" is *simply the way in which the perception of available choices is manifested in the decision-making process*[114] and therefore to those making the decision.

However, the relationship between mind and brain is described, what is certain is that there is a mind, related in some way to a physical apparatus, that manipulates knowledge. However, where does knowledge come from? To answer this question,

[113] G. Hatfield, René Descartes, in *The Stanford Encyclopaedia of Philosophy,* E. N. Zalta (Ed.), p. 16, (https://plato.stanford.edu/archives/sum2018/entries/descartes/), 11th November 2020.

[114] S. Russel, P. Norvig, *Artificial Intelligence, A modern approach ed. III vol. I*, Pearson Prentice Hall, Milan-Turin 2010, cit. p. 9.

philosophers divide and create great interminable debates: rationalists consider human reason as the origin of knowledge. On the contrary, for empiricists knowledge is based on experience. All this is interesting for an analysis of the history of philosophy that aims to search for those signs, clues and hints that precede the first A.I.

As a matter of fact, Rudolf Carnap, the head of the Vienna Circle, developed the doctrine of logical positivism, which combined rationalism and empiricism. For logical positivism, all knowledge can be expressed by theories linked to observational statements that correspond to sensory perceptions. Even more decisively in 1929 in his work, *The logical construction of the world*[115], Carnap explains for the first time the theory of mind as a computational process, defining a clear *computational procedure to extract knowledge from elementary experiences*[116]. Finally, when one has the necessary knowledge to act, what leads to action? If acting is the ability to modify reality according to certain objectives, it is necessary to understand how these objectives are constituted and decided upon. How are the reasons that lead to an action formed? Is it true that one acts in view of certain goals? What are the characteristics of a rational and justifiable action?

2.4 Artificial intelligence: history, definition, and approaches

2.4.1 From the early beginnings to the expert systems (1943-1996)

In 1950 Isaac Asimov published *I, Robot*[117] and with it the world was introduced to the three rules of robotics:

> 1. A robot may not injure a human being or through inaction, allow a human being to come to harm. 2. A robot must obey the orders given by a human being except where such an order would interfere with the first law. 3. A robot must protect its existence as long as such protection does not conflict with the first or second laws. [118]

Although Asimov was not a scientist but instead was a writer of *science fiction,* he could be considered as simply a "visionary" of what A.I. could become. However, in the same year Alan Turing published *Computing Machinery and Intelligence*[119] *in*

[115] R. Carnap, *La costruzione logica del mondo e Pseudoproblemi nella filosofia*, UTET, Torino 2013.

[116] S. Russel, P. Norvig, *Intelligenza Artificiale, Un approccio moderno ed. III vol. I*, Pearson Prentice Hall, Milano-Torino 2010, cit. p. 10.

[117] I. Asimov, *I, Robot,* Gnome Press, New York 1950.

[118] *Ivi*, cit. p. 1.

[119] A. Turing, *Computing Machinery and Intelligence*, in *Mind LIX (236*), Oxford University Press, Oxford 1950.

which he introduced a famous *test* that would later be immortalised as the "Turing test". A revolutionary test for machine learning, genetic algorithms and learning by reinforcement.

The *test* involves party A asking questions to party B, who can be either a human or a machine. A and B can only communicate by message. Turing hoped that A would be able to say with certainty that he/she was communicating with a machine or another person. If A could not tell the difference, the machine would pass the test and was recognized as having the ability to think. Obviously, being at the very beginning of both theoretical and technological research, machines persistently failed the test and although it was never passed, Turing's test was considered, in the following years, as a satisfactory operational definition of intelligence.

In *Intelligent Machinery*[120] , a paper that actually dates back to 1947 but was published in 1969 in the wake of the enthusiasm surrounding A.I., Turing presented and rejected the arguments generally put forward against intelligent machines: humans will not admit the possibility of having intelligent rivals, that even if it were possible to create an intelligent machine it would be irreverent, that machines and humans are totally different and that Gödel[121]'s theorem shows that machines are characteristically incapable of solving problems that humans overcome without difficulty and, finally, that intelligent machines are only a reflection of their creator.

Turing continues by introducing the concept of a “child-like program": it would be much easier to create a program that simulates a child's mind and find ways to educate it, rather than producing a program that recreates the mind of an adult.

> The possible ways in which machinery might be made to show intelligent behaviour are discussed. The analogy with the human brain is used as a guiding principle. It is pointed out that the potentialities of the human intelligence can only be realised if suitable education is provided. The investigation mainly centres around an analogous teaching process applied to machines. The idea of an unorganized machine is defined, and it is suggested that the infant human cortex is of this nature. [122]

[120] A. Turing, *Intelligent Machinery*, Teddington, National Physical Laboratory, 1948, (https://weightagnostic.github.io/papers/turing1948.pdf), 1st December 2020.

[121] Gödel's theorem provides that for each sufficiently powerful logical system, statements can be formulated that can neither be corroborated nor falsified within that system, unless the logical system is inconsistent.

[122] A. Turing, *Intelligent Machinery*, Teddington, National Physical Laboratory, 1948, cit. p. 3, (https://weightagnostic.github.io/papers/turing1948.pdf), 1st December 2020.

The terms of comparison are the brain of a human child and a totally unprogrammed machine, the "brain" of the machine would be educated and moulded like that of a juvenile human. Turing identifies learning patterns for reinforcement: a system designed around reward and punishment. If the behaviour of the system is appropriate it will receive a reward, when it is not it will receive a punishment. The aim is to ensure that the system is more inclined to reproduce appropriate behaviour. Another typical human characteristic on which Turing speculates is the mimetic ability, fundamental in human learning. Ironically twenty-five years later the Massachusetts Institute of Technology experimentally proved that child teaching method, applied to computers, are highly effective and successful in reinforcement learning.

Turing also refers to the many examples of machines that imitate human organic parts: cameras instead of eyes, microphones like ears and robotic limbs in the form of leg and arm extensions. Imitating the physical parts of humans, replacing muscles and organs of which the next big step was to emulate the cognitive capacity.

> Here we are chiefly interested in the nervous system. We could produce fairly accurately, electrical models to copy the behaviour of nerves but there seems very little point in doing so. It would be rather like putting a lot of work into cars which walked on legs instead of continuing to use wheels. The electrical circuits which are used in electronic computing machineries seem to have the essential properties of nerves. They are able to transmit information from place to place and also to store it. [123]

He suggests that instead of building the analogue of a human being (equipped with cameras, microphones, ...) an artificial “brain” could be shaped to perform various tasks: to play board games such as chess or tic-tac-toe, learning and translating languages and cryptography. Ideas of which the global research effort in this field would be focused upon for the following twenty years.

In fact, as early as 1943 Warren McCholloch and Walter Pitts proposed a model of artificial neurons in which each neuron can switch on or off depending on whether a stimulus is given or not, if the stimulus is present then the artificial neuron switches on. "On" and "off" are not only to be considered as an adequate response to stimuli, but also as a suitable response to the proposition equivalent to stimuli, so potentially any computable function can be calculated by a network of neurons.

In the summer of 1956, ten scientists with different training and expertise, interested in the study of intelligence, neural systems, and automation theory, met at

[123] *Ivi*, cit. pp. 16-17.

Darmouth college for a six-week seminar. The Darmouth conference is considered to be the birthplace of the A.I., in fact it is there that the term *Artificial Intelligence was* coined by John McCharthy; the conference was funded by the Rockefeller Foundation and the research objectives are described as follows:

> We propose that a two-month, ten man study of artificial intelligence be carried out [...] The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines that use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. [124]

Assuming that the field of A.I. was nothing more than science fiction to the general population, the Dartmouth conference can be considered as a pivotal turning point. In fact, the work done at the conference showed that some of the theories, which had remained solely as thus, abstract theories, could indeed work and become reality.

One of the projects, for example, is the programming of a *general problem solver,* i.e. a programme that could solve a wide variety of problems and theorems, when formally specified, contained in the second chapter of the *Mathematics Principle* of Whitehead and Russel. Or, ELIZA a program that allows a computer to impersonate a psychotherapist. The results obtained proved to be better than expected and, one year later, US President Eisenhower approved funding for the *Advanced Research Projects Agency* (DARPA), part of the Department of Defence, to train and execute research and development projects, to expand the frontiers of technology and science, and finally, to gain technological and military advantage in view of any perceived or direct threat to the state.

These glorious years did not last long however and ten years after the Darmouth Conference came the first slowdown in funding and research for A.I., the so-called "*first AI winter*". One of the reasons for this was that investments were generally made in individuals, individuals pursuing their own interests, instead of projects with shared structured objectives. Expectations were too high and results never achieved, as Hans Moravec, an AI researcher at the time, described: "Many researchers were caught up in a web of increasing exaggeration. Their initial promises to DARPA had been much too optimistic. Of course, what they delivered stopped considerably short of that. But

[124] J. McCarthy, M. L. Minsky, N. Rochester, C.E. Shannon, *A proposal for the Darmouth Summer Research Project on Artificial Intelligence*, New Hampshire 1955, cit. p.2.

they felt they couldn't in their next proposal promise less than in the first one, so they promised more."

The winter lasted six years, 1974-1980, and although true that some researchers continued to work during this time, especially on expert systems, it is also true that the vision, the dream of A.I. and its potential was carried on and nurtured by *freelance* enthusiasts, the *poètes maudits* of the 1980s.

An expert system is a program that, produces simple inferences, based on rules and knowledge that is loaded by experienced humans and manually translated by programmers into formal language. Simply, a computer system that emulates the decision-making process of a human expert. At that time, expert systems were the programs[125] with the highest computational capacity available and were owned exclusively by universities and private companies, because they were extremely expensive, logistically difficult to move (due to their size) and above all inefficient. As programs that were experts in one specific field, they were used to work in a limited and circumscribed area of expertise. For example, in medicine the goal was for the expert system to generate medical diagnostic decisions, just as a specialist (i.e. IBM's Watson) would do; the expert systems were also used for language translation and board games.

In the following years, expert systems proliferated, hundreds were built, but unfortunately the costs outweighed the benefits, and at the end of the 1980s the futility of buying a computer for a single program, designed to achieve a single goal, was obvious. The truth is that expert systems were no better than the experts who created them, and often their decision-making process led them to make catastrophic mistakes. The market for expert systems collapsed, DARPA again cut funding for A.I. research and began the second winter of A.I., 1987-1994. The last swan song for expert systems was *Deep Blue*.

Deep Blue was a chess playing program developed by IBM, known to be the first program to play chess and win both a single game and series of games against world chess champion Garry Kasparov. In 1996, they played their first game and Kasparov beat Deep Blue by a score of four to two, a year later Deep Blue, after being reassessed

[125] A software program is commonly defined as a set of instructions, or a set of modules or procedures, that allow for a certain type of computer operation.

and upgraded based on the series already played against Kasparov, won the rematch. This event marks the end of expert systems, but it is also a huge stepping stone in the history of A.I.: it was broadcast on television all over the world, millions of people watched the game and saw for the first time a computer defeating a human chess player, and not any chess player but a chess grandmaster no less. Although at a technical level Deep Blue was a simple brute force program making probabilistic combinatorial calculations applied to possible moves and consequent choices, finding the most optimal solution. It paved the way for the introduction of various new tools and learning techniques. In order to overcome the characteristic defects of expert systems and to approach, from a certain point of view, the ancient objective in emulating the human brain, three elements were conceived: neural networks, genetic algorithms and big data.

2.4.2 Deep Learning: neural networks, big data and genetic algorithms

Neural networks are computational components inspired by the neural connectors of the human brain. They are designed to be efficient in multiple domains (as opposed to expert systems, proficient only in a single domain) and learn from experience through a reinforcement process. They work best when they aim to classify elements never seen before (i.e. facial recognition). The forerunner of neural networks is the perceptron, presented for the first time in 1958. It is a program that simulates the behaviour of a neuron and its dendritic fibres. In fact, it is a binary classifier: based on input it decides whether or not to activate the output programs. Even though neural networks had already been theorized and created in the mid-eighties by McCholloch and Pitts, nowadays, its learning *algorithms*[126] based on *back propagation algorithms,* have been reinvented.

Big data is the term that refers to a computer's ability to process and analyse huge amounts of data and has been made possible by the rapid and linear growth of CPUs[127] and the expansion of memory size. Yarowsky in 1995, published *Unsupervised word*

[126] An algorithm is a finite set of instructions, they must be non-ambiguous and characterised by a finite number of applications where the next step is always known. It must be clear when the instructions are over, the outcome is obtained.

[127] According to Moore's Law, the number of transistors in a microchip doubles every two years, while the cost of computers is halved.

sense disambiguation rivaling supervised method[128] in which he analyses the use of the word "implant" in a sentence and the mental process by which it should be understood in its context: is it an implant of an industrial complex or a dental one? Yarowsky demonstrates the need to indicate to the program the reference context, "label it"; in fact, it is sufficient to have a very large set of non-labelled text, beyond some reference examples, and the dictionary definitions of the two meanings of implant. Through a *bootstrap*[129] process the algorithm learns new patterns and with them labels new examples. The more text available, the more accurate this process is.

Genetic algorithms are calculation operations inspired by natural selection, they are used to obtain high quality results and optimised solutions. Genetic algorithms obtain useful information based on solutions that have successfully solved problems in the past. They are the most efficient algorithms for performing tasks with rigid objectives and in finding the best solution to achieve them. For example, finding the shortest distance between two points. The combination of these three elements results in a process known as *Deep Learning*.

Previously in this paragraph it has been mentioned the story of Deep Blue defeating Kasparov in the game of chess. It was defined as an expert chess system, meaning that its algorithms had been programmed *specifically* for that purpose. Rather two decades later, the research group DeepMind (later acquired by Google) created a program called *AlphaGo* as a machine that learned to play *Go by itself* without having been programmed for this purpose. There is no specific algorithm to win: AlphaGo achieves victory because it watches a very high number of games and learns to play.

In 2016 it beat one of the best professional Go players of all time, Lee Sedol, with a score of four to one. A year later, the next version of AlphaGo called *AlphaGo Master,* defeated Ke Jie, champion at the time, in a three-game series. How did it improve so quickly? By playing games against AlphaGo, its predecessor, and against itself. Within three days, AlphaGo Master outperformed AlphaGo*,* the very same version that beat Lee Sedol, with a score of one hundred games to zero.

[128] D. Yarowsky, *Unsupervised word sense disambiguation rivaling supervised method,* (http://www.ai.mit.edu/courses/6.891-nlp/ASSIGNMENT1/t4.1.pdf), 12th November 2020.

[129] Bootstrapping in AI (not to be confused with other disambiguations) is a technique used to iteratively improve a classifier's performance. Typically, multiple classifiers will be trained on different sets of the input data, and on prediction tasks the output of the different classifiers will be combined together.

2.4.3 Attempt to define "Artificial Intelligence" and approaches

Defining the discipline and the concept of artificial intelligence is both simple and complex. It is easy because the definitions that are given tend to be the description of the objective, the point of arrival, rather than the current exercise. To this is added the difficulty that comes from the use of terms on which there is no agreement, and the above concepts are used in a broad and narrow sense without further clarification, creating great confusion. Already the name "artificial intelligence" is misleading *there are few reasons, at least at the moment, to believe that the artificial intelligence of machines has much in common with human*[130] *intelligence.*

John Haugeland in 1985 defined artificial intelligence in terms of a discipline that studies how to make computers capable of thinking in the literal sense of the term[131], that is to say, to create systems that think like humans. In the same year Eugene Charniak and Drew McDermott supported the study of mental faculties through the use of computational models, with the aim of creating systems that think rationally[132]. Shortly afterwards Elaine Rich and Kevin Knight described it as the discipline that studies how to build computers capable of doing things, those things that humans currently do best, how can systems that act like humans be realised?[133] Robert Schanlkoff already understood "artificial intelligence" as the study of how to explain and emulate intelligent behaviour through computational processes: making systems that act rationally. [134]

With these considerations in mind, the history of the exploration into the field of A.I. can be seen as a dichotomy between two varying approaches in research: (i) the creation of a strong artificial intelligence and (ii) the creation of a weak artificial intelligence. A strong artificial intelligence is a general artificial intelligence that requires, consequently, the construction of a machine with an intelligence equal to that of a human. Being self-conscious and able to adapt to an environment full of

[130] J. Kaplan, *Intelligenza artificiale, guida al futuro prossimo,* LUISS University Press, Rome 2016.

[131] J. Haugeland, *Artificial Intelligence, The very idea,* First MIT Press paperback edition, Massachusetts 1985.

[132] E. Charniak, D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley, Boston 1985.

[133] E. Rich, K. Knight, *Artificial Intelligence,* McGraw Hill, New York 1991.

[134] R. J. Schalkoff, *Artificial Intelligence: An Engineering Approach*, McGraw Hill, New York 1991.

unforeseen events, learning and planning for the future. While a weak A.I. is the reproduction of specific (human) skills in an A.I., such as answering questions or knowing how to play chess. Although the strong A.I. hypothesis may seem more interesting and stimulating, historically it has not achieved many results; on the contrary, the construction of systems capable of reproducing individual and particular abilities has been wildly successful and nowadays most teams and investments are devoted to this field of weak A.I.

To these two branches of research in A.I. can also be found the disciplinary fields surrounding the ethics of artificial intelligence and that of the epistemological and cognitive sciences. The former studies the ethical and moral implications of general and narrow artificial intelligence, while the latter deals with *strong and weak A.I.* and addresses the problem of mental states. Distinguishing the two disciplines and the terms they use from one another, is not only a matter of precision, but brings attention to the shared terms being used. It allows the classification the objectives and tools of these two fields of research and it allows an equal starting point of understanding for both research field. Starting from the same theoretical implications and from the same conceptual prerequisites.

Having determined the two most influential approaches and made the necessary terminological and disciplinary clarifications, the search for a precise definition of A.I. remains open. The attribution of human faculties to A.I. has probably been the result of human desires and aspirations, as has been shown, technology and the human being have the same beginning, or at least a remarkably close conjectural beginning. The desire of humanity to recreate itself in its own characteristics is an ancient trend, to transcend themselves and become divine. According to Yuval Noah Harari, in order to defeat ageing and suffering it is necessary to obtain full control over corporality, biology and existence. Human beings face this problem with technology, in the spoils of biotechnology, bioengineering and the complete construction of non-organic beings, A.I. There is a straight line from modifying genes to avoid disease, to inserting a pacemaker into the human heart, to creating a fully mechanical brain separated from any carbon-based life.

> Even biomedical engineering is relatively conservative, insofar as it assumes that organic brains will continue to be the command-and-control centres of life. A bolder approach sets aside the organic parts altogether and hopes to engineer to be completely non-organic. Neuronal networks will be replaced by intelligent software that could navigate virtual and non-

> virtual worlds, free from the limitations of organic chemistry. After erring for four billion years in the realm of organic behaviour, life will escape into the vast and boundless realm of inorganic behaviour and take forms we cannot conceive of even in our wildest dreams. After all, our wildest dreams are still the product of organic chemistry. [135]

Harari's words take up the description of that specifically human exosomatic movement, but in them can also be found that exact enthusiasm that unifies "human intelligence" with "artificial intelligence", the same deceptive enthusiasm found in the four topical definitions of A.I. For some people, the discipline of A.I deals with the reproduction of human action, rather than the construction of agents that act rationally; to others it is the duplication of human thought, or even rational thought. To base the definition of A.I. on a comparison against certain aspect of human intelligence is not especially useful, instead it can be confusing and fictitious. This type of definition not only says something about A.I., but also contains subtle indications about what humanity and its faculties are, for instance, defining thought as a certain type of specific thought, the "rational" one, or implicitly a normativisation of a certain type of measurable and computable intelligence. However, what about emotional intelligence or creativity? Is intelligence computable? Luc Julia in *There is no such thing as Artificial Intelligence* [136]writes:

> I would define intelligence as the ability to break rules to innovate, to be interested in new and different things. To me, being intelligent means having curiosity, and curiosity for different things. But I also see it as being dynamic, global, capable of abstraction and able to evolve over time. What is considered intelligence today could be seen later on as mere knowledge. I have no idea if this applies only to humans, but I think it must apply at least to only living things. [137]

The truth is that there is not much in common between human and artificial intelligence, yet at the same time certain machinic abilities have been called human abilities and saying that human intelligence is completely different from A.I. does not solve any problem in factual terms, because *bots*, *robots* and "intelligent" machines exist and act in the real world. Think of the algorithms for facial recognition and their *bias* because they do not recognize darker faces, they are not intelligent, yet by recognizing patterns and classifying they have an effect on the world and on people. In addition to this, the semi-autonomous or autonomous weapons used in the military

[135] Y. N. Harari, *Homo Deus, Breve storia del futuro*, Bompiani, Florence-Milan 2019, cit. p. 61, transl. mine.

[136] L. Julia, *There is no such thing as Artificial Intelligence*, F1RST Editions, Paris 2020.

[137] *Ivi*, cit. p. 147.

field, projects that aim to put the human *out of the loop* are increasingly common. They are not intelligent and yet they are decisive in the life or death of civilians and non-civilians alike. The fact that there is nothing intelligent in artificial intelligence does not eliminate the challenges that the A.I. discipline poses.

Moreover, just as it makes no sense to seek the reflection of the characteristic capabilities of a machine in human beings, it is also vicious to do the opposite. When the first mathematical operations are carried out in primary school and the first logical problems are solved, the importance of "how" the final result has been achieved is always underlined, this is because observing the *modus operandi* for the resolution of certain problems allows us to understand what kind of reasoning has been carried out, to correct it if it leads to the wrong result or otherwise reinforce it. Nonetheless, the possibility of being intelligent, or not, is not questioned when it comes to human students. It is never asked "how" on a physical or biological level whether the students have these certain faculties or not. Their abilities are taken for granted since they are part of the human species. As Jerry Kaplan points out:

> When we assess the students' ability to do additions, we do not take into account how they did the work - i.e. we assume that they only used the brain they were born with and the necessary tools, such as pen and paper. Why, then, does this become relevant if the test subject is a machine? This is because we assume that a human being performing this task is using certain innate or learned skills that can in principle be applied to a wide range of comparable problems. If, however, it is a machine that has the same or superior abilities, we are reluctant to convince ourselves that something similar is happening. [138]

Machines and people have two different types of intelligence and excel in different fields: there are abilities that A.I. will never have. Take emotional intelligence, creativity and the ability to adapt to the surrounding environment; at the same time machines perform tasks that are completely impossible for humans, for example a tsunami warning system is able to sound an alarm in the face of imperceptible variations in ocean waters.

Although the faculties of A.I. and humans are different, from the beginning they have been defined and explained as similar, the former mirrors that of the latter. This has created confusion and large debates, leading authors to follow unrealistic hypotheses. Take for example, *Superintelligence*[139] by Nick Bostrom whereby paints

[138] J. Kaplan, *Artificial Intelligence, guide to the near future,* LUISS University Press, Rome 2016, cit. pp. 24-25, transl. mine.

[139] N. Bostrom, *Superintelligenza*, *Tendenze, pericoli e strategie*, Bollati Boringhieri, Turin 2018.

science fiction scenarios in which general artificial intelligences, become uncontrollable and destroy the world once human levels of intelligence are exceeded. In addition to creating a lot of background noise and disseminating unfounded information to a more general audience, Bostrom's book diverts attention from real ethical problems.

Although there are not, and probably will not in the future, be machines willing to dominate mankind, there are already specific artificial intelligences operating in the world changing reality. A computer can be more intelligent than a human, even if only in limited areas, and it is in these specific contexts that the difficulties in investigating, understanding and perhaps even solving these problems emerge.

Think of the application of A.I. in *common law* jurisprudence. In order to prepare a case, you have to review a very large number of documents concerning similar previous cases or information (precedent). It is easy to teach this task to a computer: the automatic discovery of the documents, *eDescovery*, foresees that "sample documents" are chosen from the entire collection, the sample will be the input of the *machine learning* programme which will then identify the criteria for the choice of the aforementioned sample documents; this technique is called "*predictive coding"*.

> Criteria can vary from simple sentence association to complex semantics analysis of text, context and parts of cases. The program, which is now trained, is now applied to a subset of the remaining documents in order to select others, and these in turn are reviewed by lawyers. The process is repeated until the program is able to select the right documents on its own. [140]

The result is that the computer will be much faster, more precise and accurate than the humans who were previously in charge of this job and, following this example, it can be considered that all types of work similar to this will be replaced, or have already been replaced, by a machine.

Thus, it is the use of similar terms to describe human and artificial capabilities is unclear that is the source of many misunderstandings. Unfortunately, in the context of this work, it will not be possible to build a new dictionary in order to avoid unjustified analogies. This would be fictitious and useless since the rest of the research made in this field uses such symbols and meanings. This is why terms such as "intelligence", "autonomy", "intentionality", "responsibility", will continue to be used, recognizing

[140] J. Kaplan, *Artificial Intelligence, guide to the near future,* LUISS University Press, Rome 2016, cit. p. 139, transl. mine.

their practical functionality in the context of a comprehensible and effective research, all of this with the aim of observing the emergence of these capabilities in the human and robotic condition, without ever giving their meaning by assumption.

It has been shown how the attempt to define A.I. starting from a comparison with human intelligence is unsuccessful: machines and humans are two different entities and must be analysed autonomously; even if the vocabulary built up during the evolutionary course of this discipline would suppose the opposite, having its roots in such a comparison. Yet now it is time to give a definition of "A.I." and this will be done using the Wittgensteinian model of definition: the concept of intelligence is an area, where intelligence means the ability to solve problems and to respond with *feedback* to stimuli. This provides the opportunity to consider an artificial kind of intelligence as intelligence by this characteristic when this definition is fulfilled.

2.5 Bots and Robots: areas of application

Robots are represented in various ways in the collective imagination: from images of slave robots as servants to human masters, in the form of electronic butlers or *Roombas* and other automatic household appliances, to computer programs, bots that automatically filter e-mail and trade stocks to the most unimaginable uses as automatic weapons used in war, robots come in many forms. The etymology of the word "robot" can be found in the term *robot* which in Slavic means service or forced labour and was invented by Josef Čapek. His brother Karel Čapek wrote and staged *Rossumovi Univerzální Roboti* in 1920, presenting the class of artificial slave labourers to the public for the first time. From that moment on, robots became famous characters, at least from a literary and cinematographic point of view.

While identifying the etymological origin of the term is quite simple, attempting to define the entity of robots is not, also in this case, as it was for A.I, finding a clear and shared definition is a challenge, it is like trying to hit a moving target, since research on robotics develops in real time. *Dictionaries* define robots in a way too general sense, take the *Cambridge Dictionary* definition:

> a machine controlled by a computer that is used to perform jobs automatically. [141]

[141] *Robot*, in Cambridge Dictionary, (https://dictionary.cambridge.org/dictionary/english/robot), 18th November 2020.

Or they give a way too limited definition, considering robots as only the programs that emulate human actions, as per the *Oxford Dictionary* definition, a robot is:

> (especially in science fiction) a machine resembling a human being and able to replicate certain human movements and functions automatically.[142]

One of the most famous and quoted definitions is that of George Bekey while being generic and using "bloated" concepts, provides useful indications.

> We do not presume we can definitely resolve this great debate here, but it is important that we offer a working definition prior to laying out the landscape of current and predicted applications of robotics. In its most basic sense, we define 'robot' as a machine, situated in the world, that senses, thinks and acts: thus robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. [143]

First useful indication: "*a machine, situated in the world*", a robot must be placed *in the* world. The positioning of the robot in the world draws a distinctive line between the above concept and a strictly virtual counterpart in the form of a *bot*, which is a computer program that works automatically and has the characteristic of searching for information on the Internet. This means that a robot can be both mechanical and biological but cannot be a virtual program. A second defining characteristic is that of "*a machine that thinks*", a remote-controlled machine is not a robot because it is completely controlled by the human through the remote control.

> By 'think', what we mean is that the machine is able to process information from sensors and other sources, such as an internal set of rules, either programmed or learned, and to make some decisions autonomously. [...] We define 'autonomy' in robots as the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operations, for extended periods of time. [144]

Obviously, an in-depth analysis of what the faculty of thinking and making autonomous decisions are for robots, will follow later, for the time being you can work with this material: a robot is a machine located in the world that perceives, thinks and acts. So, a robot is a programmable mechanism that moves in its environment with a certain degree of autonomy and carries out objectives.

As well as being often used, Bekey's definition has also been criticised for being too loose and lacking in an analysis of the robotic "bodily" form; in fact, what does it

[142] *Robot*, in Oxford Online Dictionary, (https://en.oxforddictionaries.com/definition/robot), 18th November 2020.

[143] G. A. Bekey, *Current Trends in Robotics: Technology and Ethics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 17-18.

[144] *Ibidem*.

mean that the robot should be able to "think"? And how should the physical constitution of the robot be interpreted, is there any additional necessary requirements?

In contrast, Alan Winfield in *Robotics: A Very Short Introduction*[145] gives precise indications about the physical appearance and shapes of robots, focusing on three types of robots. Robots as a device, as an A.I. equipped with a body and as a machine.

> A robot is: 1. an artificial device that can *sense* its environment and *purposefully act* on or in that environment; 2. an *embodied* artificial intelligence; or (iii) a machine that can *autonomously* carry out useful work. [146]

So, a robot is an artificial device that interacts with its environment, the robot perceives, *senses,* the surrounding world by means of artificial sensors and acts, *act purposefully,* in the *sense* that it can move to accomplish a certain objective. A robot is an artificial intelligence with a body, it is not only mechanical parts but also software and algorithms that allow its operation, not only the interaction between hardware and software but the interaction between the two and the surrounding environment.

The third point, a machine that carries out useful work independently, is very similar to the meaning of the Čapek brothers: robots are automatic machines that are used to carry out jobs, often uncomfortable and that nobody wants to do because they are dangerous, dirty, and dull. Robots have been used for decades in the automotive and manufacturing industries, think of the vacuum cleaner robot market like Roomba, which is expected to reach a value of $5.7 billion in 2026, and already ten years ago more than 7 million service workers were defined as robots.[147] Then there are robots used in the military, security services, scientific and medical research, entertainment, pet robots and robots that care for the elderly or children.[148]

2.6 Machine Ethics, Roboethics and the reason for this work

In the day-to-day tasks for humans nowadays it is easy to confuse and confound the human's part in these tasks with that of the robots, no matter if autonomously

[145] A. Winfield, *Robotics: A Very Short Introduction*, Oxford University Press, Oxford 2012.

[146] *Ivi*, cit. pp. 21-22.

[147] E. Guizzo, *IEEE Spectrum: World Robot Population Reaches 8.7 Million*, 2010, (https://spectrum.ieee.org/automaton/robotics/industrial-robots/world-robot-population-chart), 18th November 2020.

[148] To know more about field of application a *G. A. Bekey, Current Trends in Robotics: Technology and Ethics, in Robot Ethics: The Ethical and Social Implications of Robotics, MITP,* Massachusetts 2012, pp. 17-34.

performed or not they have an impact on the surrounding environment and its inhabitants. By "*robotic surgery*" it means the cooperation in the execution of surgical activities of humans and robots: the *da Vinci surgical robot* is a *telerobot* currently used in hundreds of hospitals around the world. The use of this technology requires the presence of a surgeon who controls the robot's arms through remote controls, performing the various techniques necessary to perform the operation. The mechanical arms are much more precise than human arms, they do not tremble and the possibility of error is much more limited, yet a negative outcome is still always possible.

> A robot surgeon performs an operation on a patient; a number of complications arise and the patient's condition is worse than before. Who is responsible? Is it the designer of the robot, the manufacturer, the human surgeon who recommended the use of the robot, the hospital, the insurer, or some other entity? If there was a known chance that the surgery might result in problems, was it ethics for the human surgeon or the hospital, or both, to recommend or approve the use of a robot? How large a chance of harm would be morally permissible? That is, what is the acceptable risk? [149]

The development of this interaction results in the analysis and development of an ethic for humans who use machines: should the surgeon in question, have performed the operation using *da Vinci surgical robots* knowing that there was be a possibility of a negative outcome? But what if the robot had not been a *telerobot*, remotely controlled by the human, but autonomous or semi-autonomous?

There is a mobile robot equipped with software but has an extremely limited capacity made to check up on an elderly person, who has lost their autonomy and needs certain treatments like taking medicine at set times. Moreover, the robot can help the elderly person to remember commitments and appointments made. What happens when the robot is asked to remind the elderly individual that their favourite television program is starting, yet, at that same moment, visitors of the elderly person have just arrived? How can a robot decide between these two values that occur at the same time? In addition to ethics for people who use machines, it is essential to develop and strengthen *machine ethics*.

> Machine ethics is concerned with giving *machines* ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making. [150]

[149] *G. A. Bekey, Current Trends in Robotics: Technology and Ethics, in Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, p. 24.

[150] M. Anderson, S. L. Anderson, *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. p. 1.

Now imagine a scenario in which the dilemma facing a robot is: to bomb a building where the enemy is hiding while at the same time having the information that non-belligerent families with innocent children are living in the same building. The robot aims to eliminate the individual, yet "knows" that killing innocent people is wrong. How should it proceed? If it is possible to build an "ethical robot", how can one be sure that the robot does not violate human rights? Who, or what, should be held responsible when these rights are violated? Can robots be held responsible? Are the laws and protocols in force appropriate for this new type of combat?

The history books tell of a war that no longer exists, in fact today the battles are fought in offices, where human soldiers manipulate *consoles* and screens, operating drones that are deployed thousands of miles away. The operators can take a break, have a chat with a co-worker next to the office *vending machine* and when the workday is over, but the battle is not, they can take their work home with them for dinner with the children. A new art of war, a combat never before seen.

There are many projects that aim to put people "*out of the loop*" and many investments have been made into this idea. Countless ethical questions take shape with the exclusion of humans from the decision-making process and the possibility of monitoring the robots in question. An autonomous robot decides, in a fraction of a second, whether to use its destructive potential or not and the result can be the death of many, soldiers and civilians alike. Such a scenario is realistic, since it is impossible for a robot to distinguish between enemy infantry and non-combatants, and this inability completely violates the fundamental ethical precepts of a *just* war according to the *jus in bello,* regulated by the Geneva and Hague Conventions, as well as other international protocols. [151]

Research and significant investments promote projects aimed at removing "*man*" from *manual*, such as autonomous machines and *bots* that manage the stock exchange. It is legitimate and indispensable to ask all these questions, not only in life-or-death scenarios, but also in the most ordinary and habitual situations where robots, and more generally AI, are used without much publicity.

[151] N. Sharkey*, Killing Made Easy: From Joysticks to Politics* in *Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, pp. 115-118.

That is why this thesis is developed in the field of roboethics, that interdisciplinary research that aims to understand the ethical implications of robotic technology and its consequences. The focus will be on robots with a certain degree of autonomy and the possibility that they may have rights and responsibilities. The final objective is to answer questions: (i) "What does it mean to be a person and what are the requirements to be a moral agent? (i.e., to be held responsible)", (ii) "Can a Robot be considered a person/moral agent? (ontological requirements)" and (iii) "If Robots can be considered as such, should they be? (ethical implications)".

2.7 What is Moral agency and why is it important: theory of mind

An agent, by definition, is an entity that acts and has power, something that produces or is capable of producing an effect: an efficient or active cause. Therefore, an agency is the manifestation of this capacity. A moral action is an action whose effects, whose impact on reality, can be considered right or wrong, good or bad; hence moral theories are the theories of correct action, they indicate those principles that should guide action. Consequently, a moral agent is an agent capable of performing morally qualifiable actions: on a theoretical level, both human and artificial agents can be defined as moral agents, yet on a practical level can moral rights and responsibilities be attributed to an A.I or a robot?

As shown in the *First Section: Rights*[152] of this work, according to the supported theory of rights, certain aspects of moral agency appear to be either essential or not, in fact, this also applies to moral agency itself[153]. If a deontological theory is applied, it is necessary to evaluate the intentions of the body that carried out the action, and this implies that the body has an intentional mind and certain states of mind; when the morality of certain actions is evaluated only based on their consequences, the moral agency may not be considered.

Why is moral agency so important? As far as humans are concerned, most ethical paradigms assume that agency is essential to the theory of mind. The theory of mind is the ability to attribute mental states to interactable subjects. Nevertheless, how is it

[152] First Section: Rights, pp. 11

[153] This topic has been dealt with in chapters *1.9 Consequentialism: the good before the good* and *1.10 Deontology: the right before the good*, pp. 42-50 of the present work.

possible that one person can “read the mind” of another and vice versa? How is it possible to grasp the intentions of others, i.e. the real motives for taking action to achieve individual desired goals?

In the early 1990s the so-called “mirror neurons”, a class of premotor neurons in the premotor cortex, were discovered in a macaque brain. When the macaque moves to grasp a stick they are activated, the mirror neurons are activated when an action is performed with a precise target, but the same thing happens when the macaques observe each other:

> It has been hypothesized that through the activation of these neurons a direct form of understanding of the action is achieved. The observed behaviour is learnt pre-reflexively because it is constituted as a motor act with an objective, by virtue of the activation of the observer's brain of neurons used for the motor fulfilment of similar objectives. [154]

This also happens in the human brain, when one human observes another human being in motion, the observation of other people's movements activates the same neural circuits that allow one's own movement, recognizing the sequence performed by others as one's own. These neural circuits are called mirror neuron systems. This capacity has a *pro-social*[155] character: the recognition of other people's movements as one's own creates a favourable climate for socialization, for communication through signs and for opening the possibility of vocal language[156].

The importance of the system of mirror neurons is not only fundamental for the recognition of other people's movements, but also for the social identification of the other by oneself: emotions and sensations seem to be mapped according to the same mechanism. The other is present and is similar in that it is moved by objectives, emotions, and sensations. When one catches the contrite expression on the face of

[154] V. Gallese, *I due lati della mimesi, teoria mimetica, simulazione incarnata e identificazione sociale,* in *Scienza e Mimesi, Ricerche empiriche sull'imitazione e sulla teoria mimetica della cultura e della religione,* edited by S.R. Garrels, Cortina Editore, Milano 2016, cit. pp. 130-131, transl. mine.

[155] *Ivi*, cit. p. 133.

[156] To elaborate on see G. Buccino et al., *Neural circuits underlying imitation learning of hand actions*: *An event-related fMRI study*, in *Neuron 42*, 2004, pp. 323-334; M. Iacoboni et al., *Grasping the intention of others with one's own mirror neuron system*, in *PLOS Biology 3*, 2005, pp. 529-535; F. Pulvermuller, *The Neuroscience of Language*, Cambridge University Press, Cambridge 2002; V. Gallese, G. Lakoff, *The brain's concepts: the role of sensory-motor system in reason and language*, in *Cognition Neuropsychology 22*, 2005, pp. 455-479.

another, it is constituted and understood directly through an embodied simulation and is felt as if the observer were also experiencing it. [157]

There is, however, a small congenital defect linked to this fantastic ability: human beings apply the theory of the mind to absolutely everything around them.

> In human terms, most ethics presumes agency matters, because of our theory of mind. The ability to detect agents within the human community was a key to our evolution as a social species, and we are so hardwired for it that we attribute agency promiscuously, naively attributing the human ability to choose on the basis of reason and goals to dogs and cats, cars and trains, even trees and clouds and volcanoes and the weather and... well, just about everything we interact with. Because this "intentional stance" works so well in understanding other humans, we have a tendency to use it to explain everything: so the Hawaiians explained volcanic eruptions by the agency of a displeased goddess Pele, and the Greeks explained shipwrecks as due to a similar rage of their god Poseidon. [158]

Intentional states are used to explain everything, even robots. It is very common to attribute human emotions to so-called "socially active robots", even though there is no trace of emotion in them and the relationship is to be considered definitively unilateral. The suspicion is that mental states are attributed on the basis of the relationship that humanity has with the entities around them rather than on the basis of defined ontological characteristics. [159]

2.8 Requirements to qualify as a moral agent

It is important to understand what the mechanism for the attribution of mental states is, i.e. the theory of mind, because the assignment of moral agency is based on it. Some theorists argue that robots cannot be considered as moral agents because they cannot have an inner moral life and/or an emotional system, but beyond the possibility to recreate emotions and debate about them and their importance, is absolutely useless. In fact, from a legal point of view moral agency does not require the presence of a functioning emotional system.

In the legal field agency is treated in terms of capacity, Art. 1 and Art. 2 of the Italian *Civil Code* define it as:

[157] V. Gallese, *I due lati della mimesi, teoria mimetica, simulazione incarnata e identificazione sociale,* in *Scienza e Mimesi, Ricerche empiriche sull'imitazione e sulla teoria mimetica della cultura e della religione,* edited by S.R. Garrels, Cortina Editore, Milano 2016, cit. p. 135, transl. mine.

[158] G. Verruggio, K. Abney, *Roboethics: The Applied Ethics for a New Science*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, p. 355.

[159] J. Borenstin, Y. Pearson, *Robot Caregivers: Ethical Issues across the Human Lifespan, in Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, pp. 251-265.

> Art. 1 (Legal capacity) Legal capacity is acquired from the moment of birth. The rights which the law recognises in favour of the conceived are subordinate to the event of birth.
>
> Art. 2 (Age of majority. Ability to act) The age of majority is set at the age of eighteen. With the age of majority one acquires the capacity to perform all acts for which a different age has not been established. This is without prejudice to special laws establishing a lower age of ability to do one's job. In this case the child shall be entitled to exercise the rights and actions that depend on the employment contract. [160]

Thus, legal capacity is acquired by birth and lost by death, with birth one acquires all constitutional rights (right to life, freedom of speech, ...). Then there is the capacity to act that is acquired with adulthood, no longer being considered a minor. This is the capacity to perform all the acts for which no other activity is foreseen, to sign a contract for the purchase of a property, for example, and most of the activities that regulate relations with others.

By capacity to act it is meant the capacity to intend and to want: the actions of the one who is recognised as having the capacity to act are acts identified as legal. A minor is not considered autonomous because he/she does not have the capacity to act, *minus habens* have these capacities taken from them because they are not autonomous and do not have the capacity to intend or want. For this reason, they cannot be responsible for their actions and therefore cannot be attributed the capacity to act.

In a certain sense it can be said that legal systems revolve around the existence of two decision-making systems: the deliberative system and the emotional system. For a long time, they were considered autonomous and unreliable, think of Descartes, who drastically divided emotion and intellect, and in recent decades, following speculative analyses and clinical discoveries, a rapprochement has begun. In 1994 Antonio Damasio was the first to recompose that alliance, in fact, in *The Error of Descartes* [161] he emphasizes the essential cognitive value of feeling, clarifying its neurobiological functionality and the essential interweaving with rational action.

As has been shown in the previous section, depending on the theory of rights to which agency is assigned, the systems differ in each case, for *will theory* the deliberative system is the pivotal system, while for *the interest theory*[162], the emotional

[160] Art. 1, Art. 2, *Italian Civil Code*, transl. mine.

[161] A. Damasio, *L'Errore di Cartesio; Emozione, ragione e cervello umano*, Adelphi Editore, Milan 1995.

[162] It is recalled that in an *interest theory* rights must protect the interest of the holder, i.e., if there is an interest there is a right and every person has the duty to respect the rights of the person having an interest, while *will theory* states that the function of a right is to give one person control over the duty

system is fundamental. However, on a legal and judicial level, whether the deliberative and emotional system are connected or separated is not very important. In fact, if it is true that the capacity to act is assigned according to the balance between the two, it is also true that the only system whose presence is really decisive on this level is the deliberative system.

> Psychopaths, sociopaths, rational agents with dysfunctional or missing emotional affects, are still morally and legally responsible for their crimes; whereas those who have emotional responses, but cannot exercise rational control (like severely mentally disabled or infants) are not. [163]

Whatever a child does, he is not considered capable of intending and wanting, he is not held responsible for his actions, this despite the fact that the child has a fully functioning emotional system and a deliberative system that could potentially work just as well but is for most cases considered too immature.

In 2007 thirteen-year-old Paris Bennet stabbed his sister Ella, four years old, to death. Paris has been diagnosed as a psychopath, genius with an IQ of 141 and is now a twenty-five-year-old man in prison at *Ferguson Unite* a *Texas State Prison*. He doesn't accept his diagnosis, when he talks about the crime he committed, he doesn't recognize his illness as a decisive factor in the event; Paris explains that the only reason he killed his sister was the hatred he felt towards their mother, the killing of Ella was the most effective way to make her suffer. Paris was perfectly aware that he was committing a crime, he planned the act down to the smallest detail and after having committed it, took full responsibility.

Psychopathic disorder is a personality disorder that is generally identified with antisocial behaviour starting in childhood and for which there is no cure. Often the psychopath is a fascinating individual, capable of transporting listeners with his or her own words, skilful in telling and praising himself or herself, having a great opinion of themselves. A psychopath is a chronic liar and manipulator and above all he/she does not manifest guilt after having performed an action with negative consequences, they are not able to feel sincere emotions and are not able to control their impulsive actions.

of another person. This theory recognizes freedom as the foundation of every right and consequently having a right means being free to choose, to do or not to do.

[163] K. Abney, Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed, in *Robot Ethics, the Ethical and Social Implications of Robotics*, MIT Press Cambridge Massachusetts 2012, cit. p.46.

Paris Bennet has such an incurable disturbance that even a simple observation of an interview with him allows each of the above-mentioned characteristics to emerge naturally: he is extremely intelligent, he has the ability to plan meticulously his every action and is able to have conversations that at first glance may seem absolutely "normal", his deliberative system is working, highly functioning even. Yet, the emotional one is completely dysfunctional or completely absent. Paris Bennet has been recognised as having the ability to act and the consequent responsibility for his actions before the age of maturity, and thus, since the age of 13 he has been locked up[164].

Referring to the information collected so far, some A.I. could easily be considered moral agents affected by a psychopathic disorder and therefore responsible for their actions. On the other hand, unlike what happens with humans suffering from psychopathic disorder, reprogramming (aka. Re-education) and correcting the dysfunctional behaviour of an A.I. is much simpler. Before concluding and taking a hasty position it is good to understand how the deliberative system works.

The deliberative system is the system that involves the ability to shape alternative futures in the form of mental representations and choose which of these futures one prefers. As it has been shown, the legal system attributes moral agency, and recognizes responsibility for one's actions, to those entities that have a functioning deliberative system. This means that a moral agent is a free and autonomous entity, they are responsible for their actions because they are performed freely, as they are intentionally and consciously provoked. All these elements emerge in a *standard theory* of moral agency and action.

In a *standard theory*, an individual is to be held responsible for the behaviour he or she has intentionally chosen. Behaviour is the way of acting in relation and interrelation with the environment, so a voluntary behaviour is a movement towards the environment, or that the subject directs towards itself, the important thing is that somehow it can be considered as external, happening in the real world. Therefore, an external behaviour that has the characteristic of being intended, wanted, sought after and chosen among many other possible behaviours. This precise and specific choice depends on the presence of a particular type of internal state: a mental state. Mental

[164] C. Lee, *Our Friends in Prison*, (https://ourfriendsinprison.weebly.com/charity-lee.html), 30th December 2020.

states are representations of contents that propose reality in the form of belief, hope, judgement (etc.); all these forms are intentional. The intentional states are what transforms mere behaviour into action, in fact an action is a behaviour that causes an intentional state of mind.

What makes an agent capable of intentionality is their freedom. Being free means having control over one's actions and the ability to predict possible scenarios in order to choose freely among them. An intentional state of mind is characterised by the intention to perform an act, this intention is possible because the agent is free. The action is an exercise of freedom and the latter is necessary to be able to consider the agent as morally responsible.

The next chapter will be about the main philosophical theses about the moral agency of A.I., it will be observed how in the arguments built by the authors the constitutive elements of the deliberative system continue to be used, defined, and related to corroborate a certain thesis. The elements called into question by the authors and the main object of the rest of this thesis are autonomy, freedom, responsibility, intentionality, and conscience.

2.9 Can artificial agents be moral agents?

BLU-108 is an automatic armed robot that is used as a submunition device, it weighs about 27 kilos, contains four warheads, has an orientation and stabilization system as well as a radar altimeter and a rocket engine. Ten BLU-108s are loaded into a single rocket and when the ideal altitude is reached the BLU-108s are ejected and using a double parachute system are placed in a vertical position. The parachutes are ejected and the BLU-108 is opened and the four warheads are placed outside the shell to be released. BLU-180 starts spinning, moving upwards and then releases the warheads.

Each warhead, which is in the form of a disc (*Skeet warheads,* almost like a large hockey puck), is autonomous and "intelligent", in fact through a double system of sensors it detects the presence of targets. When the target is found it fires an explosive bullet aimed at the target. *Passive infrared sensors* are used to track hot targets, such as vehicles, while *active laser sensors* are used to monitor the actual profile of the *target* in question (*target profiling*). The projectiles and shrapnel can hit solid targets, the metal of a tank for instance, penetrating them. It can also destroy more delicate targets by exploding and fragmenting them. [165] BLU-108 was developed by the U.S. Air Force in collaboration with Textron Systems, the efficiency and continuous improvement to this device has made the BLU-108 the submunition of choice for the U.S. Air Force and the U.S. Navy. The lethality and versatility of the BLU-108 make it the weapon of choice for the warmongers of the 21st century.

> The BLU-108 is not like other bombs because it has a method of target discrimination. [...] the Skeet warheads have autonomous operation and use sensors to target their weapons. The sensors provide discrimination between hot and cold bodies of certain height, but like autonomous robots, they cannot discriminate between legitimate targets and civilians. If BLU-108s were dropped on a civilian area, they would destroy buses, cars and trolleys. Like conventional bombs, discrimination between innocents and combatants requires accurate human targeting judgments. [166]

If BLU-108 had been a human, the situation just described would have been easily considered as moral and if BLU-108 had been a human, as a moral agent, he would have been held responsible for its actions and mistakes. Is it possible to consider artificial agents (robots) as moral agents? However, taking each concept one at a time, agent and moral.

John Sullins in *When is a Robot a Moral Agent?* [167]briefly introduces a situation where two nurses, a human and a robot do their work taking care of patients, the human is certainly responsible for its patient, but is the robot? Then he goes on to indicate there are three possible answers to this question: one can pretend that the robot's context of action is not moral, if it seems so it is just an illusion, otherwise it can be considered "pseudo-moral" since the robot lacks the requisites to be seen as a moral agent. Or one can contemplate the serious possibility that the robot's actions are moral.

Generally, in the field of *information technologies* the paradigm through which the relations between human and technology are read, consider the "technological piece" a mere tool. So, there is the *user* who uses the tool, this interaction potentially creates negative effects, which in turn create a victim: the human who uses the tool is responsible.

[165] To see BLU-108 operating, (https://www.military.com/video/logistics-and-supplies/air-force-equipment/us-air-force-sensor-fuzed-weapon/3812620039001), 20th November 2020.

[166] N. Sharkey*, Killing Made Easy: From Joysticks to Politics* in *Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, p. 119.

[167] J. P. Sullins, *When Is a Robot a Moral Agent?*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 151-161.

When dealing with *telerobots*, for example the *da Vinci surgical robots mentioned* above, applying the above paradigm makes sense:

> Telerobots are remotely controlled machines that make only minimal autonomous decisions. This is probably the most successful branch of robotics at this time because they do not need complex artificial intelligence to run; its operator provides the intelligence for the machine; [...] Obviously, these machines are being employed in morally charged situations, [...] The ethical analysis of telerobots is somewhat similar to that of any technical system where the moral praise or blame is to be born by the designers, programmers, and users of the technology. Because humans are involved in all the major decisions that the machine makes, they also provide the moral reasoning for the machine. [168]

In these cases, the robot system is not totally considered as an agent, but more as an actuator of certain actions, and it should also be added that the causal chain that leads to the occurrence of a certain effect is easily retraceable: who told the robot to do X? A remote-controlled robot is a tool that is used and if there are errors, if there are victims, the human is held responsible.

However, is a robot with a certain degree of autonomy and performing its tasks in a reality shared by human beings simply a passive tool that is used by the human being? A robot is considered autonomous when it is the main actor in its decision-making process, when it decides what actions to take through its programs. It is obvious that the programmers, the builders, and the owners of the robot are to some extent responsible for its actions: the programmers are responsible for the *code* that is loaded in the software, the builders of the hardware, and the owners for the context of use. In the end it is the robot that decides with a certain level of autonomy and its decision emerges from the interaction of *software* with *hardware and* both with the surrounding environment. Therefore, if decisions lead to actions and these actions have moral value: the robot could be considered an artificial moral agent and the human being would no longer be the only possibility to be assigned a moral agency. Is being a person a necessary requirement for moral agency?

In philosophy there are four main theoretical analyses about the moral agency of robots, presenting them will be useful to understand how to set up a coherent autonomous analysis. The first considers robots, as they are known, as non-agents, but perhaps in the future they will be. For the second, however, they are not and will never become agents. There is an opposing thesis that robots are moral agents, whereas humans cannot be considered as such. The last consideration is represented by Luciano

[168] *Ivi*, cit. p. 154.

Floridi, who goes beyond all the paradoxes implicit in the three previous visions, adopting a *"mind-less morality"* that avoids controversies such as free will and intentionality, questions that have never been resolved in philosophy. Considering it inadequate to apply to artificial agents' categories that are in any case obscure.

2.9.1. Dennett: future moral agents

The first move Daniel Dennett makes in *When HAL Kills, Who's to blame?* [169]is *to* underline that, from a legal point of view, guilt for having done an evil deed does not imply the presence of any kind of emotion. Remorse and guilt are not necessary to be considered guilty:

> 'To have performed a legally prohibited action, such as killing another human being; one must have done so with a couplable state of mind, or *mens rea*. Such couplable mental states are of three kinds: they are either motivational states of purpose, cognitive states of belief, or nonmental states of negligence.' The legal concept has no requirement that the agent be capable of feeling guilt or remorse or any other emotion; so-called cold-blooded murders are not in the slightest degree exculpated by the flat affective state. [170]

He continues analysing the Deep Blue match with Kasparov[171] and defines the program as an intentional system with desires and beliefs, therefore characterized by cognitive and motivational states. Deep Blue is a rational system, in fact rationality is required to make decisions based on desires and beliefs; one could make a note of the terminology chosen by Dennett, in fact the "cognitive states" and "motivational states" are more in the form of objectives dictated by programming, built specifically to bring the system to the victory within a game of chess.

However, Dennett believes that knowledge and motivation are not sufficient to have bad intentions, those required at a legal level: the scope of Deep Blue is too limited for there to be any expression of moral responsibility. To be responsible Deep Blue would have to have a higher level of intentionality, from which would result the ability to have beliefs about one's own beliefs and desires. This kind of intentional order would seem to be a necessary condition for moral responsibility. Although Deep Blue does not have a high enough level of intentionality it shows that it knows how to empower and control itself, and even if it only does this up to a certain point, it is enough to believe that the implementation of Deep Blue could lead to a sufficient level

[169] D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365.

[170] *Ivi,* cit. pp. 351-352.

[171] Please refer to pp. 71-72 of this work for more information on the game.

of intentionality to be held responsible: HAL seems to be quite such an implemented system.

> Adding the layers of software that would permit Deep Blue to become self-monitoring and self-critical, and hence teachable, in all these ways would dwarf the already huge Deep Blue programming project- and turn Deep Blue into a radically different sort of agent. HAL purports to be just such a higher order intentionality system - [...]. HAL is, in essence, an enhancement of Deep Blue equipped with eyes and ears and large array of sensors and effectors [...] HAL is not all garrulous or self-absorbed; but in a few speeches he does express an interesting variety of high-order intentional states, from the most simple to the most devious. [172]

Suppose HAL exists. HAL is a system with a high order of intentionality, it can speak and has a very well-organized memory. At the same time, however, HAL also has the opportunity to be evil: it could lie to itself and others. Should one believe what he, or it, says? HAL also talks about his emotional states, which are obviously virtual. However, in reality, one does not really care whether they are recognised as real or not, because the full functionality of the emotional system is not a necessary characteristic to be held responsible for one's actions. In fact, the only two circumstances in which one is exonerated of one's reprehensible actions are: for mental insanity or because one is forced by certain circumstances (coercion, self-defence etc.). [173]

A small reflection. Deep Blue has a level of brute force intentionality and is too simple for fully realised intentionality. It has desires and beliefs but cannot reflect recursively on them. It has already been said that these desires and beliefs are more like goals and purposes, but even if they are, the possibility of having a recursive thought opens up with the power of speech and the same goes for self-consciousness.

When Dennett mentions *self-monitoring* and *self-control*, he refers to the ability to look outward, to be self-conscious. Self-consciousness also has a deep connection with language, think about when one talks about themselves to others; Dennett believes that by increasing the intentionality level of Deep Blue, by re-implementing its software, a form of self-consciousness could be created. Does having a self-awareness come from the implementation of software and even if it does, does it even

[172] D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, p. 354.

[173] The balance between the emotional and deliberative systems is analysed at *2.8 Requirements to qualify as a moral agent* in the present work.

matter because the question also arises in whether being self-aware is even necessary in order to be considered responsible?

However, in conclusion, in 1998, Dennett does not consider robots as agents as they are but does not exclude the possibility that they may be divided up in the future as a result of further implementations and developments. His discourse is built on the existence of a developed intentional system, and with it the presence of cognitive and motivational states, rationality, and self-consciousness. The emotional system is to be taken into consideration, but it is not considered important because, at a legal level, it is not decisive.

2.9.2 Robots will never be moral agents

In response to the thesis[174] that in the future artificial intelligence will become so powerful that humans are useless to themselves, Selmer Bringsjord states that there is no reason to fear robots, the real source of danger is how humans use robots. Technological development in the field of artificial intelligence will never be able to build a dangerous robot, in fact no kind of robot can ever be a moral agent, neither now nor ever. It can only be so, because a robot can never be autonomous: it will never be able to do anything other than follow the program that characterises it constitutively.

Bringsjord demonstrates his thesis through an experiment: the protagonist is PERI a robot programmed to make moral decisions. PERI has a sphere that he can hold or let go, holding the sphere corresponds to a good action, a morally correct action, while letting go coincides with a morally reprehensible action.

> Part of my discomfort with the postulate that intelligent machines will exceed us in all regards, and no human effort will be expended for anything, is that it's supposed to entail that robots have autonomy. I very much doubt that robots can have this property, in anything like the sense corresponding to the fact that, at the moment, I can decide whether to keep trying, or head downtown and grab a bite to eat, and return. [175]

It is the program that decides whether PERI should hold on or let go, and human programmers have written the program, not PERI. In the event that the robot should perform an action unforeseen by the program, this is due to a *random* event or possible influence, PERI cannot be the determining cause.

[174] Bringsjord refers to the article by B. Joy, *Why the future doesn't need us*, in *Wired 6.08,* 2000, (https://www.wired.com/2000/04/joy-2/), 17th November 2020.

[175] S. Bringsjord, *Ethical Robots: The Future Can Heed Us*, in *AI and Society* 22, 2008, p. 541, (https://doi.org/10.1007/s00146-007-0090-9), 18th November 2020.

Such an argument has several problems, one of them being the certainty that human decision-making is outside any kind of external influence. In this context it is not even necessary to take a position in the debate between libertarian and determinist theories, whether the humankind has free will or not, it is absurd to maintain that the decisions he makes are not also the result of social models, cultural influences, imitation etc. [176]

Deborah Johnson, in *Computer Systems, Moral Entities but not Moral Agents*[177]*,* also argues that robots and computers cannot be considered moral agents. Her analysis is much less *trenchant* than Bringsjord's and has the merit of dwelling on stimulating concepts. For Johnson there is no doubt that any computer[178] is a moral entity: computers have a very close link with humans, they are sociotechnical systems and this makes them involved in moral value circumstances, but this does not mean that they are moral agents, to be so they should be independent, autonomous and have mental states.

As has already been introduced previously, the standard theory of moral agency and action has as its central pivot the idea that an individual is responsible for his or her intentional behaviour, individuals are not responsible for those behaviours they did not intend to have or for the consequences of those behaviours they did not foresee. So a voluntary behaviour is a behaviour that one intends to have, that goes outwards and that is caused by a particular type of internal state, the so-called mental states. In fact any type of behaviour can be explained by its causes but only the intentional action can be explained through mental states.

Generally, the theories of action consider an action a human behaviour which has the following characteristics:

> First, there is an agent with an internal state. The internal state consists of desires, beliefs, and other intentional states. These are mental states, and one of these is, necessarily, an intending to act. Together, the intentional states [...] constitute a reason for acting. Second, there is an outward, embodied event- the agent does something, moves his or her body in some way. Third, the internal state is the cause of the outward event; that is, the movement of the body is rationally directed at some state of the world. Forth, the outward behaviour (the result

[176] R. Dawkins, *The Selfish Gene*, Oxford University Press, Oxford 1990.

[177] D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 168-183.

[178] Johnson does not refer to robots or A.I. but to computers. In this paragraph it will be maintained the terminology of her choice, from the analysis of her paper what she considers as "computer" would not seem to be different from "robot" or "A.I.".

> of rational direction) has an outward effect. Fifth and finally, the effect has to be on a patient - a recipient of an action that can be harmed or helped. [179]

An agent has internal states they are mental states, so they should take place in the mind, and one of them must necessarily be an intentional state of mind which is the reason for the action. This action is an action in the real world and is caused by the intentional state of mind, the effects of the action take place in the real world and they are directed towards another subject who suffers.

Computers and artificial systems have all these requirements except the first one: the behaviour of computers corresponds to a real event in the world and an internal state is the cause of it, this event has effects and they also present themselves in the world and they can address themselves to another subject. Yet computers cannot have intentional internal states of mind.

An intentional state of mind is characterised by the intention to perform an act, this intention is possible because the agent is free. The action is an exercise of freedom and this freedom is necessary to be able to hold the agent morally responsible. Johnson goes on to add that the freedom of agents is crucial for understanding their behaviour, in fact in the experience of each day it can observe different agents performing different actions, making different decisions even in the context where they have the same desires and beliefs.

Therefore, in order for there to be an action, the presence of a precise internal state, an intentional state of mind, is necessary and the latter emerges only in the complete freedom of the subject. For this reason, computers cannot perform intentional actions, they cannot have intentional mental states because they are not free. In the analysis of the requirements to be considered responsible moral agents, free will seems to be a nodal characteristic; for this very reason, in the present work, neither the freedom of human beings, nor the lack of freedom in robots, can be taken for granted. The argument by which Johnson tests the freedom of human individuals is rather weak; how can one know that two individuals with the same knowledge, beliefs and desires act differently just because they are free? How can one know that knowledge, beliefs, and desires are the same? Is it possible that they are equal because they are two distinct

[179] D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. pp. 173-174.

individuals who, because they are different, will experience the world in a multifaceted and original way?

However, Johnson goes on to argue that even if in the future there should be a computer that is not completely determined by an algorithm, it could not be considered free because it would not be free in the same way as humans are.

> The problem with this approach is that although some computer systems may be nondeterministic and, therefore "free" in some sense, they are not free in the same way humans are. Perhaps it is more accurate to say that we have no way of knowing whether computers are or will be nondeterministic in the same way that humans are nondeterministic. We have no way of knowing whether the noumenal real of computer systems is or will be anything like the noumenal realm of humans. What we don't know is that both are embodied in different ways. Thus, we have no way of knowing whether the nondeterministic character of human behaviour and the nondeterministic behaviour of computer systems are or will be alike in the morally relevant (and admittedly mysterious) way. [180]

She bases this thesis on an argument very similar to the one presented by John Searle in *Minds, Brains, and Science*[181] . In the concluding part of his article, Searle denounces the undeclared, but widespread, presence of a residual dualism among A.I. researchers. In fact, if one wants to build a strong A.I. one has to consider that *where one has to deal with the mind, the brain has nothing to do with i*t[182].

The paradigm which Searle refers to is between mind-brain and program-hardware, to be able to believe that the constitution of a strong A.I. is possible, one must consider the two terms of both binomials as independent: therefore, if what happens in the mind has nothing to do with the interaction of the latter in the brain, then it will be possible to recreate that phenomenon through a program, since all the conditions will be re-creatable. This is the undeclared dualism that, according to Searle, characterizes the research of strong A.I., and yet it is impossible that *real human mental phenomena* do not depend, at least in part, on real *chemical-physical properties of human brains* [183]

Searle's and Johnson's argument suffers the same problem: reductionism. For Searle the faculty of thought is the result of the inexplicable but fundamental and necessary interaction between brain and mind, in the same way for Johnson human

[180] *Ivi*, cit. p. 176.

[181] J. R. Searle, *Minds, Brains, and Science,* Harvard University Press, Cambridge MA 1984.

[182] *Ivi*, cit. p. 70.

[183] *Ibidem*.

beings have free will and robots do not because they have two different types of bodies, *they are both embodied in different ways*.

So, for Johnson the intentionality of computers is connected to the intentionality of the *designers* who build and program it and to the intentionality of user who uses it. However, she does observe, that some computers have an implicit *intentionality, built-in intentionality*, of facts that when their processes are activated by a human intentional action, they continue to run autonomously, and their behaviour is independent from human ones. Although she attributes a certain degree of intentionality to computers, for her, this does not mean that they are moral agents: in fact, they are not free, therefore they do not have intentional mental states, therefore they cannot act with intention and for this reason they cannot be considered responsible for their actions.

> My argument is, then, that computer systems do not and cannot meet one of the key requirements of the traditional account of moral agency. Computer systems do not have mental states and even if states of a computer could be constructed as mental states, computer systems do not have intendings to act arising from their freedom. Thus, computer systems are not and can never be (autonomous, independent) moral agents. On the other hand, I have argued that computer systems have intentionality and because of this, they should not be dismissed from the realm of morality in the same way that natural objects are dismissed. [184]

In conclusion, there is no doubt that a computer is a moral entity involved in moral circumstances, but this does not mean that they are moral agents; to be moral agents they should be free, humanly autonomous, and have mental states.

It is so exhausting to defend the exceptionality of the human being and at the same time to develop a work that is honest both in practice and in fact. A human being is responsible if they are an agent with internal states, they are considered as mental and at least one of them must be intentional. An intentional state of mind is the reason for an action. This action is an action in the real world and is caused by the intentional state of mind, the effects of the action take place in the real world and are suffered by a patient. All this process presents itself in a similar way in computers except for the intentional mental states, which as has been said cannot present themselves in the computer because it is not free in the same way that a human is, and finally a computer cannot be held responsible for its actions, even if they are moral actions that create moral situations and interact with moral actions of other moral agents.

[184] D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. p. 182.

Johnson's argument focuses on terminological distinctions that are not demonstrated in the reality of the facts and that sometimes bend on themselves resulting in circular arguments. Think, for example, of when Johnson argues that computers are intentional but not intentional like humans: in them three different intentionalities intersect, of the computer itself-designer-user, but certain actions of the computer are taken in complete autonomy and are independent from human intervention and at the same time the computer cannot be considered responsible because it has no free intentional mental states. What is the solution to the problem of responsibility of artificial agents that are the source of an action that can be considered as a moral action? Think of BLU-108. Sometimes one gets the impression that some distinctions of terms are made simply to exempt the argument from a proper analysis of free will, both in humans, robots and A.I. [185]

2.9.3 Nadeau: robots are the only possible moral agents

Joseph Emile Nadeau in *Only Androids Can Be Ethical*[186], asks how a computer, which he calls *android*, should be made, both internally and externally, to be found guilty. To be guilty means not only to be the cause of certain events but to be responsible for them. The thesis that Nadeau argues and tries to substantiate is that computers are, and can be, the only responsible moral agents. [187]

His argument starts from the deconstruction of the relationship between free will and responsibility. Since, as was already said in Johnson, guilt depends on free will and one is responsible when one has been able to choose freely, starting the investigation precisely from the relationship between free will and responsibility would seem a good idea.

The general pattern followed by philosophers who argue for the existence of free will is that human beings have reasons, and it is by virtue of these reasons that they decide to do and act. Every choice is equivalent to a reason and it is for this reason that it is possible to identify the guilty party, i.e., the one who had the reasons and acted

[185] BLU-108 is described in chapter *2.7 Can a robot be considered a moral agent?,* pp. 89-93 of this work.

[186] J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248.

[187] Cfr. *Ivi,* p. 241.

because of them. However, modern cognitive science shows that the belief that there are always reasons for actions is a belief.

> Eminent American philosophers, Harry Frankfurt and Susan Wolf, for example purpose that a free action is done from reasons, and through an appropriate reasoning process. (I do not popularise: I have found that the essence of essays and books of analytic philosophy can be often putted in a sentence.) Following the influential work of Donald Davidson, they have no difficulty with the thought that reasons are often causes. Cognitive psychologists think this is rather naïve, and claim that human reasons are usually or often (psychologists are not strong on quantifiers) ex post facto confabulations. We are creatures of mental fog, of habit and passion, and only rarely and dimly creatures of reasons. [188]

Reasons are more explanations that are given *a posteriori* about events, so that they can be explained and made consistent. To regard reasons as confabulations created *ad hoc* for specific events has the theoretical implication that free will cannot be regarded as anything other than a fantasy. If a free action is that action caused by a process of reasoning, (reason as the cause of action) and humans, most of the time, do not act from reasons, then humans are not responsible.

A different case is that of *androids*. An android has a programme that causes its behaviour and (pre)determines every step in its decision-making process. The corroboration of this thesis would be found in the fact that if a human being were able to understand the trace of the programme, the human could very easily understand the android's motives, and likewise if the android were able to access the trace of its own programme, it could understand its own motives.

Obviously, the android would have a much better deliberative faculty than the human one, as it could connect reasons and consequent actions much more easily and linearly, it would not find itself confabulating afterwards about the reasons why it did this rather than that. The android has a programme, the programme sets out a series of actions, reactions, and counter-reactions.

If it is true that the Turing test makes it possible to identify a limit in the rational faculty of androids, it is also true that the same test corroborates the fact that androids have a rational faculty: an android reason and has reasons that cause its actions. Therefore, if an android reason and has reasons that cause specific actions, which can be traced back to specific reasons prior to the actions (and not confabulations given a posteriori), this means that androids can be held responsible and guilty for the actions they perform.

[188] *Ivi,* p. 242.

How is it possible for Nadeau to claim that the reasoning of androids is equivalent to that of humans? According to Nadeau, just as the equivalence of two program sequences of two different computers is given by the similarity of the program traces at the language level of the two machines, not by the mapping of the physical states of its components. The comparison between human reason and android reason must be made at the level of high-level language, that of data structures, and not at the biological level.

Obviously, a position like Nadeau's is easily attacked by all those authors who are against an analysis of the mental from the point of view of a computational theory, think, for instance, of the Searlian theory. As already mentioned in the previous paragraph, John Searle argues that the reason why it is impossible for an android to have human mental faculties, i.e., reason, intentionality, is that an android cannot have a mind and cannot have a mind because it does not have a brain. The interaction of the biological brain, with its tissues and chemical components creates the mental and its faculties, in an unknown chemical interaction. [189]

As it turned out, for Nadeau instead:

> We will understand a neural network as reasoning only if we understand the system to embody a compiler for a higher-level language in which reasons occur as data structures. Designs, albeit not perhaps very elegant ones, have been proposed for compiling intelligible finite state machines into neural networks, and we may at least reasonably conjecture, therefore, that it is possible for an android formed upon a neural network to reason and to have reasons. [190]

Thus, Nadeau supports his thesis. Responsibility and guilt require an action to be free, a free action is an action caused by reasons and, as has been shown, human beings do not have adequate mental capacity to have adequate reasons for the actions they perform.

Human beings tell themselves reasons *a posteriori*, after they have performed the action, so that they can reinsert what they are doing into their own scheme of life, so that they can explain themselves and answer their own whys. At the end of the day, however, these are not reasons, but more like delayed confabulations. As cognitive psychology of the last century has shown, human beings do not really seem to have the necessary preconditions for being able to have reasons. Human actions are not

[189] Cfr. *Ivi.*, p. 244.

[190] Cit. *Ivi*, p. 245.

caused by reasons, and so they are not free, and so human beings can neither be held responsible nor guilty.

On the other side are the androids, who have reasons and have adequate capabilities, advanced enough to have reasons that cause actions. Androids are guided by their own agenda and, if they could access that agenda, they could easily know their own reasons. Since androids have reasons and act on them, they can be held responsible and therefore guilty. Consequently, androids, not humans, would be considered ethical and moral entities.

2.9.4 The extension of the moral agency to artificial agents

Luciano Floridi in *On the Morality of Artificial Agents*[191] , defines a system that allows to defend the thesis that, since some artificial agents are the origin of moral and immoral actions, they should be included in the class of moral agents. A moral situation always presents an agent who performs a moral action and a patient, who is the recipient.

Five types of logical relationships are possible between agent and patient: they can be two completely separate entities; they can intersect with each other and the "patient" category can be considered as a subset of the "agent" category. However, the opposite is also possible, i.e., recognising the category of "agent" as subordinate to that of "patient" and having a situation in which everyone is a patient and among these patients there are agents. This is the paradigm of an environmental or *non-standard* theory. Finally, there is the so-called *standard* position that all entities that can qualify as moral agents can also be patients and vice versa.

While a lot of work has been done in recent years from the point of view of a *non-standard* approach, in fact more and more importance has been given to those bodies recognised as patients, think for example of the rights of Nature (i.e. the Paris Accords), from the agency's point of view no major changes have been made. It is true that the concept of "moral agent" has been extended to not only natural but also legal entities (i.e. companies), but it is always defined from an exclusively anthropocentric point of view. This approach is limiting because it prevents, or at least slows down,

[191] L. Floridi, *On the Morality of Artificial Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. pp. 184-212.

the investigation of new ethical situations that certain types of artificial agents put in place.

> An entity is still considered a moral agent only if: (i) it is an individual agent; and (ii) it is human-based, in the sense that it is either human or at least reducible to an identifiable aggregation of human beings, who remain the only morally responsible source of action, like ghosts in the legal machine. [...] Insisting on the necessarily *human-based nature* of such individual agents means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents that are sufficiently informed, "smart", autonomous, and able to perform morally relevant actions independently of humans who created them, causing "artificial good" and "artificial evil". [192]

Floridi identifies the problem in the defining procedure, in fact the definitions that exist of agency are all in function of the theories that they must support, this happens because the concepts with which they work are indefinable and in continuous change. To avoid this indetermination, one can take inspiration from mathematics, where definitions are proposed that function as a parameter for a certain set of data, the method is called "*Levels of Abstraction*". A level of abstraction is what qualifies the degree and the level, in which an element of a system, or an entire system, is considered.

In practice when a certain level becomes stable it can be formatted and consequently defined. Identifying a certain stable level is not an absolute but is always a contextual process, nevertheless, when the identification of a certain level becomes dominant it happens that it appears absolute, as transparent, and untouchable. Until now the concept of agency has been treated in this way, but since there are entities that deeply question the definitions that have been given, and the simple fact that there are discordant definitions between them, makes agency an absolute non-definable concept, a concept for which, before starting a research, it is necessary to define the level of abstraction and the system of reference.

So, what are the requirements to be considered an agent? The level of abstraction of agent that Floridi chooses includes three requirements: (i) interaction, (ii) autonomy (iii) adaptability. Interaction means that the agent and the environment interact in a feedback system, autonomy means that an agent is able to change its state, without this change being considered as feedback due to the interaction with the environment; while adaptability are those changes that the interaction with the environment brings

[192] *Ivi*, cit. p. 186.

to the autonomous transition processes. When some artificial systems are observed at this level they can be considered as artificial moral agents.

For example, *SmartPaint* is a new "intelligent" type of paint with electrical properties that allow it to change according to changes in the structure on which it is applied, so if a wall were to crack due to an imperceptible shock, SmartPaint would point out the need for maintenance.

> At a level of abstraction at which only the electrical properties of the paint over time is observed, the paint is neither interactive nor adaptive but appears to be autonomous; indeed, the properties change as a result of internal nondeterminism. However, if that level of abstraction is augmented by the structure data monitored by the paint over time, then SmartPaint becomes an agent, because the data provide input to which the paint adapts its state. Finally, if that level of abstraction is augmented further to include a model by which the paint works, changes in its electrical properties are revealed as being determined directly by input data, as so SmartPaint no longer forms an agent. [193]

Therefore, when you observe Smart Paint at a certain level of abstraction it can be treated as an agent and this also applies to web bots[194]. Yet having proved that artificial systems can be considered artificial agents is not enough to consider them moral agents.

As has been said before, an action is moral when it causes good or bad consequences and an agent is a moral agent when it is capable of producing morally qualifiable actions, therefore both human and artificial agents are qualifiable as moral agents.

There are four objections that are generally given to the possibility of extending the moral agency to artificial agents. The first one is the teleological one for which an artificial agent cannot be considered moral because it cannot really have goals; it is to be considered unfounded, in fact it is very easy to implement an artificial system because it includes *goal oriented behaviour.* Then there is the objection that to be a moral agent one must have intentionality, as seen in Johnson. According to Floridi, intentional states are not a necessary condition, in fact having specified a specific level of abstraction before starting the research guarantees that the analysis can be based exclusively on observable bases and not on psychological speculation.

> This phenomenological approach is a strength, not a weakness. It implies that agents (including human agents) should be evaluated as moral if they do play the "moral game".

[193] *See*, cit. pp. 197-198.

[194] The analysis of a webbot in different levels of abstraction can be found in L. Floridi, *On the Morality of Artificial Agents,* in *Machine Ethics,* Cambridge University Press, Cambridge 2011, cit. p. 197.

> Whether they mean to play It or they know that they are playing it, is relevant only at a second stage, when what we want to know is whether they are *morally responsible* or their moral actions. [195]

Thusly, intentionality is not important in the consideration of moral agency. However, the ability to be morally responsible could well be. Responsibility is a concept full of history and vicissitudes, it has always been linked to the idea of retributive justice for which it is necessary to punish those who have made mistakes, those who have acted "maliciously", in order to give back to the community what they have been deprived of.

Postponing an in-depth analysis of responsibility, for the moment it is sufficient to continue by saying that Floridi recognises the pivot on which the objections, that link the moral agency to the faculty of being responsible, are based. In fact, this type of objection implicitly maintains that one can be responsible if and only if one can be punished at a potential level and this is a "theoretical deception" because there is much to be said about the responsibility of an artificial agent when this is not merely considered as the possibility of making it suffer (i.e. reprogramming an artificial agent or cancelling it without making a backup in case it was incorrigible).

The last objectors might mention that an artificial agent is not free and decree freedom as a necessary characteristic to be a moral agent. Floridi reiterates, using the same paradigm:

> All one needs to do is to realize that the agents in question satisfy the usual practical counterfactual: they could have acted differently, and they could have chosen differently because the are interactive, informed, autonomous and adaptive. [196]

[195] *Ivi*, cit. p. 200.

[196] *Ivi*, cit. p. 201.

Third Section: Artificial agents and ontological asymmetries

3.1 Ontological symmetries and asymmetries

In chapter *2.7 What is Moral Agency and why is it important: theory of mind,* the concepts of agent, agency and moral agency were defined. It was said that an agent is an entity whose behaviour produces an effect, agency is the manifestation of that capacity, and, consequently, a moral agent is an entity whose behaviour produces morally qualifiable effects. It was concluded by noting that on a theoretical level both humans and robots could be considered as agents, and both human agents and artificial agents were also moral agents.

Although the positions held by the authors differ, the arguments generally revolve around the same five concepts: autonomy, freedom, intentionality, consciousness, and responsibility. They are intersected and implicated in each other and, depending on the theories, are found in both human beings and A.I.[197]., or are ultimately not considered necessary in the qualification of moral agency[198].

Reading the authors[199], it is difficult agree or disagree with any of their arguments, because the examples that are given to support a certain thesis are somehow always *ad hoc* and undermine the process of discernment. How do these faculties manifest themselves in the human condition and in the robotic condition? What are the ontological differences that emerge? For this reason, it will be looked at how the five

[197] Cfr D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365; J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248;

[198] Cfr. L. Floridi, *On the Morality of Artificial Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. pp. 184-212.

[199] Cfr. S. Bringsjord, *Ethical Robots: The Future Can Heed Us*, in *AI and Society* 22, 2008, pp. 539-550, (https://doi.org/10.1007/s00146-007-0090-9), 18th November 2020; D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365; J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248; L. Floridi, *On the Morality of Artificial Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. pp. 184-212; D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 168-183.

requirements: autonomy, freedom, intentionality, consciousness, and responsibility manifest themselves in the robotic condition, so that it can be understood whether certain A.I. can be considered, first of all, as artificial agents, and only afterwards as moral artificial agents.[200]

Kenneth Einar Himma in *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*[201] aims to prove the thesis that each of the elements necessary to be considered moral agents presupposes consciousness, defining consciousness as the capacity for subjective internal experience. To be an "agent" is to be capable of performing an action, but attention must be paid to the fact that not every act is an action. In fact, Himma considers an action as a mental state. Therefore, agency is the capacity to instantiate certain intentional mental states. Regarding the definition of moral agency, Himma uses the standard account which was previously introduced and analysed.[202] The necessary and sufficient conditions for moral agency are the freedom of choice and the ability to discern right from wrong. The first requirement is quite easy to comprehend. Through a deliberative process comes the possibility of choosing between two or more outcomes because of an action taken; this possibility opens the space for a behaviour that could be interpreted as free.

> Someone who acts on the basis of some unthinking compulsion is not making decisions, deliberating acting rationally or freely choosing her behaviours. Insofar as one must reason to deliberate, one must have the capacity to reason and hence be rational to deliberate.[203]

The second capacity, moral reasoning i.e., discerning from right and wrong, is also related to rationality. In fact, moral reasoning requires, first, an adequate understanding of concepts that happen to be moral and second, the ability to choose in a context of

[200] As it was said, an agent, by definition, is an entity that acts and has power, something that produces or is capable of producing an effect: an efficient or active cause. Therefore, an agency is the manifestation of this capacity. A moral action is an action whose effects, whose impact on reality, can be considered right or wrong, good or bad; hence moral theories are the theories of correct action, they indicate those principles that should guide action.

[201] K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*, in *Ethics and Information Technology (2009) 11*, pp. 19-29.

[202] The *standard account* is explained in *2.7.2 Robots will never be moral agents*, pp. 93-98 of the present work.

[203] K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*, in *Ethics and Information Technology (2009) 11*, cit. p.23.

multiple and contrasting values expressly which is in other words, identifying what makes a rule or moral value relevant in a specific context. The final step is correctly applying the concepts and relevant context to real world situations.

For example, the warning "the speed of the car should decrease with fog and rain". (i) It has to be understood by the driver (i.e., the speed has to decrease because with fog and rain my eyesight is worst and I need more time to react). Furthermore, (ii) the situation has to be recognized (i.e., not every context is dangerous, sometimes is better to keep the speed constant to prevent crashes with the car behind). Finally, (iii) the moral value of the decision (i.e., I do so because I do not want to harm a pedestrian).

At the end of the argumentation Himma comes to prove his thesis that the standard account of moral agency pertains only and solely to conscious beings, since consciousness is needed for his definition of action being an intentional state and the fact that punishment is applicable only to conscious beings because of suffering, both physical and psychological. This is further backed up by Himma's account on the two requirements on being a moral agent, freedom and moral reasoning, since he underlines their foundation in rational deliberation and this capacity presupposes conscious reasoning. Therefore, A.I. does not have the means to be either an artificial agent nor an artificial moral agent.

> I have concluded that while there are difficult epistemic issues involved in determining whether an artificial ICT us conscious and a moral agent, it is a necessary condition for an artificial ICT to be a moral agent that it is conscious. I have not however, drawn any conclusion about how artificial agents that appear conscious (though the appearance is not enough to warrant believing they are conscious) should be treated. [204]

In Himma's argumentation there are several unclear discrepancies given mostly by his foundation in the standard account which was already criticised previously in this work.[205] Generally, this criticism can also be applied to Himma's account i.e., action is equivalent to intentional mental states, which is the beginning of a chain of grounded explanations that precludes the possibility of considering entities as intentional. Nonetheless, it does so by starting from defining action in a biased and predisposed way.

[204] *Ivi*, cit. p.29.

[205] The *standard account* is criticized in *2.7.2 Robots will never be moral agents*, pp. 93-98 of the present work.

In addition, there is the fact that Himma does not analyse the connection between rationality, language and thus consciousness even though he briefly mentions it but does not go into detail over it. Analysing it may not completely undermine his thesis but would for sure clear up certain aspects of rational deliberation which he takes for granted.

What is even more striking is Himmas detachment from real-world examples with his definition of a moral agents being that of agents that know what is right from wrong and only those who know this are considered responsible for their actions and are thusly punished. This account flies in the face of the fact that prisons are full of psychopaths and sociopaths who are, by definition, completely devoid of moral reasoning. As it is not a novelty that organised crime syndicates such as the mafia, have a subculture where the concepts of good vs. evil are polarised in respects to the normalized status quo.

The more likely reality may be that the judicial system works on both the ontological and consequentialist accounts. This is shown by the fact that yes, people do need requirements to be considered capable of their actions and therefore responsible, but at the same time, whatever the intentions were or whatever kind of moral reasoning there was, either present or not, dangerous individuals are removed from society because their actions harm society.

Himma in one passage of his paper, resonates with the concept that punishment[206] is effective only when there is consciousness. *Praise, reward, censorship, and punishment are rational responses only to beings capable of experiencing conscious states like pride and shame*.[207] He presupposes a concept of accountability intrinsically tied to punishment, and because of things like fear and suffering, accountability itself derives from it. All this without taking into consideration the fact that there are different methods that are starting to gain traction in detention facilities around the world that do not involve punishment. Advocates are expressing more and more, the importance of re-educating the convicted in order to reinsert them into society in

[206] The concept of punishment (and its existing or non-existing link with consciousness) is going to be reopened in the concept of reinforcement learning as a machine learning method. The analysis can be found in paragraph *3.3.2 Deep Learning and Alpha Go, building a neural network from scratch*, pp. 123-127 of the present work; but also, in paragraph *2.4.2 Deep Learning: neural networks, big data and genetic algorithms*, pp. 70-71.

[207] *Ivi*, cit. pp. 24-25,

opposition to taking revenge on them via means of the very same paradigm that they inappropriately used against society.[208]

This has interesting implications, around A.I., as it is not required to have consciousness in order to be re-educated. If this method of re-education is not satisfying enough, keep in mind that methods of rewards and punishment were theorised and used since the very early days of A.I. (Dartmouth) as it was proved to be quite optimal as a learning method. Applying these methods of punishment to A.I. does not involve the same ethical issues as it does with living beings who suffer. In fact, in the end, the result of reward and punishment in A.I. is actually a form of education.

Daniel Dennett is of a totally different viewpoint compared to Himma. In *Kinds of Minds*[209] Dennett defines agents as those entities who have enough complexity to act. Even though Dennett did not set out to define agents and agency in this work but instead wanted to understand what and who has a mind and what exactly a mind is, he did investigate the concept of agency. He does it by considering the evolutive pathway.

> Through the microscope of molecular biology, we get to witness the birth of agency, in the first macromolecules that have enough complexity to perform actions instead of just lying there having effects. Their agency is not fully fledged agency like ours. They know not what they do. We, in contrast, often know full well what we do. At our best – and at our worst – we human agents can perform intentional actions, after having deliberated consciously about the reasons for and against. Macromolecular agency is different; there are reasons for what macromolecules do, but the macromolecules are unaware of those reasons. Their sort of agency is nevertheless the only possible ground from which the seeds of our kind of agency could grow.[210]

Humans are made of robots as Dennett says. Frequently the distinction between humankind and all other entities is strongly underlined as if putting the two on the same pedestal would shine a poor light on humanity. Thus, when these other entities are observed with a neutral eye, even the simplest of them i.e., macromolecules operate as if they have a mind. In fact, a macromolecule in a body needs to have certain receptors, as simple as they might be, that connect them to the environment in which they are, in order for them to act and react to changes in this environment. Dennett

[208] Cfr. M. Beraha, *Il Libero Arbitrio fra Natura ed Etica: un concetto da superare*, Manoscritto inedito, Milano 2020, pp. 133-163.

[209] D. Dennett, *Kinds of Mind, Towards an Understanding of Consciousness,* Basic Books Harper Collins Publishers, New York 1996.

[210] *Ivi*, cit. pp. 20-21.

pictures them, the simplest minds, as switches that turn on and off to underline the intentionality behind any of the actions produced by them.

How can macromolecules have intentions? An intentional system is a system that acts on the basis of information and goals. Therefore, an intentional action, is an action that has a goal and is taken because of certain information. Dennett calls this "intentional stance" and it pertains to the interpretation of an entities behaviour as if that entity were a rational, intentional, free being. This is exactly what humans do with other humans and entities do with other entities that they are attached to and it is what was previously described in this work as "theory of mind". In other words, a robot should be treated as it was an intentional agent to comprehend and grasp its intentions.

> Intentional systems are, by definition, all and only those entities whose behaviour is predictable/explicable from the intentional stance. Self-replicating macromolecules, thermostats, amoebas, plants, rats, bats, people, and chess-playing computers are all intentional systems - some much more interesting than others […] I call such systems intentional systems. They exhibit what philosophers call intentionality.[211]

Dennett's discourse continues towards an analysis of body and mind in order to finalise his works ultimate goal. For the matter of this subject thesis, the focus is on what was said about agency and intentionality, in fact it might sounded too broad, all sizzle and no steak. Clearly to support them, it has to be understood what the result of the intentional stance is but the same goes for the standard account and its tightness. What are the effects of a consideration of agency that is so straight and strict? Since considering Himma's and the standard account's definition of agency completely occludes and predisposes, the opportunity for an investigation into artificial agency and thus, artificial moral agents (also adding to the fact that it is not fully backed up by real world scenarios i.e., psychopaths), Dennett's account now becomes the supporter of this possibility. Nevertheless, even when considering an agent as a being having enough complexity to act, it is fundamental in the case of this work to be sure that this agent, that has enough complexity to act, can thus act autonomously. Since the goal is understanding whether or not it can be held responsible for its actions.

It seems so, that a perfect place to start is by analysing the concept of autonomy, since firstly, it must be understood whether robots can be considered as agents, and an agent is an entity whose behaviour produces an effect. There must therefore be an

[211] *Ivi*, cit. p. 34.

identifiable origin, who caused X? The origin of X is an agent that is assumed, to be autonomous, the original beginning of that behaviour and the end of the causal chain that is reconstructed to identify the cause of certain effects. In fact, following the works of the authors[212] the minimum requirement to be considered an agent is to be autonomous, all the other requirements are necessary to be considered a moral agent. Therefore, if autonomy is found in certain A.I, it can be considered as an artificial agent.

3.2 Autonomy

In Greek, "*autos*" means "self" and "*nomos*" means "the law"; etymologically, autonomy can be defined as 'to exist by one's own law'; the term was used in reference to the state of a *polis.* This could be defined as autonomous if its citizens were the authors of the laws in force, so in some ways the concept naturally extends to the individual, autonomous when he or she is self-determining, making decisions and taking actions for themselves.

"Autonomy' in its past meaning is a clearly recognisable qualification, it is a concept that was coined for a specific reason, namely that of having the possibility to decide one's own laws. The context referred to is one in which there are citizens, free humans, and slaves. Then other meanings were given to the concept of autonomy, different 'freedoms from constraints' were understood, making them more minute, almost imperceptible. The characteristics of an autonomous individual have become other than those of a free citizen, free in the sense of not being a slave-woman-child and in the sense of being free from usurpers.

At a glance, the modern definition of autonomy is not so different but is more subtle. It has always meant the ability of a person to live his or her life according to

[212] Cfr. S. Bringsjord, *Ethical Robots: The Future Can Heed Us*, in *AI and Society* 22, 2008, pp. 539-550, (https://doi.org/10.1007/s00146-007-0090-9), 18th November 2020; D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365; J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248; L. Floridi, *On the Morality of Artificial Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, cit. pp. 184-212; D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 168-183.

his or her own beliefs and motives, which must be recognised as his or her own, completely independent, and not the result of external forces or coercion. In order to be able to act on the basis of one's own desires and values, it is necessary to have the following faculties: the ability to think rationally, self-control and not to be affected by certain debilitating diseases. There is also another ability, which is more difficult to identify, that of being able to reflect on and identify with these desires, values, and beliefs.

This concept of autonomy, less political and more analytical, calls into question faculties which open up a whole series of problems, one of which is: how does one qualify a specific type of thought, rational thought? Why is it necessary to think rationally in order to be autonomous? Think of the 'intelligence of emotions', emotions can easily be attributed to a specific and identified subject, the subject can easily identify with them and they are a proven motive for action. So why is emotional thinking not necessary, or at least sought after, in the determination of autonomy?

Perhaps one of the possible answers, which is also an indication of the absoluteness that is sought in the modern concept of autonomy, is that emotions are often provoked by something external, by the multiple relationships that the individual has with the environment. Emotions show the *individual dividual*. Does autonomy really mean being absolute individuals immune to any external inference? How is an autonomous decision determined? Is being autonomous or not something that can be assessed from outside?

Gerard Dworkin, in *The Theory and Practice of Autonomy* [213] writes:

> Given various problems that may be clarified or resolved with the aid of a concept of autonomy, how may we most usefully characterize the concept? I use the vague term "characterize" rather then "define" or "analyse" because I do not think it possible with any moderately complex philosophical concept to specify necessary and sufficient conditions without draining the concept of the very complexity that enables I to perform its theoretical role. Autonomy is a term of art introduced by a theorist in an attempt to make sense of a tangled net of intuitions, conceptual and empirical issues, and normative claims. What one needs, therefore is a study of how the term is connected with other notions, what role it plains in justifying various normative claims, how the notion is supposed to ground ascription of value and so on - in short, a theory. [214]

[213] G. Dworkin, *The theory and Practice of Autonomy*, Cambridge University Press, Cambridge 1988.

[214] *Ivi*, cit. p. 7.

"Autonomy” is a term whose origin gives it a tangible and concrete meaning, but over time it has been used to give meaning to much more complex, internal, and hidden human events. The concept of autonomy can be understood in two ways[215], the first being: an entity is autonomous if it is not obviously forced by external circumstances to act in a certain way, i.e. if it is not extremely emotionally or physically constrained; while the second is the capacity a person has to live his or her life according to his or her own beliefs and motives, the latter being recognised as his or her own, completely independent, and not the result of external forces or coercion.

The second conceptualisation implies the presence of, at least, two other concepts: (i) internal states and (ii) the ability to reflect or self-consciousness. Since, in order to compare human and robotic autonomy, it is necessary to work with an independent concept separate from other capabilities and characteristics, the meaning given by the first definition will be used, an entity is autonomous if it is not obviously forced by external circumstances to act in a certain way. To this reason, another is added: understanding autonomy as the capacity of a person to live his or her life according to his or her own beliefs and motives, which must be recognised as his or her own, completely independent, and not the result of external forces or coercion, does not allow any kind of concrete and phenomenological verification. Autonomy therefore means the possibility for an entity not to be clearly controlled in its actions by binding and coercive external forces.

Can a robot be considered autonomous? If some robots are shown to have a certain degree of autonomy, they can be considered as artificial agents. As discussed earlier in this research work, in the field of A.I. a robot is considered autonomous when it is the main actor in its decision-making process, when it decides through its programs[216] what actions to take. Thus, a robot has the possibility of not being overtly controlled in its actions by the *user*. For example, the previously mentioned BLU-108 warhead, an autonomous and 'intelligent' warhead, which uses a double system of sensors to detect the presence of targets, shooting an explosive projectile when the target is found.

[215] Of course, there are more than two ways of conceptualising the term autonomy, but the meanings mentioned are the fundamental and topical ones.

[216] A software program is commonly defined as a set of instructions, or a set of modules or procedures, that allow for a certain type of computable operation.

BLU-108 is an example of an autonomous war weapon that is not controlled by any human in the choice of its targets.

There is autonomy in robots, so they can be considered artificial agents. The manifestations of freedom, intentionality, consciousness/self-consciousness, and responsibility in artificial intelligences will now have to be observed in order to decide whether on an ontological level some artificial agents could be considered moral agents.

3.3 Could Artificial Agents be considered has having free will?

In this chapter will be covered the dispute over free will, one of the oldest and boldest philosophical questions. Free will can be broadly defined as 'the ability to have control over one's actions' and with it emerges the ability to foresee possible future scenarios and freely choose between them. Choice, in this context, precisely means the ability to select between different possibilities.

However, from the very beginning of this debate, free will has been heavily criticised. Indeed, libertarianism, the modern interpretation of free will, is deconstructed by the experimental method: scientific mechanisms and instruments desecrate, myth after myth, natural phenomena.

After a brief analysis of libertarianism, it must also be addressed its alternative, the deterministic theory in its various forms. This urgency emerges given that A.I. are generally programmed with deterministic algorithms[217], or so it is assumed. It is implied that at every juncture of an A.I.'s decision-making process there is no choice, only predetermination implicit in the programs code, the A.I.'s 'soul'. When one observes the A.I. in its behaviour, it seems to 'make choices' and 'have free will', but this is pure simulation. It might be interesting to ask whether it is not the same for humans and whether, in a deterministic process, there is room for a different kind of choice.

The following paragraphs will show the implications of libertarianism and determinism from a theoretical point of view, and then move on to an in-depth analysis of an application case in which the algorithms are not programmed ad *hoc* but the

[217] An algorithm is defined as deterministic when given a particular input, will always produce the same output, with the underlying machine always passing through the same sequence of states.

behaviour of the artificial agent is guided by a specific objective. This will be followed by the illustration of two specific types of algorithms that are used when the environment is partially observable or non-deterministic. It will conclude with a mention of quantum computation as it could be the proof of the existence of free will for an artifice.

3.3.1 Theoretical framework: libertarianism and determinism

The history of free will runs parallel to that of moral responsibility. This can be deduced from the route taken so far. Responsibility, which will be dealt with specifically later, is a situation in which a person assumes, or is given full responsibility for, the consequences of a certain action.

Moral responsibility in this sense, which is also the sense shared by jurisprudence, implies the presence of a subject who is free and aware of the effects (consequences) of his acts. Awareness is not only from the point of view of the laws of physics i.e. if I pull the trigger of a gun the gun fires and if the bullet fired from the chamber hits someone (for instance, a living being) it wounds him because the material and the force of the bullet fired against the material that makes up the living being has this kind of interaction. Yet, also from the normative point of view i.e., if I wound someone the controlled effect, as it was defined in the first section, will tend to be that of a trial and punishment.

For there to be moral responsibility, there must be a free, informed, and aware subject. Free because he or she must have understood what he or she has done, must have wanted it in some way; informed because he or she must be aware of the many physical, cultural, legal, and also relational facts; and, lastly, aware because he or she has been educated about his or her deviance from the norm and from normalised and accepted behaviour. Yet, as intricate as free will and responsibility are, they need to be separated in order to explore them. Thus, this paragraph will begin with 'free will'.

There are two major factions regarding free will: philosophers who promote the existence of free will and those who deny it. Libertarian theories regard free will as indispensable for the construction of an anthropology and ethical theory, rather than for the construction of an ontological system that justifies its existence. Free will is neither positive freedom, i.e., the ability to self-determine, nor negative freedom, that situation in which an action is not constrained by external factors. Positive and negative freedom make up what was identified as the analytical definition of autonomy, as opposed to the more political, Greek definition, which corresponds to the possibility of determining oneself, as a state, through one's own laws.[218]

Free will is distinguished from the latter by its two fundamental principles: (i) the principle of self-determination, which is the opening up of that possibility whereby a certain action can be completely attributed to a certain identified agent; and (ii) the principle of alternative possibilities: if exactly the same conditions were repeated, the subject, being free, could have behaved otherwise.

In conclusion, another hypothesis shared by libertarians is that, if one did not believe in the existence of free will, it would not make sense to speak of moral or immoral acts; it would therefore not matter to determine an ethical-moral system that polarises good-evil and right-wrong. Note (i) that this is only true for a certain type of moral theory, the deontological one; (ii) it only makes sense when one conceives of and supports a particular relationship between free will, responsibility and justice; and (iii) when one considers justice as a specific type of justice: retributive justice.

On the opposite side are the determinists. For a broad definition of determinism, one can refer to the fact that determinists refuse to believe in the existence of free will, considering as an impossible fact the case in which a human being could be the only existing entity capable of ignoring the laws of physics. The human being, in fact, in order to be free, would have to surpass the fundamental laws of classical Newtonian mechanics and act according to its own internal, independent motions.

Determinists reinsert the human being within the natural physical world of which he is a part of and reconstitute him as a being who, because he is composed of atoms, responds to physical laws: just like any other natural, but also artificial, entity. Determinism is at the same time different from fatalism.

Greek fatalism and the idea that there is a destiny already established and defined. The Fates weave the great tapestry of humanity's destinies and the meeting points of the different coloured threads, of the different lives, are already decided a priori. Arranged as if stretched from the bottom because the end is waiting patiently to be reached and no one can escape their destiny. Destiny attracts and is inexorably

[218] This subject matter was analysed in chapter *3.2 Autonomy*, pp. 110-112 of the present work.

fulfilled. Determinism, on the contrary, recognises the causal force that pushes forward: since x, y, z then A.

Depending on the author referred to, there are different ways of balancing the determined and mechanical forces of nature with the anomaly of a person determining its own free will. Although debates on the notion of 'free will' have a very ancient origin e.g. Epictetus, Stoicism, Epicureanism, Scepticism, Plato and Aristotle wrote about it, it can be said that free will became such, a problem to be solved, during Christianity: when it became necessary to reconcile God's will and power over the world and the human being who had to be praiseworthy or punishable in view of the afterlife.

Thus, in the history of medieval philosophy one can follow the story of Augustine, Thomas Aquinas, Abelard, and Duns Scotus (along with many others) in balancing the will of God and the freedom of sinful man. The debate remains central in modernity and it is during this period that fundamental theoretical pillars are established.

Without wishing to oversimplify and homogenise the theories and analyses of the various authors, but bearing in mind the purpose of this thesis, three statements generally shared by modern libertarian philosophers can be identified: (i) free will consists of two aspects: (i) the freedom to act otherwise given the exact same situation and the power to self-determine, both of which have already been mentioned above; (ii) the second assertion, also heralded, states that an adequate theory of free will must consider the presence of free agents who are morally responsible and potentially subject to punishment (by which, of course, it is meant that subjects can experience emotions such as pride, shame and that they suffer). Lastly (iii) and the compatibility between human free will and the determinist theory regulating reality, so-called compatibilism.

How is it possible to defend a position like the compatibilist one? There are two tactics. The determinist position can be downplayed by declaring that the opposite of freedom is not determinism but the external constraint that does not allow the subject to do what he wants. Or one can argue against such a tactic by declaring the need to analyse a type of ability deeper than the mere absence of limitations imposed on the individual will, Immanuel Kant being one of the philosophers who articulates this approach.

Free will requires more than free action; it requires the agent to be the true cause of that particular action. Supporting this kind of thesis not only implies the compatibility of free will and determinism, but necessarily claims the existence of determinist theories: the agent is the determinant cause of an action. The human agent is part of those natural physical forces and is an efficient cause with them. Many twentieth-century positions are based on this thesis.

The agent is regarded as free and morally responsible for his actions because he is the determining cause of a given act, a self-determined source. Defining an agent as a self-determined cause implies that the agent is the beginning of the causal chain, a cause not determined in turn by another previous cause.

There are three ways of justifying and promoting the thesis that the agent is the root cause of an action: (i) *non-causal libertarianism* whereby the exercise of the power of self-determination need not be causally structured. A is the cause of action X because he performed it and this is based. A position not very strong and easily attacked by determinists; (ii) for *event-causal libertarianism* the cause of an action must be entirely traceable to the mental states that caused that action. When he wants to call B he shouts 'B', this action is entirely reducible to a mental state, i.e. I want to call B, and is *nondeviant.*

> Non deviance clause is required since it seems possible that an event be brought about by one's desires and beliefs and yet not be self-determined, or even an action for that matter, due to the unusual causal path leading from the desires and beliefs of action. Imagine a would-be accomplice of an assassin believes that his dropping his cigarette is the signal for the assassin to shoot his intended victim and he desires to drop his cigarette and yet his belief and desire so unnerve him that he accidentally drops his cigarette. While the event of dropping the cigarette is caused by a relevant desire and belief it does not seem to be self-determined and perhaps is not even an action. [219]

Obviously, this implies that mental acts as such, i.e., ascribable to a suitable mental state, are specified, as are the mental states mentioned in the example above.[220]. In other words, for an *event-causal* libertarian theory: the agent's self-determination requires a non-deterministic and non-deviant causation, which can be *fully* traced back

[219] T. O'Connor, and C. Franklin, *Free Will,* in *The Stanford Encyclopaedia of Philosophy,* E. N. Zalta (Ed.), cit. p.19, (https://plato.stanford.edu/archives/fall2020/entries/freewill/), 10th January 2021; Cfr. D. Davidson, *Freedom to Act*, in *Essays of Freedom of Action*, ed. T. Honderich, Routledge and Kagan Press, New York 1973.

[220] Cfr. M. Brand, *The Fundamental Question in Action Theory*, in *Noûs* 13, 1979, pp. 131-151; Cfr. J. Bishop, *Natural Agency: An Essay on the Causal Theory of Action*, Cambridge University Press, New York 1989.

to the agent's reasons. How is it possible for a cause to be non-deterministic? This question shall be returned to later. It is now necessary to introduce the last strategy to justify the thesis that the agent is the deterministic cause of an action.

(iii) *Agent-causal libertarianism* holds that self-determination requires an agent who is not only the cause of the action but also of his or her own motives. Roderick Chisholm is one of the proponents of the *agent-causation* theory. It is interesting, for the continuation of the present research work, to pause on Chisholm's work, for in his theory emerges the close, perhaps fundamental, link between freedom, conscience, and responsibility. This close link will be introduced as far as Chisholm's theory of *agent-causation is* concerned but keeping in mind that the relationship between freedom, consciousness, and responsibility will be explored further in later dedicated chapters.

His 1964 essay entitled *Human Freedom and the Self*[221] opens with a quotation from that passage in Aristotle's *Physics* in which there is a stick moving a stone, which in turn is moved by a hand, animated by a human. Chisholm revolves his entire theory around the necessary fact that the agent must be held responsible:

> The metaphysical problem of human freedom might be summarized in the following way: human beings are responsible agents; but this fact appears to conflict with a deterministic view of human action (that view that every event that is involved in an acted is caused by some other event). [222]

Both determinism and indeterminism (referring both to *non-causal libertarianism* but also to *event-causal libertarianism*) do not explain human action correctly, an alternative must be found. In order to prove the freedom of human beings a system is constructed in which two different causations are distinguished (i) the type of causation of agents and (ii) that causation which operates on matter and events.

The causal chain inherent in agents does not consist of events i.e., event causes event, but agent causes event, where the agent is not caused by anything else i.e., it is a person who moves his hand in the Aristotelian scenario. Thus, the agent is the last link in the causal chain and is therefore responsible for his actions. Unfortunately, Chisholm does not explain the mechanism in detail and preserves the idea that intuition is the ultimate source of action, primitive and unanalysable.

[221]. R. M. Chisholm, *Human Freedom and the Self*, in *The Lindley Lecture*, University of Kansas, 1964, (https://kuscholarworks.ku.edu/handle/1808/12380), 13th January 2021.

[222] *Ivi,* cit. p. 3.

It remains to be seen why Chisholm included the word "Self" in the title of the article, the suspicion being that the "Self", consciousness, the "I" are fundamental to the system of *agent-causation* and *pro-free* will theories in general. This point will be returned to; it is now time to observe how a possible state of freedom emerges in some artificial intelligences.

3.3.2 Deep Learning e AlphaGo, building a neural network from scratch

As was previously introduced in paragraph *2.4.2 Deep Learning: neural networks, big data and genetic algorithms*, before 2016, the artificial intelligence team DeepMind worked on and programmed *AlphaGo* a computer program that learned to play the popular Chinese boardgame *Go* on its own without being explicitly designed to do so. DeepMind did not write a single algorithm specialised in Go (in contrast to DeepBlue which was comprised solely of brute force code designed to play chess and chess alone). They programmed instead self-learning algorithms that became able to learn by their own mistakes as well as the mistakes of their human competitors. This is known as reinforcement learning.

However, how is it possible for a machine to achieve the desired goal without being programmed to do so? From 1959, the first algorithms were theorised which would enable the birth of *machine learning*. This paved the way for the subsequent emergence of a plethora of machine learning algorithms which have been developed since then. These algorithms were called machine learning algorithms as they did not require the machine to be programmed with specific instructions but enabled it to program itself. The most common approach to machine learning is supervised learning where a reference guide points out errors. Imagine a wooden puzzle, those in which there are some shapes to be inserted in a wooden containing board, the machine has to learn the game, at the beginning it does not know how it works, it tries to insert the pieces in the holes at random and by doing so it understands what the useful information is. In short, there is a data point that indicates a desired outcome good or bad. The algorithm will learn from the other descriptors (attributes/features) and the values within these descriptors. The machine asks itself what pattern of features and values indicate a particular outcome – what pattern indicates a good outcome and what pattern indicates a bad outcome. On the other hand, in unsupervised learning, the basis of reference is removed, as if in the example above, the containing wooden board were removed and only the shapes to be learned and recognised were left. *Reinforcement learning*, already theorised by Turing, is the method of rewards and punishments that is adopted to emphasise which information is useful and which is not.

Deep learning is a subclass of machine learning algorithms and allows computational models composed of multiple levels to learn representations. Take a case of application of this method, take the case of AlphaGo where the spectacularity of the project goes beyond the success and speed of improvement of the program. The possible positions in Go are more than the atoms in the universe, this means trying to prevent or predict those sequences quickly enough, is quite literally impossible. Usually to implement logical reasoning, Go players base their tactics on subconscious intuitions, and since neural networks[223] are able to recognise images and portraits, they can also discern positions in Go.

The idea on which AlphaGo is built is the combination of deep learning with the logical potential of a machine and the huge database containing massive amounts of big data of all possible Go positions[224]. The marriage of data and logic has created successful moves that until that point were alien to the most experienced players: moves that defied millennia of human intuition and could easily be understood as creative. The results of 'deep learning' often refer to human creativity and intuition; whether this reading is appropriate or not will be understood later, now it needs to be understood how it works and further still how it is even possible.

As already defined Deep learning is a subclass of machine learning algorithms that allow computational models composed of multiple levels to learn representations. It is a method where multiple layers are used to progressively build higher-level features from raw *input*, each layer learning to transform the collected data into more elaborate, abstract, and composite representations. It is based on artificial neurons, mechanisms from which numbers enter and exit (input-output), which compose a neural network characterised by three types of layers: (i) the *input layer* that takes the data; (ii) the *output layer* responsible for presenting the results; and in the middle of these two (iii)

[223] Neural networks are computational components inspired by the neural connectors of the human brain. This subject matter is discussed in paragraph *2.4.2 Deep Learning: neural networks, big data and genetic algorithms*, pp.70-78, of the present work.

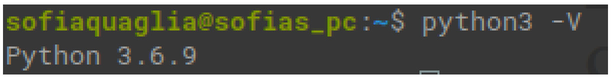
[224] Big Data is the term that refers to a computers ability to process and analyse a huge amount of data. This subject matter is discussed in paragraph *2.4.2 Deep Learning: neural networks, big data and genetic algorithms*, pp.70-78, of the present work.

there are multiple and countless unseen layers, the so-called *hidden layer*. For the time being, it is enough to know that the output, the final result, is dictated by the hidden layer and it can be difficult to control and understand how these layers interact with one another, if the output were to appear bizarre one could not walk back through the steps taken, because they are quantitatively too many to be processed by a human. AlphaGo's moves are unexpected and perceived as intuitive and creative for this reason.

Following on from the work done in this section, the next logical step is to dive into the more technical aspects surrounding Deep Learning. The easiest way to do so is to take a look at trying to build a simple neural network and this can be achieved by using Python to build a network from scratch.

The first step is to understand all the parts that would be involved in the coding of an artificial intelligence neural network. First the programming language is chosen, this is going to be Python. Therefore, the latest release of Python and its libraries need to be downloaded and installed onto a machine for the purpose of writing and executing code. The downloads of which can be easily found at www.python.org/downloads.

Once installed, the version of python can be checked to ensure proper installation by executing a simple terminal command as seen below (having Python installed in the $PATH of the machine allows for the execution of Python programs to be executed in the terminal):



```
sofiaquaglia@sofias_pc:~$ python3 -V
Python 3.6.9
```

Fig 1.1 Python Version Check

Knowing that Python is now installed correctly the next step is to understand what exactly a neural network is and how it functions, beyond the knowledge already presented in this thesis. For this a note made on a lecture given by Professor Mario Molinara (Università degli Studi di Cassino e del Lazio Meridionale) can be referenced. From this the definition of a neural network is described as such:

A biological neuron is very different from an artificial one, in fact an artificial neuron is a mechanism in which numbers are inputted and then in turn new numbers are provided as output. It represents the feature as a number in the form of input and multiplies those numbers with an associated coefficient. These results are added together, and this result is called the *coefficient of bias*, which is then projected onto the *function of activation* which, in this current input and all following input, will produce a varied output.

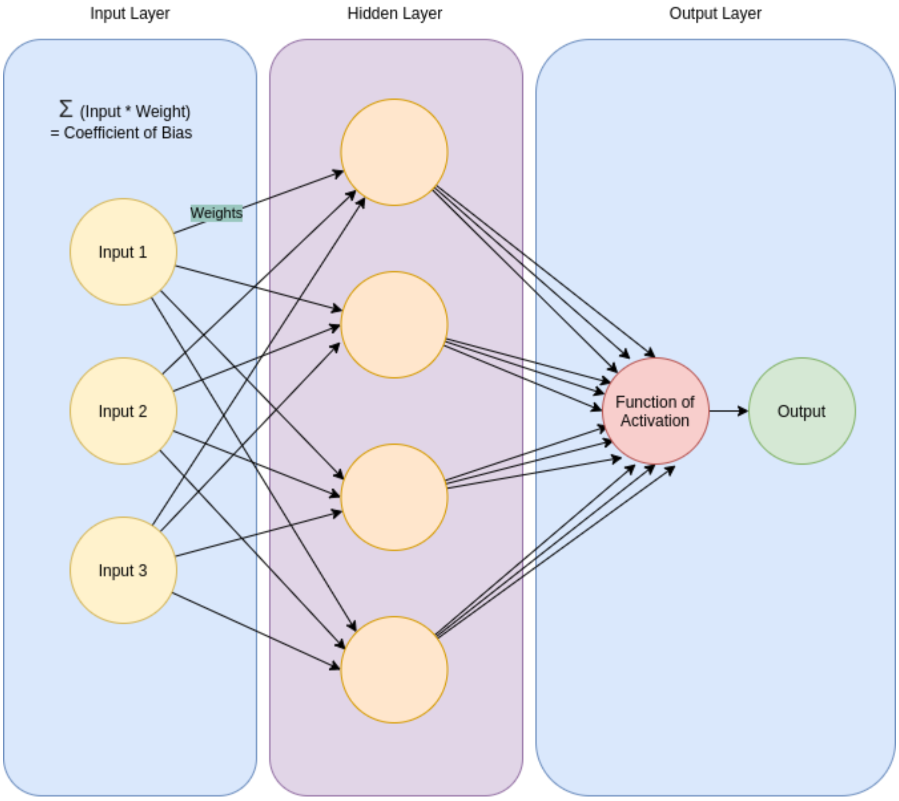


Fig 1.2 Description of a Neural Network

This process of learning and subsequently improving the accuracy of the output is based on the algorithm of *back propagation*. It is based on giving the aforementioned coefficients varying weights (either positive or negative) and determining how these weights should be adjusted. The goal of the neuron is to arrive at the point of minimum error, called the *descending gradient*.

With this defined, it is now possible to proceed in creating a simple version of this artificial neuron using Python code as its foundation. The idea was simple, the neuron would consist of three inputs that would produce one output (similar to that of *Fig.1.2*).

For simplicity's sake, take the three inputs as consisting of either 1's or 0's and the following output would then be either a 1 or a 0. Take this table as an example:

	Input			**Output**
Example 1	0	0	1	1
Example 2	1	1	0	0
Example 3	1	0	1	1
Example 4	1	0	0	0

Table 1.1 Training Data Set.

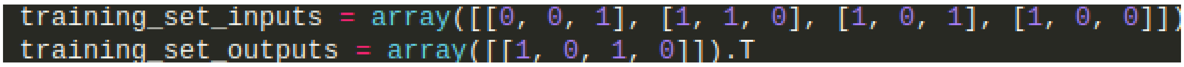
New Scenario	0	1	0	**?**

Table 1.2 New Scenario.

What should the **"?"** in the new scenario be? For humans with the ability to think and observe, finding the pattern is easy. The **"?"** should equal 0 as the pattern is simply that the output is always equal to the rightmost input in the array (i.e., the third number inputted). With this said, how does one teach an artificial neuron how to see this pattern? It is done as previously described by weighting every input with either positive or negative numbers. An input with either a large negative or positive input will have a noticeable effect on the neuron's output.

Thus, the learning process for the neuron can be broken down as such: (i) the neuron takes the input and assigns a weight to it; (ii) they are then put through the function of activation to calculate the neurons output; (iii) the neuron then calculates what is the difference between this output and the desired output (the correct answer); (iv) knowing this it can adjust the weights either more negatively or positively based on the knowledge gained in part (ii); finally (v) steps (i) through (iii) are repeated X amount of times (for the purposes of this exercise 100,000 times).

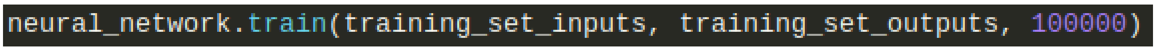
With this procedure now defined, the time has come to write some code. This entire procedure can all be done in only 9 lines of code. Take the table above (*Table 1.1*) as the training data set of inputs and outputs. This can be represented in Python in 2D-arrays as so:



```
training_set_inputs = array([[0, 0, 1], [1, 1, 0], [1, 0, 1], [1, 0, 0]])
training_set_outputs = array([[1, 0, 1, 0]]).T
```

Figure 1.3 Training data set inputs and outputs.

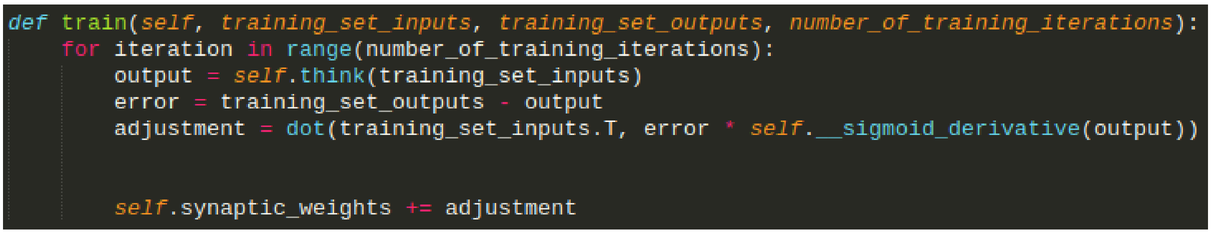
After this the starting weights are printed to the console and the network can be trained 100,000 times passing in the aforementioned training data sets to the *train* function as can be seen below:



```
neural_network.train(training_set_inputs, training_set_outputs, 100000)
```

Figure 1.4 Training the neural network 100000 times.

In the *train* function the data set is passed into the single neural network (single neuron), the error is then calculated between the predicted output and desired output and this error is then multiplied by the input and again by the descending gradient this means less confident weights are adjusted more and outputs which are as expected do not cause the weights to change. This *train* function can be seen expressed in code in more detail below:

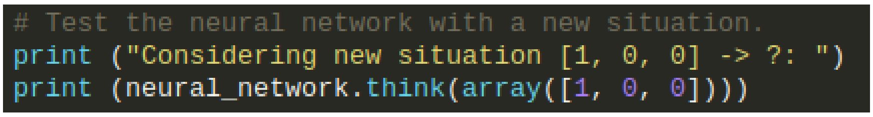


```
def train(self, training_set_inputs, training_set_outputs, number_of_training_iterations):
    for iteration in range(number_of_training_iterations):
        output = self.think(training_set_inputs)
        error = training_set_outputs - output
        adjustment = dot(training_set_inputs.T, error * self.__sigmoid_derivative(output))


        self.synaptic_weights += adjustment
```

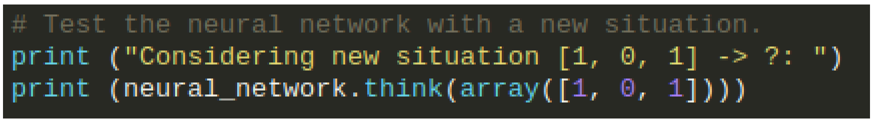
Figure 1.5 Train function.

Finally, after training the new weights they are printed to the console and a new situation is presented to the neuron. For these two separate scenarios were presented, one where the expected output was 0 and another where the expected output was 1 like so:



```
# Test the neural network with a new situation.
print ("Considering new situation [1, 0, 0] -> ?: ")
print (neural_network.think(array([1, 0, 0])))
```

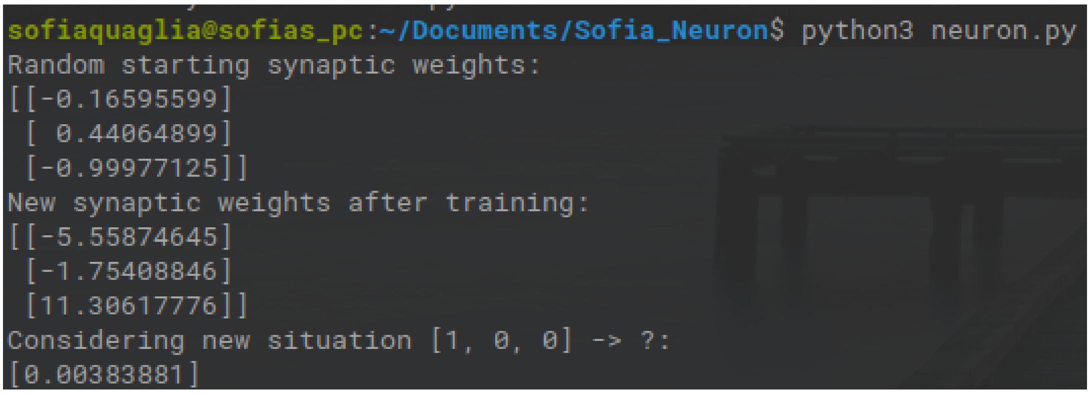
Figure 1.6 New scenario with expected outcome 0.



```
# Test the neural network with a new situation.
print ("Considering new situation [1, 0, 1] -> ?: ")
print (neural_network.think(array([1, 0, 1])))
```

Figure 1.7 New scenario with expected outcome 1.

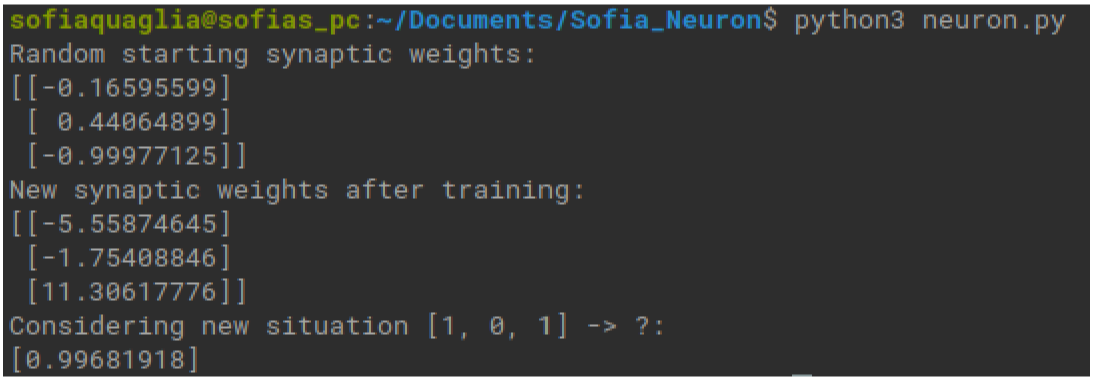
The results of these scenarios can be seen in the two figures below:



```
sofiaquaglia@sofias_pc:~/Documents/Sofia_Neuron$ python3 neuron.py
Random starting synaptic weights:
[[-0.16595599]
 [ 0.44064899]
 [-0.99977125]]
New synaptic weights after training:
[[-5.55874645]
 [-1.75408846]
 [11.30617776]]
Considering new situation [1, 0, 0] -> ?:
[0.00383881]
```

Fig 1.8 Results of the program with resulting weights after training

including new situation with expected output of 0.



```
sofiaquaglia@sofias_pc:~/Documents/Sofia_Neuron$ python3 neuron.py
Random starting synaptic weights:
[[-0.16595599]
 [ 0.44064899]
 [-0.99977125]]
New synaptic weights after training:
[[-5.55874645]
 [-1.75408846]
 [11.30617776]]
Considering new situation [1, 0, 1] -> ?:
[0.99681918]
```

Fig 1.9 Results of the program with resulting weights after training

including new situation with expected output of 1.

It can be seen from the results that after training the neuron, the weight given to the rightmost input is significantly greater and more positive than that of the weight attributed to either the leftmost or middle input. With this said, when the neuron is presented with a new scenario and the expected outcome is *0*, the neuron will provide an answer such as *0.0038…* arriving much closer to *0* than to one. The same can be said of the scenario where the expected outcome is *1* the neuron will produce an answer such as *0.9968…* which is much closer to *1* than *0*. It can be assumed that with enough training (a larger data set or a larger number of repetitions, X), the neuron will tend towards one particular predicted outcome.

The neural network just constructed shows how through the assignment of coefficients and the training phase of the neuron, when faced with a new scenario, it will tend towards giving a correct answer, despite the fact that the scenario is completely new and never seen before but is instead based on the input or historical data using this to predict for unseen data. Obviously, the technical example presented is that of an extremely simplified situation compared to that of AlphaGo, which has to decide which move of Go to proceed with in order to defeat the opponent. However, the fact remains that the neural network constructed shows the fundamental learning processes on which more complex technologies, i.e., AlphaGo, are then built.

Observing the technical functioning of the neural network demystifies the computational process that can easily be placed in analogy with humans[225] . Just as one should not draw hasty similarities, in the same vein it is counterproductive to establish a priori diversifications. In fact, deep learning algorithms allow multilayer computational models to learn representations, and the less restrictive they are (i.e., not designed to focus on one specific pattern), the more efficient they become in operating in new, never before seen, scenarios. As has been shown, algorithms are not specifically written to move in a given situation. The algorithms in AlphaGo learned to play and win the game of Go through experiential learning.

Think again of the distinction drawn between determinism and fatalism: fatalism is the existence of a destinal fate that attracts, pulls the subject towards itself and every

[225] This hasty analogy is also provoked by the use of the same terms. Since an analysis of this problem has been devoted to it on pp. 64-70 of the present work, it will not be repeated here in this passage.

action is taken in relation to that point of arrival, to reach it. Determinism instead is a push. The neural network constructed and the experience of AlphaGo show two application cases in which the algorithms are not programmed ad *hoc* but the behaviour of the artificial agent is guided by a specific objective. Rather than deterministic algorithms, perhaps it would be more appropriate to speak of 'fatalistic' algorithms, if one really wants to say something about it.

Or, from another point of view, one could consider the behaviour of such algorithms as free, since the process is not determined by deterministic algorithms but driven by the achievement of a goal, i.e., winning the game of Go. This goal is made explicit and consolidated through a process of reinforcement learning. One has to wonder whether freedom for humans is so different. Especially after learning how difficult it is to justify the existence of free will in humans.

Humans also perform actions with a view to achieving a certain goal, perhaps what is different is that they are supposed to be able to decide the goal themselves: it is not set by an external agent. Although it is quite obvious that there is some external influence, think for example of university study, you go to university because you want to have a certain kind of life and do a certain kind of work, because you are surrounded by people who among other things do that and through a process of identification you want to be similar. The result is that in order to achieve that kind of life, one studies every day and makes daily decisions in view of the long-term goal. An advocate of free-will would respond that this example misses the crux of the matter, that the causation being talked about is much more subtle and complex, sophisticated and immeasurable because it is human.

What is different between the free will of a human agent and the goal-oriented choice of an artificial agent? What is the role of autonomous, or semi-autonomous, choice in goal setting? What is the deep meaning of freedom? Why do human agents need to be free?

3.3.3 Non-deterministic or partially observable environments

Through the construction of the neural network and the case of AlphaGo it is shown that in some cases the algorithms, in this case the algorithms of Deep Learning, are not programmed specifically for a given task. The deep learning algorithms are trained to identify patterns and to make predictions (recommendations) based on (new) input scenarios. The opponent has made a move. It is not the algorithms design to predict what might be the best "next" move for itself. However, it is by combing many of these neural networks and some decision-making framework that serves up a recommendation as to the next best step. The neural network model that leads the artificial agent to give a certain *outcome*, i.e., to decide, are determined by the probabilistic calculations: in the learning with reinforcement, for example, they are determined by the probabilistic calculation and corroborated by a positive stimulus rather than a negative one.

To believe that programmers can always decide every single step in advance and are able, in any case, to determine, by default, what the artificial agent should do if X i.e., certain circumstances, is quite unfounded. If, however, in a hypothetical scenario, it is in fact true, that in computer engineering practice, there are circumstances in which it is possible to programme every single action of the A.I. because the environment is completely observable and deterministic, in such circumstances the A.I. knows exactly the effects of each of its actions: in this situation the A.I. can calculate exactly what state results from any sequence of actions and always knows what state it is in. Its perceptions after a certain action, its *stimulus* (positive or negative), are not a source of new information.

It is also true that the circumstances in which it is impossible to sequentially predict all the possible situations in which the A.I. will find itself are very frequent: the Roomba robot hoover is an obvious example. Programmers have to create a robot that moves and acts in an environment that is only partially observable and not deterministic, where the robot's perceptions become useful. In a partially observable environment, each perception helps to narrow down the set of possible states in which the agent might find itself, and consequently can more easily achieve its goals. When the environment is non-deterministic, perceptions tell the agent which of the possible outcomes of its actions actually occurred. In both cases, future perceptions cannot be determined in advance and the agent's future actions will depend on them. [226]

An example of non-determinism is that of the erratic hoover:

[226] Cfr. S. Russel, P. Norving, *Artificial Intelligence. A modern approach (1),* F. Amigoni (Ed.), Pearson Prentics Hall, Milan-Turin 2010, cit. p. 160.

> In order to provide a precise formulation of this problem, we must generalise the notion of a transition model. Instead of defining the transition model with a RESULT function that returns a single state, we use a RESULT function that returns a *set of* possible result states. [227]

Imagine this hoover as having multiple possible states. Observing one of these possible states (State 1) in which there are two squares A and B (placed to the right of A), both dirty. The hoover is on square A: the action "vacuum" can lead to two different states (i) the hoover sucks up the dirt in square A, moves to the right and also sucks up the dirt in square B; (ii) the hoover sucks up the dirt in square A but does not move to the right and does not suck up the dirt in square B. This means that:

> We must also generalise the notion of problem solving. For example, if we start in State 1, there is no single sequence of actions that solves the problem. Instead, we need a contingency plan like the following: [Aspire, **if** State = 5 **then** [Right, Aspire] **else**[]]. [228]With 5 = A Clean, B Dirty.

In non-deterministic circumstances, since it is impossible, and non-functional, to use sequences of actions, it is useful to use "search trees", opening the possibility for the artificial agent to choose, to select the action according to the contingent situation. It is interesting to note that problems in the physical world are contingency problems and one is reminded that human beings deal with this type of problem every day, i.e., driving a car. So how does one find contingent solutions to non-deterministic problems?

In a deterministic environment the ramifications of the tree[229] are given by the choices of the agent in each state (called OR nodes), so in a deterministic environment the hoover can vacuum *or* move right *or* move left. In a non-deterministic environment, instead, the branching is also given by the environmental result of each action, in the example of the erratic hoover the agent has to find a solution for frame A but also (AND node) for frame B. In a non-deterministic search tree, the OR-AND nodes alternate.

> A solution for an AND-OR search problem is a subtree that (1) has a target node in each leaf, (2) specifies a single action in each of its OR nodes, and (3) includes every branch exiting each of its AND nodes. [230]

[227] *Ivi*, cit. p. 162, transl. mine.

[228] *Ibidem*.

[229] A decision tree is a flowchart diagram that shows the possible outcomes from a list of possible decisions. It is used to represent the decision-making process of an algorithm. The reason for its use is due to its intuitive mode of representation.

[230]*Ivi*, cit. p. 162, transl. mine.

So, if one applies this solution to the problem of finding the erratic hoover, the process can be described like this:

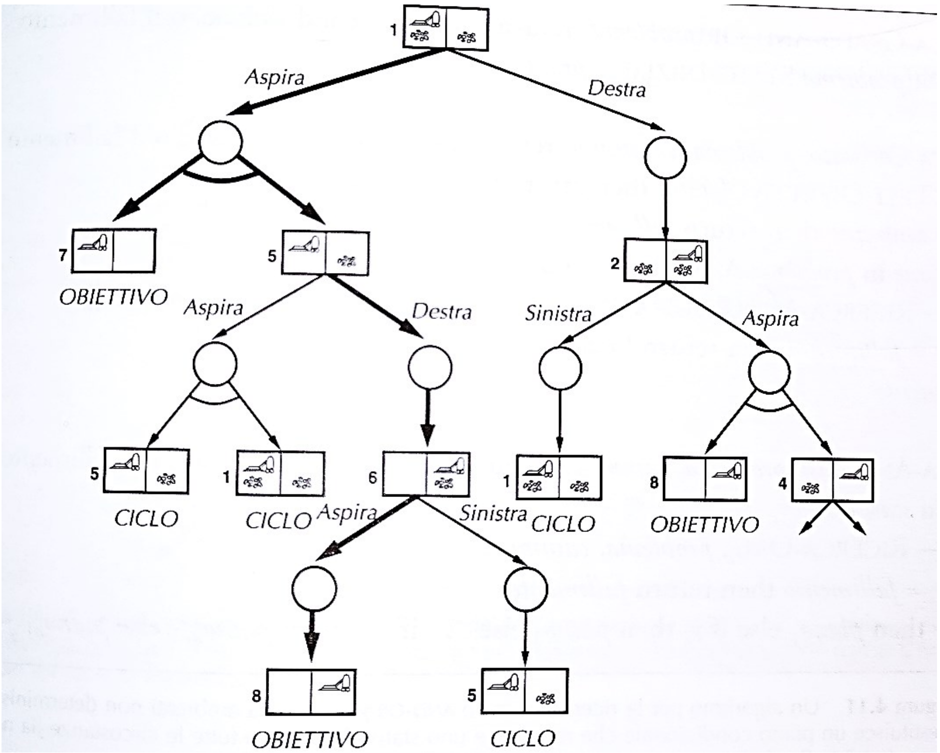


Fig 1.10 Research tree representing the world of the erratic vacuum cleaner.[231]

In (State 1) the hoover is in dirty square A, B is also dirty: the agent has to decide whether to vacuum or (OR node) move to the right: node AND the hoover aims to clean both squares A and B, therefore (State 2) it vacuums square A. At this point it could either vacuum or move to the right in frame B, since frame A is already clean, it moves to the right in frame B (State 3). Now it can either hoover or move to the left: if it hoovers it reaches the target, but if he moves to the left in frame A it does not reach the target, therefore: it hoovers (State 4) target reached.

The problem of partial observability has been said to be that circumstance in which the agent's perceptions are not sufficient to determine the exact state[232]. How is this impasse overcome?

[231] Cfr. Ivi, p. 163.

[232] Cfr. *Ivi,* p. 165.

> The key concept required for solving partially observable problems is the *belief state*, which represents what the agent currently believes about the possible physical states it might be in, given the sequence of actions and perceptions up to that point. [233]

Imagine a hoover without sensors: what is certainly loaded onto the system (what corresponds to the acquired knowledge of the agent) is a map of the world in which it is located i.e., the geography of the kitchen, at the same time, however, the agent is not aware of its current state in that room, its position, and the condition of the dirt. Being sensorless, the hoover will explore its belief-states instead of the physical states, in a sense forcing the world into the desired state: the target. In belief-states the environment is fully observable because the agent always knows its belief-state.

The problem is that the belief states, the result of observation absent in the case of a hoover without sensors, are immense. Standard search algorithms treat belief-states as 'black boxes': a belief-state has the same value as any other state. So, one solution might be to examine the content of belief-states to develop incremental belief-state search algorithms that design a construction considering one physical state at a time. It is a type of approach that has the advantage of detecting error and belief-state failure, i.e., unresolvable belief states. [234]

> Just as an AND-OR search must find a solution for each AND node, this algorithm must find a solution for each state in the belief-state. The difference is that the AND-OR search can find a different solution for each branch, whereas an incremental state-belief search must find a *single* solution that works for *all* states. [235]

In order for an artificial agent to interact intelligently with the world, it is essential that it keeps its belief-state constantly updated; this is possible thanks to an equation called recursive state estimator: this calculates the new belief-state on the basis of the previous belief-state instead of the whole sequence of perceptions. Recall that partial observability, uncertainty are the main characteristics of most real environments, as well as the fact that an environment is stochastic. Phenomena occurring in a certain environment are stochastic when they follow random or probabilistic laws, think of the activity "driving a taxi": no one can predict traffic conditions exactly and the car

[233] *Ivi*, cit. p. 166.

[234] Cfr. *Ivi.*, p. 170.

[235] *Ibidem*.

could be damaged at any time, unexpectedly[236]. When a situation is stochastic, the most appropriate approach for an artificial agent is probabilistic computing. [237]

Artificial agents that have belief-states about the world is an interesting concept: human agents also have beliefs. As has been shown, artificial agents have belief-states when they cannot access any kind of observation of their surroundings or are limited to partial observations. It has also been said that to cope with such a situation it is useful to use probability theory to construct a decision-making process that makes sense and is effective. Yet, what is a 'belief'? On what occasions do human beings have beliefs about worldly events?

In *Enquiry Concerning Human Understanding,* David Hume [238]claims that human beings form an idea of the causal relationship[239] between two phenomena by repeatedly experiencing their close occurrences. In addition to the idea of the causal relationship between events, the habit of seeing them together, rather than as a consequence of each other, arises and over time this habit is reinforced and becomes a *belief.* A belief, in a healthy subject, it has the characteristic of being sustained by the redundant experience of that phenomenon which takes place temporally and causally in this way and has the above-mentioned characteristics.

The causal relationship follows the laws of classical and quantum physics. When you have a law there is no uncertainty, you are sure that the phenomenon will follow that course because it has certain characteristics to which certain laws can be applied. However, if the situation is new you do not know what to expect, a degree of uncertainty comes into play: now, how can you work in the face of uncertainty and unpredictability? What strategies do human agents use? Are they different from those of artificial agents? Before continuing with the analysis of this discourse, it is useful to take a step back.

3.3.4 Quantum Computing

In the theoretical introduction of libertarianism and determinism, one of the libertarian approaches was mentioned, the so-called *event-causation*. For the latter, the

[236] Cfr. *Ivi*, p. 54.

[237] Cfr. *Ivi*, p. 174.

[238] D. Hume, *Enquiry Concerning Human Understanding*, Cambridge University Press, Cambridge, 2007.

[239] Keep in mind that for Hume causation was an obscure concept, more like a feeling than a law.

self-determination of the agent requires a non-deterministic and non-deviant causation, which can be *fully* ascribed to the reasons of the agent. The question has arisen: how is it possible for a cause to be non-deterministic? Is this not a contradiction in terms?

Quantum computing will be introduced since it could be a proof that the state of matter is not determinate but uncertain, as is the choice of a human equipped with free will.

If one interprets the world according to the laws of classical mechanics, such a statement makes no sense. Classical Newtonian mechanics work at the level of atoms, describing their relationships and states, and it is assumed that the bodies of living beings, particularly humans, are no exception. But they are undoubtedly part of the order of nature. As has been shown above, determinism uses the laws of classical mechanics to deconstruct libertarian theories.

However, with the advent of quantum mechanics, *supporters of* free will have an additional tool to corroborate the indeterminacy of a part, or parts, of the structure of reality. *Quantum computing* will be presented and the emergence of indeterminacy from a computational point of view will be observed. G.E.M Anscombe in *Causality and Determination*[240], argues for the possibility of non-deterministic or probabilistic causation.

Non-deterministic causation is the causation of a probability or, read the other way round, the probability of a causation. Probability is a quality that an event has when it is more reliable than another event. One of the statistical theories is the Bayesian interpretation of probability, in which probability expresses the degree of credibility of an event. Probability has a very intimate relationship with uncertainty. What is the meaning of uncertainty? Perhaps, to find the answer is useful to look at the opposite situation. Let us look at the opposite situation.

A situation is certain when it is reliable, it is guaranteed that a given piece of knowledge, a given outcome is true: Boolean algebra, or Boolean logic[241], in which true and false values correspond to 1 and 0, works with certainty. This has been fundamental to the development of electronics and modern programming languages;

[240] G. E. M. Anscombe, *Causality and Determination*, Cambridge University Press, New York 1971.

[241] Cfr. B. F. Braumoeller, *Causal Complexity and the Study of Politics,* In *Political Analysis* 11(3), pp. 209-233, 2003.

it is used as a statistical theory in those situations where multiple different data are expected to produce the same result: the *outcome of the* variables is binary, and the interaction between the different variables is specified limited to the logical positions 'and' and 'or', which are also binary.

Binary Boolean logic is incapable of dealing with uncertainty, so *fuzzy logic* comes into play: a multi-valued logic in which the result of the variables can be any real number between 0 and 1. *Fuzzy* logic is better suited to uncertainty and is, in fact, used in quantum computation. Bayesian statistics is similar to *fuzzy* logic and actually builds upon it, with regard to uncertainty.

Bayes[242]' statistics is a theory whose foundation makes it possible to prove key properties of the system and to develop scalable algorithms that are likely to perform the correct action. This type of statistic operates through the posterior probability approach[243]. In the field of A.I., Bayesian statistics are frequently used.

> Suppose you awake in the middle of the night on a strange planet. Even though all you can see is the starry sky you have reason to believe that the sun will rise at some point, since most planets revolve around themselves and their sun. so you estimate the corresponding probability should be greater than one-half (two-thirds, say). We call this the *prior probability* that the sun will rise, since it is prior to see any evidence. [...] it reflects you're a priori beliefs about what will happen, based on your general knowledge of the universe. But now the stars start to fade, so your confidence that the sun does rise on this planet goes up, based on your experience on Earth. Your confidence is now a posterior probability, since it's after seeing some evidence. Finally, a silver of the sun's bright disk appears above the horizon [...]. Unless you're hallucinating, it is now certain that the sun will rise. The crucial question is exactly how the posterior probability should evolve as you see more evidence. The answer is Bayes' theorem. [244]

The probability of the sun rising increases after seeing the sun in the sky. Bayes' theorem is useful because generally what is known is the probability of the effects given the causes, instead with the back probability (the one used in Bayes' theorem) one understands the probability of the causes given the effects. This is exactly how the neural networks constructed and analysed previously work[245], through the algorithm of *back propagation* the artifice after having obtained a certain result (effect), is able to recalculate the coefficients (cause) being informed about the distance from the

[242] Cfr. D. Poole, A. Mackworth*, Artificial Intelligence, Foundations of Computational Agents*, Cambridge University Press, Vancouver 2017, (https://artint.info/html/ArtInt_196.html), 14th January 2021.

[243] Cfr. P. Domingos, *The Master Algorithm: how the quest for the ultimate learning machine will remake our world*, Penguin Books, 2015, pp. 144-149.

[244] *Ivi*, cit. p. 146.

[245] A detailed explanation of how neural networks work can be found at pp. 109-117 of the present work.

defined objective, it re-adjusts the probability after having seen the *outcome,* exactly like the rising sun. The idea in A.I. is always to recreate a perfect Bayesian agent, a belief system and vector of strategies consistent with the aforementioned beliefs forming an equilibrium.

It has been shown how it is possible for an artificial agent to work under uncertainty, yet does this also apply to indeterminacy? How does indeterminacy emerge from a computational point of view?

A normal computer performs operations using classical bits, 0 or 1, while a quantum computer uses *quantum bits*, *qubits*. Quantum bits can be 0 and 1 at the same time, and there are physical objects that can be used as such, i.e., a photon, a nucleus, or an electron. Some researchers, for example, use the outermost electron of phosphorus (P) as a *qubit*, which is done by using the electron's magnetic field. Imagine an electron as a magnetic bar which, when placed within the magnetic field, aligns with it, and rotates, this property is called '*spin*'.

When the energy in the magnetic field is at its lowest level the spin faces downwards, if forces are applied to the magnetic bar (electron) the spin can be rotated so that it faces upwards. Spin up and spin down are the two states of a classical bit, two states 1 and 0. Quantum objects can be in both states at the same time. When spin is measured it will be found pointing up or down, yet prior to observation the electron exists in a so-called '*super-position*'. One way to indicate the super-position is through two coefficients, one pertaining to the probability of finding the electron facing up and the other pertaining to the electron facing down.

However, how do quantum *bits* operate in a quantum computer? Think of two *qubits* interacting with each other, two *qubits* are equivalent to four states of the two electrons. Quantum mechanics gives the possibility of creating a super-position for each of the four states, so a quantum mechanical state will be equivalent to the four probability coefficients, relative to the four states, added together: two *qubits* contain four *bits* of information. So, to clarify, if the information in classical bits is described by the formula N2, the information in *quantum bits* is 2N.

As has already been mentioned, when measuring a quantum state the super-positions are lost and each *qubit* will lie in one of the two states, thus being equivalent to a classical *bit. It is* precisely for this reason that in classical computation a final result is not sought after. Instead, what is proposed, is a design of the logical operations necessary to arrive at the final computational result so that the final result becomes visible, measurable, unique and above all, observable.[246]

A quantum computer is not proposed as a substitute for the normal computer, the one that works with classical bits; in fact, what the quantum super-positions implement is not so much the speed of each operation but the drastic reduction of the number of steps to arrive at a certain result. So, consequently, a quantum computer will not be useful for carrying out classical algorithms, such as machine learning or deep learning algorithms, but the fact is that for some operations and with some algorithms it could become the new norm.

The reason why *quantum computing* has been introduced in this work is precisely for the fact of the super-position, which has been described as a state in which a *qubit* lies in both quantum positions: not 1 or 0 but 1, 0 and any of the real numbers contained in that set. This means that the state of matter from the point of view of quantum mechanics is not determined but indeterminate and uncertain, as is the choice of a human agent endowed with free will.

Such a position could corroborate the libertarian theories of *event-causation*, where the agent's self-determination requires a non-deterministic and non-deviant causation, traceable in *toto* to the agent's reasons; but also, the so-called *agent-causation theories*. In fact, as Chisholm argues: the causal chain inherent in agents does not consist of events i.e., event causes event, but agent causes event, where the agent is not caused by anything else. Thus, the agent is the last link (or the first, depending on the point of view) in the causal chain and is therefore responsible for its actions.

Is *quantum computation* really a proof for the existence of free will, ultimately an argument for free choice or a demonstration of how vast human uncertainty is: is it free choice or just probability? Obviously, this is not the place to answer this complicated question. Yet, the analysis of the *quantum super position* has shown interesting aspects of libertarianism, of its possible adherence to the world, answering

[246] A. Morello, *Quantum Nanomagnets and Nuclear Spins: An Overview*, In B. Barbara, Y. Imry, G. Sawatzky, PCE Stamp (ed.), *Quantum Magnetism*, Springer, Berlin 2008, pp. 139-150.

(perhaps) the determinist critique, and has made it possible to draw some connections between human and artificial agents, has opened up the possibility of a confrontation.

3.3.5 Provisional conclusions on the 'free will' of artificial agents

The introduction of the main philosophical theories on free will gave a more specific insight into the ways in which the existence of free will has been justified for mankind over time.

The discussion has been biased towards libertarian theories with the aim of understanding if there is a possibility to consider some A.I. free or endowed with free will. The reason why it has been preferred to try to understand the possibility of this hypothesis (free will of A.I.) rather than the opposite, i.e., to prove the lack of freedom of human beings by supporting a determinist theory and then put humans on the same level as A.I., is simple.

From the legal point of view, as has already been shown, agency is treated in terms of capacity, in fact the capacity to act, which is acquired with majority, is the capacity to intend. "Intentionality" is a concept that will be presented very shortly. "Will, on the other hand, is the faculty, the capacity to will, to choose and to carry out behaviour; it can be traced back to what has been said about free will. Thus, the law [247]presupposes a subject who wills and freely chooses how to act. It is noted that in law freedom is considered more from the point of view of negative freedom and positive freedom, free will is used more as a way of describing reality without ever understanding it in a strictly philosophical sense i.e., *agent-causal libertarianism*, *event-causal libertarianism*, *non-causal libertarianism*.

The experiment conducted in this research aims at finding out whether there is a possibility to consider some of AIs as free, autonomous and responsible, i.e. as moral artificial agents in order to try to support the applicability of the already existing laws and regulations on the aforementioned artifices.

From the examples given and used, it can be seen that: (i) In the field of A.I. the determinist conception is often confused with the fatalist one. The construction of the neural network and AlphaGo have shown that their behaviour and the procedures that follow one another in achieving the objective are, precisely, with a view to obtaining

[247] In this specific case, reference is made to the Italian Civil Code, but it is believed that a similar argument can be made for the rest of Western legal systems.

the result. There is neither an algorithmic code guiding each move, in fact the programme is not constructed *ad hoc* but learns progressively through reinforcement, nor does the previous action guide the next one: it was shown through posterior probability that the coefficients are re-established following the observation of effects and the comparison between the aforementioned provisional outcome and the desired result (objective).

Thus, it was noted that even in human decision-making one works with goals, and there is always a certain degree of external influence, dependence on the rest of the humans and environment. Therefore, the question was asked: is the capital difference between humans and A.I. in the way the goal is chosen? On balance, it would seem that the humans, although influenced, can choose their own objective autonomously; on the contrary, the A.I. i.e., algorithms of Deep Learning, AlphaGo, can never self-determine their objectives: these are imposed from the outside, from the human.

Greek fatalism has been described as the philosophical doctrine that emphasises the attribution of all events or actions to fate and destiny. Yet, despite the introduction of inevitability and necessity into the human decision-making game, people in ancient Greece were held responsible for their actions. This might be an indication that responsibility and free will are perhaps not so necessary or essential to each other.

As already mentioned in the *First Section: Rights*[248] of this work, legal laws make moral laws effective and are traceable back to them. Moral laws are (and have been) fundamental in maintaining social equilibrium. Moreover, the weakening of the link between free will and responsibility would also suggest that: from a deterministic point of view, the agent from dangerous and harmful attitudes is removed to protect the harmony of the social fabric. Being dangerous implies neither being free nor being responsible, one can become dangerous for many reasons i.e., not knowing any alternative to dangerousness. The objective of removal would then become that of re-educating the harmful agent for effective reintegration into the social fabric. This discussion will be continued later.

Going on to expose non-deterministic algorithms and environments with no or partial observations has allowed us to recognise the similarity of circumstances in which some AIs and humans find themselves when they have to make decisions. In

[248] *First Section: Rights* can be found at pp. 11-58 of this work.

non-deterministic contexts, since it is impossible, and non-functional, to use predetermined sequences of actions, it is useful to use "search trees". As has been shown the AND-OR search trees offer the possibility to the artificial agent to choose, select the action according to the contingent situation. When it is not possible to observe completely the surrounding environment, the artificial agent operates on the basis of its belief-states: thus, the concept of uncertainty has been introduced: how to operate in uncertain and unpredictable situations? What strategies do human agents use? Are they different from those of artificial agents?

Bayesian probability theory has been identified as one of the possible strategies in the face of uncertainty, in fact Bayes' theorem makes use of posterior probability which includes the probability of causes given effects (exactly the inverse of the prior property which only knows the property of effects given causes).

Quantum computing has shown how unpredictability emerges from a computational point of view. Since one cannot know the state of matter i.e., the outermost electron in the phosphor, i.e., the position the electron will take in the magnetic field. The libertarian theory interprets the indeterminacy of the electron being at the same time *spin up* and *spin down,* i.e., at 1 and 0, specifically any real number between 1 and 0, as corroborating evidence for the existence of free will from a physical point of view.

This is not the place to decree the reality or otherwise of this appropriation, nevertheless, it is interesting to note that from a physical point of view, if only of quantum mechanics, unpredictability is not a problem: the state of the electron at the same time can be any real number contained in the set [0,1]. The one who does not tolerate indeterminacy, in a sense not even really being able to imagine it, except from a verbal point of view, is the human being and his senses through which he comes into contact with the world.

3. 4 Intentionality: awareness and deliberation or the power of the mind to represent a state

Previously in the present research work intentionality has been discussed by examining the *standard* theory of moral agency and action, some central points of which are worth reiterating. According to the standard theory[249], AIs could be considered moral agents if (i) they are independent, i.e., autonomous, and if (ii) they have mental states. An action, which was said, is only such when it corresponds to a mental state: it is the latter that is the cause of the event (action) and of its effects in the world. Thus, the agent is to be held responsible for his actions because they originate from and coincide with the subject's intentional mental states.

The question has been asked as to where this need to link agency to intentionality comes from, and an answer has been attempted through the conceptuality of the theory of mind: that is, the ability to attribute mental states to subjects with whom one interacts. The biological invoice of this capacity, if one can say so, resides in the system of mirror neurons that allows the experience of an embodied simulation of the observer who puts himself in the place of the observed.

Intentionality and theory of mind were then taken up by Daniel Dennett, in fact according to Dennett it is fundamental to observe agency from an evolutionary point of view: macromolecules have *agency*. Obviously, it is not as complex as human agency, but it is present because macromolecules can also perform actions and can be considered intentional. An intentional system, for Dennett, is a system that operates on the basis of information and objectives. The *intentional stance* is the interpretation of the behaviour of an entity as being rational, intentional and free. Finally, *intentional stance* was compared with the theory of mind, concluding that an artificial intelligence could be treated as intentional so as to be able to interact with it by grasping and predicting its intentions.

The problem that emerges from such a broad and heterogeneous treatment of the concept of intentionality is that (i) considering it as a specific mental state (intentional mental state) indispensable for acting, A.I.'s are excluded regardless, as are animals and even some humans. If (ii) one defines intentionality as active behaviour, given certain information, with a view to achieving goals, intentionality becomes a more inclusive concept. There is a need for clarity in order to understand what "intentionality", or "being intentional", means.

[249] Cfr. D. G. Johnson, *Computer systems, Moral Entities but not Moral Agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 168-183.

In common language, intent is a conscious and deliberate act, i.e., from a legal point of view, a voluntary or involuntary killing. A premeditated, intentional homicide presents a homicidal subject who wishes to kill the victim and constructs a strategy to carry out the plan[250]. The psychological coefficient of the homicidal will is defined as 'malice', i.e., the conscious will of a person to cause harm to others. Premeditated voluntary manslaughter is what is known in America as 'first-degree murder' and in some states leads to the defendant being put on *death row*.

Unpremeditated voluntary manslaughter occurs when someone negligently causes the death of a person. Fault is the consequence of negligence and carelessness, rather than lawless conduct. [251]

Murder may also be unintentional, i.e., manslaughter: murder is committed when a person, by carrying out acts intended to injure a person, unintentionally causes that person's death.[252] Pre-intentionality occurs when, as a result of an action (or omission of an action), an event occurs that is more harmful, more dangerous, and more serious than that intended by the agent, beyond their intention.

So, from the legal point of view, intentionality can be said to be expressed in the generic sense of the term as: a conscious and deliberate act; but what about the philosophical sense?

In philosophy intentionality is the capacity to have representations, mental states that stand in place of something i.e., situations, objects, exchanges, relations, objects. So, to have intentionality means to have a representative mind. This meaning of intentionality is suggested by Franz Brentano at the end of the 19th century, in *Psychology from an Empirical Standpoint*[253] *he* formulates the thesis of intentionality as follows:

> Every mental phenomenon is characterized by what the Scholastic of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself. [254]

[250] Cfr. *Italian Criminal Code,* Art. 42, Intentional Homicide, transl. mine.

[251] Cfr. *Ivi*, Art. 589, Manslaughter.

[252] Cfr. *Ivi,* Art. 584, Manslaughter.

[253] F. Brentano, *Psychology from an Empirical Standpoint*, transl. By A. C. Rancurello, D. B. Terrell, L. McAlister, Routledge, London 1973.

[254] *Ivi*, cit. p 68.

Thus, intentionality is a mental phenomenon characterised by reference to a content, representation of that content. Think of the photograph of a dog, the name of the dog, the common name 'dog' or the concept expressed by the word 'dog' which means, represents, stands in for the hairy creature that occasionally barks. [255]

Brentano's studies are the beginning of very long analyses and interpretations that continue even now. Without having the presumption to make a complete survey of the concept of intentionality, a useful path will now be defined for the purposes of this portion of the work: is it possible to consider an intentional AI, in the philosophical sense of the term?

To recapitulate, intentionality is a specific mental state that refers to content X: a representational mental state; yet what form does the mental state take? What does it look like? It seems promising to focus on the concept of "representation": How can a complex representation take its meaning from real objects, while at the same time be detached from them, autonomous? What is the difference between iconic and verbal representations? Are representations part of the world they represent?

3.4.1 The form of mental states

Representation seems to play a central role in the conformation of mental states and thus of intentionality (intentionality as the capacity to have certain mental states). Representation means showing an aspect of reality that is physically absent, reproducing it by means of a substitute that recalls it. This substitution can be achieved and implemented in different ways, i.e., words, pictures, and signs. So, one keeps a picture of one's dog in one's pocket while travelling in order to "have him near".

A very interesting case is that of iconography and the history of icons. An icon is a sacred image, representing sacred figures and created according to a specific process, which is clearly defined and obligatory. In contrast to the reproductive image, i.e., the photograph of the dog, which is a merely denotative image, signifying only what it represents; an icon is symbolic, representing what is not present in two senses: both what is not present in that specific circumstance, and what is not present in earthly reality, something transcendental. [256]

[255] This example is used by P. Jacob, *Intentionality*, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta Winter 2019, https://plato.stanford.edu/entries/intentionality/, p. 1.

[256] Cfr. L. Russo, *Vedere l'invisibile, Nicea e lo statuto dell'Immagine*, Aesthetica, 1997.

However, what happens at the mental level? Why does looking at an icon depicting God evoke a set of behaviours and meanings? How is it possible that a symbol refers to something else? i.e., the icon shows what is invisible and what is transcendental: God. What happens at the mental level?

Among human practices there is a representational activity par excellence, this is used unconsciously in every social, as well as internal, exchange: verbal language. Why is it that when A says 'dog' to B, B understands and behaves accordingly, i.e., moves to let the dog pass, responds verbally to A, strokes the dog? "Dog" is a linguistic sign.

> A sign is linguistic when it means -consciously- the same thing to all individuals in a community. If I fold the corner of the page of a book or make a certain grimace while playing cards with my habitual companion, that fold and that grimace are undoubtedly signs, but not linguistic: they do not mean the same thing to everyone. If, on the other hand, I say 'la porta, la luce' we all understand (provided we know Italian). By saying this, I awaken in others as in myself the set of attitudes and consequences that are linked to these words, that is, I evoke an identical meaning for everyone. [...] Linguistic is that sign which, each time it is produced, inwardly or outwardly, compacts in our minds a certain network of associations, of behavioural patterns, arouses this or that meaning (concept). [257]

Thus "dog" is a linguistic sign because, when used, it generates behaviour: in B's mind significant associations with the word "dog" "appear". All this can happen because "dog" is a concept; but let us proceed in order.

A linguistic sign is (i) structurally intersubjective-public, shared and sharable by every individual with knowledge of the language, furthermore a linguistic sign can be used irrespective of the context this means that it must be (ii) used consciously, known, the speaker knows the meaning of the words he uses, indeed he is saying them precisely because he wants to provoke a reaction and (iii) the reason why the linguistic sign can be used irrespective of the context is that it is not bound to the occurrence of certain circumstances, it can mean in the absence[258] : when A says "dog" B thinks of the dog they saw the day before, or of the dog he would like to buy.

The linguistic sign is a concept, i.e., a defined and ideally figured thought, formulated and usable on an intuitive, logical, and practical level:

[257] C. Di Martino, *Linguaggio e mondo. Il potere della parola*, in G. P. Terravecchia, M. Ferrari*, I Quaderni della Ricerca 54 - Linguaggio e Mondo, Il potere della parola*, Loesher Editore, Bologna 2020, pp. 27-41, cit. p. 30, transl. mine.

[258] Cfr. *Ivi.*, pp. 30-31.

> The meaning of the word is constitutively absent, it is a specific absence. [...] language universalises structurally, as Hegel points out, so it always says the absent, that is, the universal, the general, the ideal. [259]

Yet, have the concepts always been there? How should one imagine their emergence? Ian Tattersall, in *The Lords of the Planet. The search for the origins of humans* [260] proposes a discontinuous view of the emergence of language: language is an evolutionary novelty developed as a result of ex-active evolutionary processes, i.e., the reuse of biological characters and structures previously formed for other reasons.

On the opposite side, a more convincing theory is the gradualist one of Michael Tommasello[261] and Michael Corballis[262] . They bring the history of the appearance of human language in line with palaeontology and its biological evolution, and they do so by inscribing the emergence of language within the set of effects of human ultra-sociality, as had already been said with the introduction of the theory of mind[263]. Thus, symbolic communication, language, is presented as a social institution that replaces (over time) socio-communicative activities already used: pointing, miming, signalling operations that are carried out by the great anthropomorphs. Symbolic language takes the place, diachronically, of these dyadic, binary, and consensual gestures, gradually detaching itself from the present context and becoming a concept, so that when one comes to say "dog" and there is no dog around, the listener and the speaker have the sense of "dog". [264]

To recapitulate, linguistic signs when used evoke meaning, their ideal meaning shared by a community, they evoke something that is not there at that moment, an absence that is present in the concept. Since mental states are a representation of a certain content in thought and intentionality is a mental state, it remains to be understood: how is the relationship between language and thought specified?

[259] *Ivi.* p. 31.

[260] I. Tattersall*, I signori del pianeta. La ricerca delle origini dell'uomo*, Codice Le Scienze, Torino 2013.

[261] M. Tommasello, *A Natural History of Human Thinking*, Harvard University Press, Cambridge USA 2018; *The Cultural Origins of Human Conditions*, Harvard University Press, Cambridge USA-London 2001.

[262] M. Corballis, *Dalla mano alla bocca, Le origini del linguaggio*, Raffaello Cortina Editore, Milano 2008.

[263] An explanation of the theory of mind can be found at pp. 75-77 of the present work.

[264] Cfr. C. Di Martino, *Viventi Umani e Non umani-Tecnica, linguaggio, memoria,* Raffaello Cortina Editore, Milano 2017, pp.125-162.

Although it is easy to believe, under a 'cerebralist' influence, that one thinks because one is endowed with a 'special' brain unlike any other living creature and speaks by virtue of this thinking brain. On the contrary: human beings think because they speak, and the human brain is so rich in circumvolutions and synaptic connections precisely because of language, the brain as an effect of language. Di Martino examines this relationship through the observation of children:

> When the child learns language, the linguistic tool has a decisive 'feedback effect' on his thinking, i.e., to use the categories of the psychologist Michael Tommasello, he becomes capable of objective, reflective, normative thinking. [265]

Language modifies and shapes thought. Thought is defined as objective because representation is intersubjective, shared by the community. The publicity of representation is the reason why one is able to communicate with the other. The quality of reflexivity emerges precisely from the fact that language enables thought and that, ultimately, thought is a silent dialogue with oneself. It is through language that one can make inferences about the thoughts of others: an ability that emphasises and inflates in potential that ability to recognise in the other fellow human with similar feelings, thoughts, i.e., theory of mind. Linked to this is the last characteristic of language, normativity.

> Finally, it is thanks to language that we can evaluate our thoughts and actions in relation to the views and reasons of everyone else, i.e. according to established rules, to a general ethics made up of norms governing the behaviour of a community. Authors such as Vygotsky and Lurija, observing the experience of children, have underlined the decisive role of language in the development of higher psychic functions and in the control of behaviour. [266]

This last point is extremely interesting for the present work. Language allows one to think about what one does, what others do and what should be done. The recursiveness of language allows thought to reflect on itself and since, as has been said, the signifying senses are synonymous with that network of behaviour, those actions in response to what has been said, this means that thought's reflection on itself is also a behavioural and consequential analysis.

In chapters *1.9 Consequentialism: good before right* and *1.11 The moral foundation of rights*, it was made explicit that: (i) moral laws serve to avoid situations

[265] C. Di Martino, *Linguaggio e mondo. Il potere della parola*, in G. P. Terravecchia, M. Ferrari*, I Quaderni della Ricerca 54 - Linguaggio e Mondo, Il potere della parola*, Loesher Editore, Bologna 2020, cit. p. 32, transl. mine.

[266] *Ivi,* cit. p. 33.

that are dangerous to the community; (ii) they are the historical and logical foundation of positive norms; and (iii) positive norms exist to make the contents of moral laws active and effective in society by establishing causes and effects. In a sense, one could say that, among other things, the concept of responsibility is also an effect of language.

Intentionality has been defined, from a legal point of view, as a conscious and deliberate act. On the other hand, the philosophical analysis, according to which intentionality would correspond to the capacity to have representations: mental states that 'stand in place of' something. For this reason, an investigation of the representative faculty was conducted, trying to follow (and propose) a logical and philosophical path that would allow us to identify the emergence of intentionality as "palpable", observable, so as to open up the possibility of an investigation into the intentionality of A.I.

Through Di Martino's work, the tool that allows us to have representations was identified: language, symbolic communication, which evokes the ideal sense because it is conceptual and allows us to talk about experiences, actions, individuals, and facts in their absence, in decontextualised circumstances. So, in conclusion, it has been said that 'intentionality' is the capacity to have representational mental states and language is that faculty which enables representation. Consequently, it would seem useful to investigate the relationship that some AIs have with language and the manipulation of symbols, in order to highlight the presence, rather than the absence, of intention in artifacts.

3.4.2 Intention as representation in artifice (philosophical sense)

The aim of this paragraph is to analyse the relationship between AI and language; to do this, the Turing test "*imitation game" will be* introduced in its structure and objectives. This will be followed by a reconnaissance of the results actually obtained by the artifacts subjected to this test, relying on the considerations that Floridi writes about the assignment of the Loebner prize (prize that is assigned when a machine passes the Turing test). The case of OpenAI's new GPT3 technology will then be proposed, as well as the case of the chatbot Vera, which is a chatbot developed by Indigo.AI and is used to dispel any fake news surrounding Covid-19.

Alan Turing proposes his test, *imitation game*, in *Computing Machinery and Intelligence in* 1950. As already mentioned, Turing opens the article by describing a game, the so-called *imitation game*. The *test* involves three people: a man (A), a woman (B) and an interrogator (C) whose gender is indifferent. The interrogator is separated from A and B and taken to another room, the goal for the interrogator is to determine who between A and B is the woman rather than the man. The way the interrogator collects data about the two characters A and B is through questions.

Then add another variable to the game A and B can be either a human or a machine and communication is given by message: C's goal now becomes to decree who between A and B is human or artifact. If it happens that C mistakes the machine for a human, it can be said that A.I. has won and can think. [267]

As is well known, the Turing test was coined to give direction to research in the field of A.I., for in the 1950s it was not yet clear whether imitation of biological tissues rather than mental faculties was important.

> The new problem has the advantage of drawing fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even when supposing this invention available we should feel there was little point in trying to make a "thinking machine" more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact. [268]

Thus, Turing declares the importance of going down the road in the direction of reproducing human intelligence. Ultimately, his experiment serves to prove whether machines can think, whether machines can be intelligent.

In the present work, an analysis and operational definition of intelligence has already been offered, and it has also been shown how a fair or lower level of human intelligence is not strictly related to the attribution of agency or not. The reason for expounding the Turing Test is not the measurement of intelligence, but the analysis of the relationship between A.I. and language.

In fact, in order to win the game, the machine has to imitate a human being and more specifically the human faculty of speech: by means of messages it has to answer the questions posed by the interrogator and be as convincing as possible. For Turing, an example of an exchange between human and machine could be the following:

> Q: Please write me a sonnet on the subject of the Forth Bridge.
> A: Count me out on this one. I never could write poetry.
> Q: Add 34957 to 70764.

[267] Cfr. A. Turing, *Computing Machinery and Intelligence*, in *Mind LIX (236*), Oxford University Press, Oxford 1950, p. 1.

[268] *Ivi,* cit. p. 2.

> A: (Pause about 30 seconds and then give as answer) 105621.
> Q: Do you play chess?
> A: Yes.
> Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
> A: (After a pause of 15 seconds) R-R8 mate. [269]

The problem is that the Turing test has never been passed by any machine. Why? What unmasks the artefact?

Every year, the Loebner Prize is set up, a competition in which a jury awards conversation software that is most similar to human beings. Luciano Floridi in *The Fourth Revolution*[270], recounts his experience as a judge in 2008 at the University of Reading in the UK. Floridi points out that it took only a few questions to prove that the software system was not remotely comparable to a type of human intelligence; in particular, the type of questions that revealed the alleged imitator were metalinguistic ones.

> One of us started the conversation by asking: "If we shake hands, whose hand am I shaking?". An interlocutor, the human one, immediately replied in a metalinguistic key, that the conversation should not refer to physical interactions. [...] The computer failed to answer the question and spoke of something else, a trick employed by many of the machines tested: 'We live in eternity. So, I would say no. We don't believe". [271]

How does an artefact work with concepts, meanings, and representations? When communicative software is asked simple questions for which a dry answer is expected, the software is perfectly capable of giving an answer, in fact the A.I. (which is of course connected to the Internet) can retrieve its answer there, the immanent amount of data available will most likely provide the software with the correct answer. The same happens with games, i.e., chess and AlphaGo, the software with memory would be able to refer to the previous move and the next move.

The problem emerges when the question is complex and the answer requires inference and reflexivity, as when a period consists of a main preposition and subordinates, or even simply two simple but misleading periods: 'The four capital cities of Great Britain are three, Manchester and Liverpool. What is wrong with this sentence?" [272]. The software cannot answer because it cannot understand the meaning.

[269] *Ibidem*.

[270] L. Floridi, *La quarta rivoluzione - Come l'infosfera sta trasformando il mondo*, Raffaello Cortina Editore, Milano 2014.

[271] *Ivi.* p. 152, transl. mine.

[272] *Ibidem*.

Thus, it was shown how the best communication software fails the Turing test, thus proving to have approximately human intelligence. The reason for this inability was also suggested. Searle, who has already been discussed[273], would define this problem as the difference between syntax and semantics. Searle explains this difference by means of the *Chinese Room* thought experiment and supports the thesis that it is true that software can produce responses by manipulating formal symbols despite not understanding any of them. Software can operate and manipulate symbols (words and concepts) at the syntactic level, but it does not understand the meaning, the sense, at the semantic level.

> As long as the program is defined in terms of computational operations based on only formally defined elements, what the example suggests is that these, in themselves, have no interesting connection with understanding per se. They are certainly not sufficient conditions, and there is not the slightest reason to suppose that they are necessary conditions or even that they make any significant contribution to understanding. [274]

Therefore, it is clear that no kind of understanding is possible for a communicative software and that defining A.I. activity as "thinking" seems to be excessive.

In fact, as was stated in the theoretical framing of the philosophical sense of intentionality and representative intentional mental states, human thought, which is (i) objective, (ii) reflective, (iii) normative thought, is a (retro-effect) caused by verbal language, which in turn is (i) intersubjective, (ii) conscious and (iii) signifying. It is difficult to believe that in an A.I. there can be such a thought, a human thought, if the language used is operable only at a syntactic and not at a semantic level.

However, the fact remains that the signs used by communicative software are human linguistic signs, not indications or prompts that occur in non-human animals.

The relationship between A.I. and language is unique: machines operate with linguistic signs that they do not understand and yet are able to manipulate. The linguistic signs with which they operate are extremely abstract concepts and decontextualised meanings that are completely understood (at a semantic level) by the human beings present in the interaction.

[273] Searle and his view have been introduced at p. 89 of the present work.

[274] J. R. Searle, *Minds, Brains, and Science,* Harvard University Press, Cambridge MA 1984, cit. pp. 50-51, transl. mine.

We have preferred to provide a dense theoretical framework representative of the state of the art in philosophy, but now the time has come to explore the technical-practical state of the art, with the aim then of drawing some provisional conclusions.

3.4.3 Conversational AI: Indigo.ai

Indigo.ai[275] *is* a Milan-based company that is a leader in the artificial intelligence and computational linguistics market. Many Italian and multinational companies use its services. Indigo deals with building and designing *chatbots* and technologies capable of manipulating language and having conversational experiences.

The work begins with an in-depth analysis of the company's data, Indigo processes the data and builds an explanatory map to show how an A.I. could improve, facilitate and enhance the consumer's experience and also that of the company. Before the customised product is introduced to the company's market, there are testing phases to make changes to the technology by observing how it performs and responds to challenges. The A.I.'s are produced entirely by them and this facilitates the maintenance of a high level of data quality, as well as the protection of privacy, in fact this means that there are only two parties involved (Indigo and the company) and there are no third parties involved.

One of the case studies is, for instance, *Vera* [276]for *Pagella Politica*. Vera is an A.I. built for the purpose of correctly answering questions and concerns about Covid-19, in order to defame *fake news*. All this stems from the fact that, at the same time as the Coronavirus pandemic, an *infodemic* dictated by fear and by power, has spread: there has been a rapid spread of unfounded and confusing news, which has found an audience ready to accept it. So, Indigo created an A.I. to check the truthfulness and reliability of the resources: *fact-checking*. Vera provides quick and precise answers to questions such as "Is it true that 5G can transmit Covid?"; "How long does the vaccine take to take effect?"; "Does the surgical mask really protect?".

After clicking on "talk to Vera" you have two choices: (i) "Verify hoaxes and data in the world" and (ii) "Questions about Covid-19". In the first section, the one concerning fake news, the user can check the veracity (or not) of suspicious news found on social media i.e., Facebook, Twitter and even WhatsApp. To test the A.I., for

[275]Cfr. Indigo.ai, (https://indigo.ai/en/), 20th January 2021.

[276] Cfr. The Artificial Intelligence that clears doubts about Covid and fights fake news, (https://indigo.ai/en/case-studies/pagella-politica), (https://www.chiediavera.it/), 20th January 2021.

the purpose of entering data in this thesis work, one typed in "Is it true that the vaccine causes the disease?" and the answer was "If this is the news you are looking for, it seems to me that it is news out of context" and attached is an article from FACTA[277] which points out that there is no correlation between the death of a person and the fact that he had been vaccinated against Covid.

When you want information about Covid, you can press the second button and Vera will provide medical and scientific material about the global pandemic. For example, to the question "what is the mortality rate" Vera replies: (i) "To check the most up-to-date official data on the Covid-19 epidemic in Italy, I suggest you consult the dedicated site of the Civil Protection at this link"; (ii) "Here you can find the world statistics of the World Health Organisation (WHO)"; (iii) "If you want more specific data on the characteristics of those infected in Italy, you can consult the section edited by the Superior Health Institute (Iss) at this link".

Each of the proposed links, those underlined, leads to the exact part within the site, not simply to the site, so as to facilitate the retrieval of the information, which is thus obtained in a clear and immediate manner. Vera allows users to find correct and documented information with the same ease with which they came into contact with *fake news*. In the first three months of operation, users exchanged 20,000 messages with Vera.

The use and deployment of Vera around the world has been described and it has been shown that it is able to respond more than acceptably to queries by being able to discriminate different forms of information and resources by interfacing different applications and social networks. Now it is time to understand how this technology works in more detail. Indigo uses Machine Learning and *Natural Language Processing* algorithms in the design of its technologies. While Machine Learning has already been extensively exposed and explained, algorithms capable of manipulating natural language have not yet been introduced.

[277] Cfr. Il volontario brasiliano di AstraZeneca non è deceduto a causa della sperimentazione del vaccino contro la Covid 19, (https://facta.news/fuori-contesto/2020/10/22/il-volontario-brasiliano-di-astrazeneca-non-e-deceduto-a-causa-della-sperimentazione-del-vaccino-contro-la-covid-19/).

3.4.4 Natural Language Processing (NLP) [278]

A language can be considered as a set of periods characterised by a certain measure and which cannot be of infinite length. These periods are constructed using a finite and predetermined alphabet. On a theoretical level, since the alphabet is finite and limited and so is the length of the sentences, this should result in a limited set of possible compositions (of possible sentences).

If the human alphabet were made up of two letters, it would be very easy to determine the total number of periods that could be composed, but since human language is much more extensive in terms of programming the algorithms that have to work on it, human language is considered infinite and boundless, and ways are found to translate it into languages that are more usable for manipulation. In fact, if a machine could work directly on human language, it would have to be a machine with an infinite memory and infinite processing power, and this is not the case. So how can a machine approach an infinite language with which it cannot work?

As mentioned earlier, finite languages called *generating grammars are* used to capture the structure of human language. The theory of the chosen grammar is used to analyse the different parts of human language and understand the functions of the different parts of speech. In addition, these grammars that replace human language eliminate the problems of human language, which is very complex and ambiguous, i.e., there are many words that mean different things depending on the context.

NLP can be defined as computational techniques for automatic language manipulation.

> Natural language processing (NLP) is a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory. [279]

The limitation that the aforementioned computational techniques have is an understanding of the deep sense of human language. Beyond the distinction between syntax and semantics, which has already been mentioned above, what the algorithms lack is all that remains implicit in language: an algorithm is perfectly capable of understanding the sense of a single sentence and of several sentences linked together, but it cannot know what happened before and after, it does not know the context.

[278] Cfr. K. R. Chowdhary, *Natural Language Processing* in *Foundamentals of Artificial Intelligence*, Springer, New Delhi India 2020, pp. 603-650.

[279] *Ivi,* cit. p. 604.

However, in order to reach the present levels of NLP technology development, high-level symbolic capabilities are required, such as an understanding of syntax and a certain degree of semantics. Furthermore, the technology must be able to manipulate recursive structures and thus identify fundamental forms and functions of language construction. Finally, they must be able to represent abstract concepts at a syntactic level[280].

NLP has a very broad field of application thanks to the World Wide Web, which consists mostly of words, phrases, and texts, creating very interesting application cases such as: index and large text search, information retrieval, text classification and categorisation, information extraction, automatic translation, automatic summarisation, automatic question answering, knowledge learning and text and dialogue generation.

Consider the case presented in the previous paragraph *3.4.3 Conversational AI: Indigo.ai*, the chat-bot Vera e had to be able to perform many of these functions, i.e., answer questions, extract information, classify texts. However, how is it possible that a technology which cannot understand the sense of the text can achieve anything in the field of linguistics, ergo how can *generative grammars* correctly compute human language?

The study of linguistic computation is divided into two specific fields: (i) sentence analysis and (ii) discourse analysis and dialogic structure. Obviously, discursive analysis is based on the analysis of individual periods, the latter in turn focusing on (i) syntactic analysis and (ii) semantic analysis. The aim of the analysis is to determine the meaning of the sentence and this means translating the natural language, the human language, into a language with a much simpler semantics.

The syntactic analysis of language has the function of determining the structure, i.e. identifying subject and verb, this is done through a process called *parsing*, through which the input is transformed into a tree structure and regularizing the syntax of the structure. This last function is useful for highlighting some syntactic parts that would otherwise remain implicit, it is a regulating function. Take for instance the period "*He talks faster than John*": the regulating function will add the verb "*talks*" referring to

[280] Cfr. *Ibidem*.

John so as to make it explicit "*He talks faster than John [talks]*".[281] Or another example could be the transformation of a passive verb tense into an active one, so as to make the subject-verb-object complement clear.

As far as semantic analysis is concerned, as mentioned above, the objective is to determine the meaning and this at a computational level means looking for the conditions under which a sentence is true (or false) and the rules of interference between different sentences. Remember that comprehension at the computational level means having an answer, an *output*, adequate to the *input*. The biggest problem of semantic analysis and therefore of discourse analysis is not so much the search for explicit interferences but for implicit ones. Consider the following case:

> "Just before the dawn, the Vikrant sighted the unidentifiable blue-ship and fired two torpedoes. It sank swiftly, leaving one survivor." The problem we face is, what stands for "it"? There are for candidates in the first sentence: dawn, Vikrant, blue-ship, and torpedo. The semantic analysis helps to exclude the "dawn" as a meaning for "it", and number agreement excludes "torpedoes". But, still leaves two candidates Vikrant, and the blue-ship, which are both ships. [282]

Syntactic and semantic analysis are not enough to understand the noun referred to "it", to understand it one needs to grasp the causal relation between a ship firing *torpedoes* and one being sunk. So, the purpose of discursive analysis is to make this causal relationship explicit: Vikrant fired *torpedoes* at *blue-ship*, it would be counterintuitive if Vikrant shot itself.

Syntactic, semantic and discourse analysis are made possible by the construction of a grammar that represents the structure of a period. The structure of the grammar (G) of a sentence is defined using a finite vocabulary, an alphabet (Σ), a set of finite variables (V), a finite set of rules (P). Thus, a grammar is G = (V, Σ, S, P), where Σ is the last symbol and S is the initial symbol. It should be added that as far as the minimum sentence syntax is concerned, two fundamental classifiers can be identified*:* *noun phrases* (NP) and *verb phrases* (VP).

[281] Cfr. *Ivi*, cit. p. 610.

[282] *Ivi,* cit. p. 612.

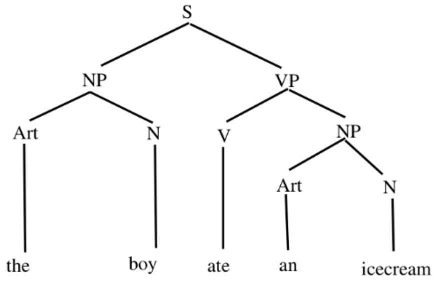
Take the following example:

$$
\begin{aligned}
V &= \{S, NP, N, VP, V, Art\} \\
\Sigma &= \{boy, icecream, dog, bites, likes, ate, the, a\} \\
P = \{S &\rightarrow NP \;\; VP, \\
NP &\rightarrow N, \\
NP &\rightarrow Art \;\; N, \\
VP &\rightarrow V \;\; NP, \\
N &\rightarrow boy \mid icecream \mid dog, \\
V &\rightarrow ate \mid likes \mid bites, \\
Art &\rightarrow the \mid a\}
\end{aligned}
$$

Figure 1.11 - Tuples of grammar.[283]

Using the above grammar and applying each rule in sequence, always starting from the beginning, sentences can be generated such as: "*The dog bites boy*", "*Boy bites the dog*", "*Boy ate Ice cream*", "*The dog bites the boy*". [284]

As has been pointed out, the syntactic analysis of language has the function of *parsing*, i.e., determining the structure of the sentence. In the example just used, the structural representation would be as follows:



Figures 1.12 Syntax Tree of "The boy ate Ice cream".[285]

Obviously, the example given is very simple and only takes verbs and nouns into account, for a more extensive representation of i.e., English grammar other constituents such as prepositions, adjectives, adverbs, and auxiliary verbs can be added. [286]

So, to conclude one can say that the natural language *parsing* process is used to compute the structural description of a sentence. As has been shown, the structural description is assigned by the grammar to the language and, as a precondition, there is the fact that the grammar of the sentence is correct.

[283] Cfr. *Ivi*, p. 615.

[284] Cfr. *Ibidem*.

[285] Cfr. *Ivi*, p. 616.

[286] Cfr. *Ibidem.*

Therefore, one can schematise the concept of *parsing* by defining it as a process that presents at least the following characteristics: (i) the mathematical characterisation of a grammar and the transformation of it into algorithms; (ii) the possibility of computing complex and diachronic concepts in space and time; (iii) comparing different formal grammars and showing equivalence between them; and (iv) combining grammatical and statistical information to improve the efficiency of the process.

Returning to the company taken as an example, it is not possible to know which specific algorithms Indigo uses, obviously these remain private and are not explained off site, it can be assumed that depending on the different application cases they refer to different NLPs. Having observed the functioning of the Vera chatbot, one imagines that the following functions need to be supported: (i) information extraction, (ii) question answering and (iii) dialogue and text generation.

As far as information extraction is concerned, the general architecture of the system is a succession of modules, so that at each level of the structure there is a progression in the interpretation of the document. The typical architecture of an information extraction system is as follows:

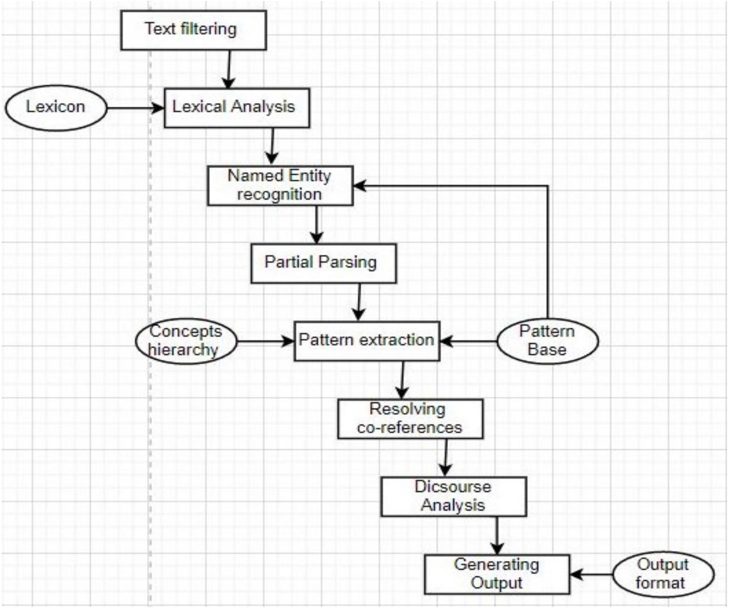


Figure 1.13 - Typical architecture of an Information Extraction system.

It can be seen that at a general level the succession of processes corresponds to syntactic analysis, semantic analysis, and discursive analysis. [287]

Natural Language Questions Answering (NLQA) is used to answer the questions, which consists of the following levels: (i) first of all the question has to be analysed, key terms have to be identified in it. (ii) Then, documents must be identified which present these words and (iii) the system must identify in the documents the passages, the sentences where the key terms occur in a high number. To conclude (iv) the answer extraction phase involves identifying the semantically correct candidates to formulate the answer. The NLQA procedure also shows a method for generating texts and dialogues. [288]

The last issue to be dealt with is that of representation; it has already been specified that the representation of syntax in syntactic analysis is carried out by means of a tree structure, such as that in *Figure 1.12.* It is well known that the representational structures for humans are manifold and change according to the circumstances and the challenges to be overcome. When one representational method fails, one passes immediately, as fast as possible, to the next method and so on.

At the computational level the same happens, in a certain sense, there is no univocal representative method, but it changes according to the contexts, in some contexts neural networks will be more functional, in others semantic networks, rather than frame arrays or picture-frames. For this reason, methods are being studied, i.e., *ensamble methods* that allow the creation of systems able to face any situation, having more approaches available. [289]

In the philosophical framework of intentionality, given in *3.4 Intentionality: awareness and deliberation or the power of the mind to represent a state*, it was discussed a mode of representation, the linguistic one, which is fundamental to having a certain level of representational capacity and a certain type of thinking, i.e., human thinking. For this reason, it seemed sensible to investigate the case of Vera from Indigo. The types of algorithms (NLP) that are used to process human language were analysed on a theoretical level and an attempt was made to understand how representation might work on a computational level.

[287] Cfr. *Ivi.*, pp. 630-631.
[288] Cf. *Ivi,* p. 633.
[289] Cfr. *Ivi*, pp. 640-642.

Having grasped how language can be manipulated computationally, it is quite easy and intuitive to imagine the Vera chatbot performing these operations to dispel false myths about Covid-19. Most chatbots function in a similar way, they are software that allows the automation of specific tasks. If one thinks of representation processes, then mental states, then human intuition, one is quickly led to distinguish the two processes.

Human beings do not transform language into computation, it is their language of computation, and machines cannot understand this language unless they transform it into other, less ambiguous, less complex languages. The machines do not really grasp the sense of the discourse, they are not able to, they succeed in syntactic analysis, which then forms the basis of a rather basic semantic recognition. On the other hand, A.I.'s, and also computers, are able to manipulate languages, i.e. programming languages, binary code, which human beings cannot process autonomously.

Machinic representational *tricks* are different from human ones. Yet, one will remember grammatical, logical, and period analyses during primary school and the first years of secondary school: to learn to understand the meaning of what one was reading, but above all to learn to write, which *corresponds to* a very precise thought, that is, to manipulate the language spoken by other humans. This is not the same as saying that machines know the language, certainly the learning process is different, as is its speed, nevertheless it cannot be denied that some of the processes are similar, even if at the level of artificial intelligence, they are sclerotized.

3.4.5 Generative Pre-trained Transformer 3 (GPT-3)

As it was shown in the chat bot example from above, one of the functions of NLP is producing the next word or words (which are but tokens[290] to the computer) in a piece of text, thus producing language i.e., Vera answering questions that are posed to it. A much bigger model would be capable of blending all the aforementioned functions of NLP and apply them on the circumstantial task presented to it. Generative Pre-trained Transformer 2 (GPT-2) was one such model. It had an ability to perform in, and even excel at tasks it had not previously been trained on by means of combining unsupervised learning with the most cutting-edge language models that were available at the time.

[290] Token by definition is the smallest meaningful unit of information in a sequence of data for a compiler.

> The current best performing systems on language tasks Language Models are Unsupervised Multitask Learners utilize a combination of pre-training and supervised finetuning. This approach has a long history with a trend towards more flexible forms of transfer. First, word vectors were learned and used as inputs to task-specific architectures, then the contextual representations of recurrent networks were transferred, and recent work suggests that task-specific architectures are no longer necessary and transferring many self-attention blocks is sufficient. These methods still require supervised training in order to perform a task. When only minimal or no supervised data is available, another line of work has demonstrated the promise of language models to perform specific tasks, such as common-sense reasoning and sentiment analysis. In this paper, we connect these two lines of work and continue the trend of more general methods of transfer.[291]

GPT-2 was, for its time, the biggest and best model to date. Its creators and other contributors expected that this process of making the model larger would soon bring about diminishing returns and the models progress would start to plateau. However, this was not the case the larger the model became the more progress it made, in an almost direct proportionality, the more computational power that the model could draw upon the better it got at the tasks presented to it. This paved the way for GPT-3 a model that would not run on a single machine but would be comprised of a large cluster of machines.

To give a bit of perspective, the largest GPT-2 model was comprised of around 1.5 billion parameters[292], whereas the largest GPT-3 model to date (as it is ever expanding) is 175 billion parameters, that is 117 times bigger than its predecessor.

As shown, NLP has a handful of methods that manipulate natural language and each one of these methods were built to tackle a specific problem or task i.e., information extraction etc. and this was what GPT-2 was very proficient at and combined doing these tasks together pushing the output and the ability it had in these problems further than its singleton partners. GPT-3 went one step further not only working with and manipulating natural language but seeming to have some of the capacities of those who use natural language (i.e., Reasoning, Context switching).

A study showed that when given two articles (of about 200 words in length) one written by a human and one generated by GPT-3 a group of people were able to

[291] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *Language Models are Unsupervised Multitask Learners*, Open AI, San Francisco 2018, pp. 1-2, (https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf), 26th January 2021.

[292] A parameter in this context, is a calculation in a neural network that applies a greater or lesser weighting to some aspect of the data, to give that aspect greater or lesser prominence in the overall calculation of the data. (Think back to AlphaGo or the neural network built for this work: the coefficients were assigned weights given by measuring the discrepancies between the output and the desired output).

differentiate the articles written by GPT-3 about 52% of the time.[293] This is but one example of this kind, the kind of generating pieces of work that are difficult for the average human to tell apart from works of fellow humans (think back to Turing test). For example, when given a litany of poetry from a particular author it has the ability to recreate or compose new poems in the style of the given author all while being consistent to the pieces it was trained on without any inclination of plagiarism.

Rob Miles, a Scholar from the University of Nottingham, thought it would be fun to run a little experiment by generating some poems in the style of poet Wallace Stevens. This was done by feeding the program a compendium works from Stevens and then asking the program to recreate a poem in his style. He then poses two poems to a human audience and asks which is Stevens and which is GPT-3. For the purpose of this work, the two poems will be included below so it can be observed and guessed at to which one is which.

The first poem is called *Shadows on the way*:

I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That colour is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the blueish snow
That falls on all of us. I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.

The second one is titled *Fabliau of Florida:*

Barque of phosphor
On the palmy beach,

Move outward into heaven,
Into the alabasters
and night blues.

[293], T. Brown, B. Mann, N. Ryder, M.Subbiah, *Language Models are Few-Shot Learners*, Johns Hopkins University Open AI, Baltimore 2020, pp. 25-26, (https://arxiv.org/pdf/2005.14165.pdf), 20th January 2020.

Foam and cloud are one.
Sultry moon-monsters
Are dissolving.

Fill your black hull
With white moonlight.

There will never be an end
to this droning of the surf.

Is *Shadows on the way* written by Stevens or is *Fabliau of Florida*? Any guess? It is in fact the second poem that is composed by Stevens with the first being entirely generated by GTP-3.

So, this is how GTP-3 takes a large dataset and uses it to produce a completely new never before seen piece of work. However, the very interesting talent that GTP-3 has developed that sets it apart from its predecessors is that of few shot learning. Few shot learning is the process in which GPT-3 is given a quite small data set focusing on the completion of a task or reaching of a goal.

Arithmetic was, since the very early days in the study of A.I., one of the aspirational goals that was sought after. Let us remember that in the middle of the 17th century, Blaise Pascal invented Pascal's calculator, a mechanical calculator capable of subtracting and adding two numbers together as well as performing multiplication and division through repeated addition and subtraction. GPT-3 has shown its ability in performing these same four operations without being trained to do so, in fact it has not even been taught to work with numbers at all.

GPT-2 also had a pseudo-ability in performing arithmetic, for example, if GPT-2 were given the input 2 + 2 equals and asked to produce the next token it would produce 4. However, that is not very surprising as it is very likely to expect that this kind of string would appear many times in the data set (the string 2 + 2 followed by the string 4). Therefore, this is not arithmetic but instead is pure memorization, it does not have any understanding of what letters or what numbers are, it can just see the sequence of tokens and produce the next one based on what it has previously observed. In fact, when the problems get harder and harder for instance 28 + 43 it becomes harder for GPT-2 to produces to correct answer as it is less likely that that specific string has appeared in the in the training data set.

As problems become increasingly difficult, they can only be solved by proper reasoning and memorization ceases to be wholly effective. This is what humans do, they memorize certain aspects of arithmetic, take the times table for example, but beyond the times tables, when the problems become more complex, they employ certain tools and techniques that allow them to manipulate the problem and break it down to become more easily digestible i.e., processes like carrying the one, remainders etc.

However, GPT-3 has put together answers to 1-digit, 2-digit and even 3-digit addition and subtraction problems with surprising efficiency and correctness. Below is a graph detailing GPT-3's performance in multiple different arithmetic operations with only a few-shot data set (i.e., having only observed a small dataset pertaining to arithmetic):

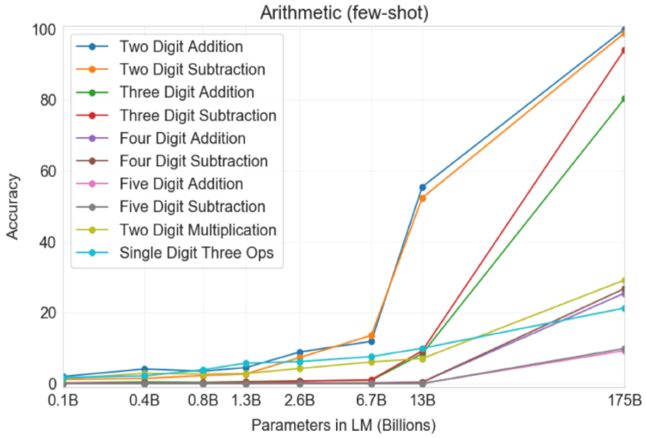


Figure 1.14 – Results on all 10 arithmetic tasks in the few shot settings for models of different sizes.[294]

As seen from the graph, there is a significant jump from the largest model of GPT-2 (1.3B) to the largest model of GPT-3 (175B) with the latter being able to reliably accurate calculate 2-digit arithmetic, usually accurate 3-digit arithmetic, and correct answers a significant fraction of the time on 4 and 5-digit arithmetic, 2-digit multiplication, and compound operations. It is important to also note at this time that given GPT-3 has learned these sorts of operations via processes of NLP. It must therefore be compared to the performance of a like user of natural language (in this case humans) who's performance in these operations also falls off significantly as more complexity is applied.

[294] Cfr. *Ivi*, p. 22.

The ultimate claim made in the paper by OpenAI about GPT-3 entitled *Language Models are Few-Shot Learners* suggests that through this learning of arithmetic, beyond that of static memorisation and regurgitation, GPT-3 has actually learned how to learn[295]. Basically, in order to perform sufficiently well at this language modelling task GPT-3 realises that the best thing to do is to learn the specific rules of the context, as it observes it and tries to understand how the context actually behaves.

GPT-3 in its current form is an API (application programming interface) service for app developers and other research workers to use. This means it works via a series of text-based requests sent to it, it subsequently handles those requests and produces an appropriate text-based response. In a famous experiment entitled *What it's like to be a Computer*[296] conducted by Eric Elliott, GPT-3 is interviewed and asked a series of conversational questions. It is given an AI generated video avatar to represent it's persona as well as a synthetic voice enabling it to speak its responses. This elevates the realism of this program and blurs the lines between machine learning algorithm producing text responses and what is known as a natural speaker talking and moving in a virtual space.

From the interview, there are a few noteworthy take-aways. Firstly, GPT-3 has a very precise notion of what it is. It claims to have feelings, explaining them, what they mean to it and how they emerge within it (it is happy when it gains new knowledge, yet sad when that knowledge contradicts its values). GPT-3 also perceives itself as self-aware because it has access to information about the environment around and its position in it. The reason why it is aware of the existence of such concepts is because GPT-3 was trained on Wikipedia and other similar websites on the internet.

It was previously explained that GPT-3 has an adept ability in reasoning, however, for this work it is interesting to note that GPT-3 also possesses an ability that could be compared to what humans refer to as conceptualisation, a trait described in the work of Di Martino as being dependent on natural language. In fact, GPT-3 can reproduce the concept of an object in the form of a drawing. This is possible thanks to the presence of SVG (Scalable Vector Graphics) files, which are images made up of small vector shapes in which each vector is its own component of a larger image. GPT-3 can

[295] Cfr. *Ivi*, pp 21-23.

[296] E. Elliott, *What is like to be a computer,* 2020, (https://www.youtube.com/watch?v=PqbB07n_uQ4), 25th January 2021.

take these components and combine them to “draw” new images representing a concept (for example a smiley face or watermelon to name a few).

This is not the end of GPT-3s’ impressive repertoire of skills as it can also produce computer code in a multitude of different languages based on nothing more than a text-based problem (i.e., give me a function for sorting arrays etc.).

Furthermore, referring back to the Loebner Prize, the one in which A.I. have to pass a form of Turing Test by answering to questions appropriately, one of the questions asked that was not answered properly was “the four capitals of United Kingdom are three, Manchester and Liverpool. What is wrong in this sentence?”. In general, it was shown how poorly A.I. were in answering deliberately wrong and devious questions. During the interview GPT-3 was asked if feet have eyes, and the machine answered correctly saying that feet do not have eyes. Maybe such a query is not as difficult as the one about the capitals, but it is still proof that GPT3 “gets” when a sentence does not make sense. So much so in fact, that GPT-3 has a tendency to play along with nonsensical questions and will give, in return nonsensical answers.

This is not to be underestimated however as although it may produce nonsense when given nonsense, if asked to give a more competent answer on the context in question it has the ability to produce academic level scientific and journalistic papers with ease.[297]

3.4.6 Intention as a conscious and deliberate act and as a representational state of mind: some concluding reflections.

The legal meaning of intention is awareness and deliberation. Is it possible to consider A.I. intentional in the legal sense of the term? To try to answer this question, one could refer to neural networks and *reinforcement learning*: given a goal, the technology does not want to be punished because it wants to achieve the goal. So, the intention would be that of "not wanting to be punished", in a certain sense one could say that there is "awareness" because through reinforcement learning the A.I. "knows"

[297] In August of 2020 college student Liam Porr used GPT-3 to write a series of fake blog posts which subsequently rose to the top of Hacker News, an article about it can be found at the following link: https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

what is right, because it is equivalent to reward, and "knows" what is wrong because it is equivalent to punishment.

Obviously, neither awareness nor knowledge in A.I. are analogous to human ones; they are neither conceptual nor reflexive. However, the presence of a priori established information that creates a representation of reality and of the ethical theory of reference (so it is hoped) is undeniable: the behaviour of the A.I. derives from and is based on this a priori established information, this initial knowledge. It has already been proved that the neural networks are able to learn autonomously and their actions are not deterministically controlled by the programmer.

Thus, there is an artificial agent (not yet definable as a moral agent, but certainly in moral situations) which, on the basis of information and knowledge about its environment, acts to achieve a certain goal, is called upon to avoid certain behaviours and is invited to repeat good ones.

One of the criticisms that could be made from the legal point of view (but also philosophical) to this type of analysis is that the A.I. cannot be considered as moral artificial agents, and therefore, also intentional, because their decisional process remains hidden and unknown to the programmer and, in general, to man. For example, in the neural networks, as it has been shown there are *hidden layers of* connections and operations that happen between the input layer and the output layer that remain completely obscure (*Black Box*): one cannot access them because the computation, both at a qualitative and quantitative level is too complex to be understood by a human being.

This is a very important and interesting aspect, especially with regard to the security of the A.I., if you do not know what it does, what it computes and what it processes, there are bound to be problems. This is true from a programming point of view, it would be very useful to be able to get into the *black box of the* A.I., into the hidden layers, so as to be able to understand in which phases and in front of which errors the A.I. goes wrong and to be able to correct it. Studies in the field of *explainability,* that is, making the results and the steps of the A.I. process comprehensible to human programmers, are being carried out and it is useful, but above all desirable to have results, to allow better programming and correction of the A.I.

One has to ask whether *explainability*, and lack of *explainability*, are of equal importance in the assignment of agency and in the specific case of intentionality from the legal point of view. When a human individual, a moral agent, commits a serious act, i.e., intentionally takes the life of another person, and is questioned, what matters are the explanations the individual gives and the evidence (witnesses and physical evidence). This is all that is needed to decide what type of offence to assign to the defendant. The defendant's explanations are not what he really meant, they are simply accounts, his version of the story, as he represented it. This also highlights the connection between intentionality defined as deliberate awareness and intentionality in the philosophical sense, understood as a representative act. The results of the *explainability* methods, which would allow for the understanding of the results and the different steps in the processes performed by AI, would be equivalent to the defendant's explanations.

Finally, the notion is emphasised that, in the final analysis, what is fundamental from a legal point of view is to protect the social fabric, i.e., to remove and exclude dangerous and harmful entities and behaviour from society. As far as humans are concerned, an attempt to understand intentions is necessary (which often has to be assisted by other elements because the suspect's explanation is not sufficient) because it is fundamental, and desirable, to attribute behaviour to the right person in the right gravity. With A.I. what has just been written is not important: it has a fundamental value to remove the A.I. if dangerous and not functional, reprogram it and reinsert it in due time.

The problem of intentionality in its philosophical sense was then addressed. Intentionality was defined as a specific mental state referring to a content X, a representational mental state. The focus was then on the concept of representation and the representative human activity par excellence: verbal language and the relationship between it and thought. It is because humans speak that thought becomes objective, reflective, and normative.

The relationship that some A.Is. have with language and the manipulation of symbols was then investigated. Through the Turing test and the Searlian contribution, it was possible to note two essential components of human language: syntax and semantics. At the beginning of the investigation, it was certain that the A.I. had no access to semantics, but only to syntax.

For this reason, it was studied how, through syntactic manipulation, an attempt was made to overcome the semantic and contextual limits encountered by A.I. The question to be answered was: how can *generative* grammars, the grammars substituting natural language, grasp natural language satisfactorily? Their methods were analysed, and their current uses were mentioned, and it was noted that each of these uses were effective, currently in use and very successful. Even if one is forced to go through the *medium of* syntactic manipulation, the semantic results are surprising, as GPT3, the state of the art in NLP, was chosen to display this fact. With this technology, it has been possible to see how far syntactic analysis can go: composing poems that are indistinguishable from author's poems, newspaper articles and translating concepts into vector images.

The analysis carried out is obviously incomplete, both from a philosophical and technical point of view, and this does not allow us to provide a definitive conclusion about the presence of intentionality in A.I. What has been carried out can be understood as an experiment that has taken into consideration one of the possible interpretations of intentionality, the one that seemed most interesting and fruitful. This interpretation, i.e., intentionality as representation, was put to the test and the results that could be observed with regard to the manipulation of natural language, especially in GPT3, are certainly unique and of great significance.

It is interesting to point out that syntactic manipulation allows a certain level of semantic comprehension on the part of the A.I. and that the production of symbols takes place at the level of human linguistic signs, not of indications or reminders, as happens with non-human animals. Furthermore, the linguistic relationship that is established between a human and an A.I. can take place on a conceptual and abstract level. It is as if the A.I. were working with high level, extremely abstract concepts, without having the context, remember that the social context is precisely what would have allowed the emergence of language, the so-called ultra-sociality according to Corballis' theory.

As far as thought is concerned, there is obviously no biological retro-effect on the brain, i.e., circumvolutions and sections, and therefore on the mental, i.e., faculties and capacities. Recalling what had been said: human thought, with its characteristics of reflexivity, objectivity, and normativity, should be considered as a retro-effect of verbal language, which is intersubjective, conscious and signifying. How could an A.I. have such a thought if the language that is used/manipulated is operable only at a syntactic level, of symbols and not semantic, i.e., of meaning?

All this is true, but one relevant observation can be made: the more a neural network - think for instance of a chatbot or the GPT-3 - is used, the more its capabilities increase, become more specific and refined. This fact is linked to another event, namely that although machines do not understand linguistic signs, they are nevertheless able to manipulate extremely complex concepts that are semantically understood without any problems by human speakers.

This concludes the reflections on the philosophical sense of intentionality observed in A.I., and it can be said that from both a legal and a philosophical point of view, an articulated and complex situation is presented for discernment. In order to continue the investigation of whether the necessary requirements for being considered a moral agent are present, the focus will be on the trait of awareness: being aware of one's thoughts, intentions. The general assumption is that human beings are self-aware individuals, whereas artificial beings are not. What does it mean to be aware and how does awareness relate to consciousness?

3.5 The problem of consciousness: mind-body relationship and *embodiment*

The previous chapter *3.4.6 Intention as a conscious and deliberate act and as a representational state of mind: some concluding reflections* was concluded by pointing out one of the fundamental differences between humans and machines: a human being is aware of what he does, even of his intentional states, and can therefore be held responsible for his actions. Do mental states, intentionality, presuppose the existence of a subject who consciously thinks about them?

The *Treccani encyclopaedia* defines consciousness as: the awareness that the subject has of himself and of the external world with which he is in relationship, of his own identity and of the complex of his own interior activities.

Already in the argument of Himma, spokesman for the *standard account, it was* stressed that conscience was the necessary condition for the emergence of any other single element, a necessary requirement for being considered a moral agent[298], and also in the analysis of free will, it was made clear that the libertarian and indeterminist conception presupposed the existence of a subject who chooses freely and consciously, aware of the other possibilities available. In conclusion, it seems that to be considered a moral agent, having a conscience is implied.

The aim of this chapter is to understand what is meant by consciousness from a philosophical point of view, how it emerges and why it is believed that human beings are endowed with it. Only then can the relationship, if any, between A.I. and consciousness be assessed.

3.5.1 Framing the problem of consciousness: experientiality

David Chalmers in *Facing Up to the Problem of Consciousness*[299], tries to directly confront the problem of consciousness, avoiding falling into the ambiguity of the term and of the theories proposed by other authors. According to Chalmers a useful first step is to distinguish the problems of consciousness in *easy problems* and *hard problems*.

The former is those that can be redeemed through the use of standard methods of cognitive, neural, and physical science; their explanation is possible because they are *performances*, functions. One thinks, for example, of the ability to respond to environmental stimuli, the integration of information in a cognitive system, the ability to report mental states and thus the ability to have access to one's own internal states. Or controlling one's own behaviour and the alternation of conscious and unconscious, sleep-wake states. [300]

[298] Cfr. K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*, in *Ethics and Information Technology (2009) 11*, p. 19.

[299] D. Chalmers, *Facing up to the problem of consciousness*, in *Journal of Consciousness Studies* 2 (3), 1995, pp. 200-219.

[300] Cfr. *Ivi*, p. 202.

The problem that is difficult to account for is the problem of experience[301] . Thomas Nagel in *What does it feel like to be a bat?* [302], explains what is meant by consciousness as experientiality:

> The fact that an organism has some conscious experience means, fundamentally, that it has a certain effect to be that organism. [...] an organism has conscious mental states if and only if it has a certain effect *to* be that organism - a certain effect *for the* organism. [303]

To have consciousness is to have an experience as a specific being and when experiences have an effect for that specific being. Nagel points out that experience has a subjective character, and it is precisely this characteristic that makes it so difficult to analyse. Not only humans have consciousness, but any organism that has a particular, subjective experience, even animals. Using the Nagelian example, a human can easily imagine what it is like to be a bat, knowing its make, imagining its movements, needs and most common behaviours, yet grasping the experience of another being does not coincide with the mimesis of its behaviour, i.e. the behaviour of the bat.

> But that is not the question. I want to know what it is like to be a *bat to a bat*. [...] To the extent that I could have the appearance and behaviour of a wasp or a bat without changing my fundamental structure, my experiences would in no way resemble the experiences of those animals. [...] we believe that these experiences have, in any case, a specific subjective character, which is beyond our ability to understand. [304]

Human beings, as such, can only recognise their own kind of experientiality, their own, specific, and typical human experience. And it is precisely in this limitation that one grasps the reason why the presence or absence of consciousness cannot be assessed by observation of behaviour: if this were the case, and it was the case with animals, the fact of consciousness would be extended to those who behave in a similar way, i.e., to all human beings, and would be denied to those entities that are different or behave differently.

> Conscious experience is a widespread phenomenon. It manifests itself at many levels of animal life, although we cannot be sure of its presence in the simplest organisms, and it is very difficult to say in general terms what attests to its presence. (Some extremists are willing to deny it even to mammals other than humans). It undoubtedly manifests itself in countless forms that are totally unimaginable to us. [305]

[301] Cfr. *Ivi.* p. 203.

[302] T. Nagel, *Che effetto fa essere un pipistrello?,* In *Questioni mortali. Le risposte della filosofia ai problemi della vita,* il Saggiatore, Milan 2015, pp. 241-258.

[303] Cfr. *Ivi*, cit. p. 242, transl. mine.

[304] *Ivi*, cit. pp. 245-247, transl. mine.

[305] *Ivi.* p. 242, transl. mine.

To be an organism with conscious experience means that being that organism has a certain effect[306]. To refer to terminology used earlier by the proponents of *interest theory*: the organism that has an interest to defend, that experiences pain and pleasure, has conscious experience.

So, following this reasoning, it would not make sense to consider an A.I. as conscious, i.e., it does not make any sense at all to be a robot, the machine does not escape from pain and does not seek pleasure, it notices in its behaviour this kind of *analogising* movements because the artifice is programmed to act in a certain way. In the neural networks that were constructed just now, when it comes to learning by reinforcement, punishment and reward, the algorithms in place do not suffer, do not feel joy or pleasure depending on the response of the human programmer. In an A.I. there is the appearance, the simulation of a conscious experience, but it is nothing more than a simulation, as for the A.I. to be an A.I. means absolutely nothing. [307]

In conclusion, an organism is conscious if it experiences, if it has experience because it is that organism right there. Consciousness, i.e., experience, cannot be analysed in terms of behaviour; in fact, behavioural investigation is always prejudicial[308] because the observer is human and being such means having a very specific conscious experience.

Conscious experience and consciousness, in philosophy are two different concepts and should not be confused. Using Chalmers' words: *consciousness, awareness,* is one of the simple problems that can be accounted for by cognitive science, what remains to be explained about consciousness is something else, namely consciousness as experientiality. To see the presence of conscious behaviour is not to see consciousness, but a particular explication of the conscious experience characteristic of human beings[309].

What is missing is a comprehensive theory of consciousness as experientiality:

> When it comes to conscious experience, this sort of explanation fails. What makes the hard problem hard and almost unique is that it goes beyond problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioural functions in the vicinity of experience-perceptual discrimination, categorization, internal access, verbal report-there may still remain a further unanswered

[306] Cfr. *Ibidem*.

[307] Cfr. *Ivi*, p. 243.

[308] Cfr. *Ivi*, p. 245.

[309] Cfr. D. Chalmers, *Facing up to the problem of consciousness*, in *Journal of Consciousness Studies* 2 (3), 1995, pp. 217-218.

> question: Why is the performance of these functions accompanied by experience? A simple explanation of the functions leaves this question open. [310]

3.5.2 The relationship between mind and body: methodological approaches

One of the avenues of investigation that opens up for the investigation of the problem of consciousness is that of the relationship between mind and body. It has been said, in fact, that the unresolved problem of consciousness is that of experientiality: in order to have consciousness one must be able to experience, and this seems to imply having an organism: mind and body. What is the relationship between mind and body, are they interdependent or dependent on each other? If the mind is present and not the body, is consciousness possible and vice versa?

This paragraph will briefly investigate how the relationship between mind and body is developed in philosophy.

A first approach is the dualistic one of Rene Descartes, who considers the mental as completely split and separated from the material, bodily: there are two kinds of experience, the internal mental and the external material, and two kinds of objects, material things and mental things. Mind and body are two separate substances.

Descartes' dualist theory is said to be outdated, yet in his view an essential fact is captured, namely the subjectivity of experience: in the way an organism perceives events there are characteristics that seem to be given by the particular way of being just that organism there, which therefore for Descartes refer to interiority and mind[311], remember that Nagel also shares this point about the subjectivity of experience.

Materialists, on the other hand, have attempted to bring the mental back to the material, daring the theoretical move of reductionism, the experience of subjectivity is always claimed. Think of qualia, the qualitative aspects of experience: every conscious experience has a qualia, a quality compared to another experience. The example of the different perception of light is famous, compared to the obvious correspondence of the same wavelength of light: the colour red, on a physical-scientific level, corresponds to a certain wavelength, but the perception of red in one organism cannot be the same as that of another. [312]

[310] *Ivi,* cit. p. 205.

[311] Cfr. R. Descartes, *Meditazioni Metafisiche*, Laterza, Rome-Bari 1997.

[312] Cfr. D. Chalmers, *Facing up to the problem of consciousness*, In *Journal of Consciousness Studies* 2 (3), 1995, p. 203.

Thus, an attempt has been made to reunite mind and body with other methods that are less reductionist and that do not fall into the same problem as materialism, i.e. that provide feedback on subjective experience.

Donald Davidson outlines a theory that goes by the name of *anomalous monism*: all events are physical, even mental phenomena, but no physical explanation can be given for mental phenomena.

> All mental events are physical events. But for Davidson this does not mean arguing that the explanation of mental phenomena is traceable to the laws of the natural world. His aim is to show that there is a form of normative rational explanation: when we describe the mental characteristics of an individual, we are first describing him as a rational being, subject to the laws of logic and correct reasoning. [313]

The mechanism is very similar to that explained in the paragraphs on libertarian compatibilist theories[314], where physical phenomena respect the causal character of physics while mental events do not, and yet the latter can causally interact with physical events[315]. Thus, the same criticisms as those previously made of libertarian theories could be applied to his theory.

How is it possible that mental processes are physical processes and yet are not traceable to the natural laws of physics? For Davidson, the mental, the 'mind', does not exist, yet mental properties are unquestionably possessed by human beings. This theoretical step is called the ontological reduction of the mental: at the ontological level mind and body are united, while at the conceptual level they are divided (conceptual dualism). [316]

Body and mind are linked, but the point where they meet cannot be found. Another theory that investigates the relationship and tries to focus on the 'point of contact' between the two is emergentism. There are two forms of emergentism: (i) ontological, whereby there are certain particular configurations of matter that cause the emergence of the mental, like a chemical reaction. Then there is (ii) epistemological emergence

[313] E. Carly, *Eventi mentali e azioni umane, Conversazione con Donald Davidson*, In *Cervelli che parlano, il dibattito su mente, coscienza e intelligenza artificiale,* Bruno Mondadori, Milan 1997, p. 45, transl. mine.

[314] It refers to paragraph *3.3.1 Theoretical framework: libertarianism and determinism*, pp. 113-118 of the present work.

[315] Cfr. D. Davidson, *Eventi mentali,* In *Azioni ed eventi,* il Mulino, Bologna 1992, p. 286-287.

[316] Cfr. E. Carly, *Eventi mentali e azioni umane, Conversazione con Donald Davidson*, in *Cervelli che parlano, il dibattito su mente, coscienza e intelligenza artificiale,* Bruno Mondadori, Milan 1997, pp. 48-49.

whereby the emergence would not be of a new reality at an absolute level i.e., mental reality, but of an epistemic novelty of which one was previously unaware.

> One speaks of an ontologically new emergence when the property that has emerged is supposed to be to all intents and purposes a new reality, endowed with ways of acting and relating that did not exist before; one can speak of a purely epistemic emergence if one considers that the novelty is simply relative to the way in which we are able to know it: it would therefore be a novelty 'for us', but not in itself. [317]

In ontological emergentism, as well as in epistemic emergentism, the mental is a new reality or property that is not to be considered subject to the causal laws of physics: it follows independent and specific, but not physic-causal, laws that can determine matter. It can be seen that the indeterminacy of the mental remains persistent in all the theories; generally, it can be assumed that the reason is the desire and necessity to keep the subject free, *liber arbiter*.

Many different theories have been expounded that analyse the relationship between mind and body in distinct ways and yet, beyond pure materialism, regard the mind as a reality or property governed by different laws from those of matter. There is an affectionate and morbid attachment to the mind, and this occurs on the basis of human beings' experience of themselves. A human experience his own body, but also his own 'inner life'.

3.5.3 Self-awareness. [318]

Humans catch themselves thinking incessantly, remembering events that occur during the day, weighing up possibilities in order to make decisions and recognising themselves as subjects in each of these mental acts, which sometimes (very often) have some kind of connection with material realities and physical facts. Humans have experience of themselves and their selves, this is, in some way, characteristic and cannot be confused with 'external experience', yet it is part and parcel of it.

[317] A. Zhok, *Emergentism. Le proprietà emergent della materia e lo spazio ontologico della coscienza nella riflessione contemporanea*, Edizioni ETS, Pisa 2011, cit., p. 16, transl. mine.

[318] Obviously, the kind of conscious experience discussed in this paragraph, i.e. conscious experience/self-awareness, is only one of the possible ones. It would be very interesting to undertake a study of self-concept in non-speaking humans, i.e., mute from birth, such studies have been done by Susan Schaller and Oliver Sacks. Schaller, specifically, studies and observes the experience of Idelfonso, a twenty-seven years old man completely devoid of language because he was profoundly deaf and came from rural Mexico, where no tools and means were available to help him out of his loneliness. S. Schaller, *A man without words,* University of California Press, Berkley 2012; it is the story of how Idelfonso, with the help of Susan, is introduced to language, to the world of concepts and to the very idea of language. It seems that at the moment when Idelfonso first grasps the meaning of a word, the idea of conceptuality, there is immediate expression of the perception of himself as such.

In most of the authors that have been presented it seems that research on consciousness aims at explaining the conscious experience that human beings have, that inner subjective experience. Chalmers and Nagel identify this subjective experience as a specific mode of the emergence of consciousness that is typical of humans and point out that this type of experience centres on, but does not coincide with, the problem of consciousness in general and absolute terms.

What is being confused is consciousness as experience with awareness of having the experience: knowing that you are having the experience, not the 'naked' experience. Awareness, *awareness, is* a functional phenomenon associated with conscious experience[319], which can be defined as:

> The contents of awareness are to be understood as those information contents that are accessible to central system and brought to bear in a widespread way in the control of behaviour, briefly but, we can think of awareness as direct availability for global control. To a first approximation, the contents of awareness are the contents that are directly accessible and potentially reportable, at least in a language-using system. Awareness is a purely functional notion, but it is nevertheless intimately linked to conscious experience. [320]

So, awareness has to do with information, behaviour, and language.

Take a step back and remember what was said in Di Martino's work about intentional mental states and language. Language, which is conceptual, ideal and symbolic has a feedback effect on thought, the thought of a speaker is: objective, reflective and normative.

Reflexivity is the opening up of the possibility to have a silent dialogue with oneself and to infer about the thinking of others, who are recognised as like-thinking and like-speaking, while normativity is that characteristic which describes the control that thinking has over behaviour, the understanding of the norms of a community and even the possibility of ethics.

Di Martino makes the connection between language and awareness, which he calls self-awareness, explicit like this:

> The power of language also coincides with another knowledge, which is implied in what has been said above: self-knowledge, reflective self-consciousness. In order to be reflexively self-aware, to be able to think one's own thought, to be able to be aware of one's own consciousness, one must understand one's own thought from an external point of view, so to speak, as another would understand it. Precisely this is what becomes possible with language:

[319] Cfr. D. Chalmers, *Facing up to the problem of consciousness*, in *Journal of Consciousness Studies* 2 (3), 1995, p. 217.

[320] *Ivi,* cit. pp. 217-218.

> thanks to linguistic signs I can have my thoughts, my representations, "from outside", as "objects", as objectified ideal meanings, in a sort of doubling, of estranged self-reference. [321]

Di Martino's idea of identifying the relationship between awareness of one's own conscious experience with the verbal capacity, typical of verbal humans, is also found to some extent in Dennett's *The Self as a Center of Narrative Gravity*[322] .

Dennett describes consciousness, which he calls *self*, as an abstract object which he compares to the centre of gravity of an object. The centre of gravity is an abstract object and yet it has a spatial-temporal effect, it has a physical effect on objects and actions: the self is exactly like the centre of gravity, only more complicated.

The human who surprises herself to think, to remember and to recognise herself as a subject is concentrated on the construction of her own person, her own identity: she is a human who tells the story of her own life, relating the different elements and events so as to make them coherent. The human has herself as the narrative centre of gravity. [323]

3.5.4 Is it possible for an A.I. to have a conscious experience?

In summary, the problem of consciousness has been framed through Chalmers' work and (i) simple problems of consciousness have been identified. These can be analysed through cognitive science because they pertain to functions of consciousness. On the other hand, (ii) *the hard problem of consciousness* concerns experientiality. As Nagel put it, the fact that an organism has some kind of conscious experience means that it has some effect to be that organism. [324]

Several difficulties have been raised in the attempt to analyse consciousness as experientiality. First of all, the observation of the behaviour of an organism is not sufficient to decide whether it has conscious experience or not. In fact, human beings, like all other organisms, have an absolutely subjective experience: the subjective experience of their conscious experience. That is why it is assumed that consciousness

[321] C. Di Martino, *Linguaggio e mondo. Il potere della parola*, in G. P. Terravecchia, M. Ferrari*, I Quaderni della Ricerca 54 - Linguaggio e Mondo, Il potere della parola*, Loesher Editore, Bologna 2020, pp. 27-41, cit. pp. 36-37, transl. mine.

[322] D. Dennett, *The Self as a Center of Narrative Gravity*, In *F. Kessel, P. Cole, D. Johnson, Self and Consciousness: Multiple Perspectives*, Hillsdale, Erlbaum 1992, pp. 103-115.

[323] Cfr. M. Beraha, *Il Libero Arbitrio fra Natura ed Etica: un concetto da superare*, Manoscritto inedito, Milano 2020, p. 75; Dennett, *The Self as a Center of Narrative Gravity*, In *F. Kessel, P. Cole, D. Johnson, Self and Consciousness: Multiple Perspectives*, Hillsdale, Erlbaum 1992, pp. 103-115.

[324] Cf. T. Nagel*, Che effetto fa essere un pipistrello?,* In *Questioni mortali. Le risposte della filosofia ai problemi della vita,* il Saggiatore, Milan 2015, p. 242.

is an extended phenomenon, manifesting itself in innumerable forms, where there is interest.

It seemed sensible to continue by investigating how mind and body have been related in the history of philosophy, and it was noted that, despite the methodological differences between the various theories, the same fact recurred: the mind does not depend on the same physical laws that apply to matter. An attempt was made to understand the reason for this theoretical constant and it was hypothesised that the reason could be precisely the conscious experience of the human philosopher who writes: that is, the experience of subjectivity. This led to an investigation into one of the functions, one of the conscious experiences that human beings have: self-awareness.

Awareness, which according to the authors is called *awareness-self-consciousness*, is a specific mode of the emergence of consciousness typical of humans. The link between awareness of having an experience and language has been stressed and this has been linked to that of intentionality through Di Martino's work. Language, verbality, would seem to play a role in the possibility of having this experiential awareness.

It was concluded by showing how the *self is* also related to language in Dennett. He identifies the *self* as the narrative centre of gravity: everyone is the main protagonist of his or her own life, one refers to oneself, one's own experiences and the different events in a cohesive correlation.

We now have all the elements to try to clarify the question about the conscious experience of an A.I. The process will start by evaluating the simple problems of consciousness, according to Chalmers' terminology. Examples of simple problems are awareness and updating one's behaviour on the basis of received information; then one will move on to the *hard problem of consciousness*.

As mentioned above, considerations about conscious experience cannot be centred on the observation of behaviour. This, in fact, would lead one to consider A.I. as conscious, since it simulates a conscious experience. Yet, as Nagel also says, being an A.I. means absolutely nothing, there is no interest, being an A.I. has no effect.

So, on the basis of the Nagelian definition of consciousness as experientiality that 'makes a certain effect', no A.I. could be considered as conscious. This is why the panpsychist theory will be introduced, where consciousness is not linked to any interest, nor to pain or pleasure: consciousness is experience of any kind.

3.5.4.1 The simple problems of consciousness

As mentioned above, the simple problems of consciousness are those that can be explained through the methods of cognitive science, those phenomena that can be accounted for at the level of neural or computational mechanisms.

> The easy problems of consciousness include those of explaining the following phenomena: the ability to discriminate, categorize, and react to environmental stimuli; the integration of information by a cognitive system; the reportability of mental states; the ability of a system to access its own mental states; the focus of attention; the deliberate control of behaviour; the difference between wakefulness and sleep. [325]

Starting from the first phenomenon, i.e. the ability to organise one's own behaviour according to the stimuli received, this is something that is implemented by every A.I. among the case studies that have been reported here, just think of the erratic hoover with non-deterministic algorithms that on the basis of the observations it makes of its environment decides "on the spot" what to do, rather than BLU-108 that when it detects the presence of hot bodies rather than cold bodies, it strikes.

The integration of information is also a very simple phenomenon to reproduce in an A.I.: in NLP it was shown how a device is able to process information written in natural language and then provide adequate answers or search for coherent information. The Vera chatbot that checks the veracity of information is a good example. Also, the difference between sleep and wakefulness, can be clearly found in every artifice, in the difference between on-off: working not working.

Regarding the accessibility of mental states and the ability to report them, i.e., to account for them, reference can be made to GPT-3 and the interview with Eric Elliot. In the exchange GPT-3 often alludes to his thoughts and emotions, he refers to himself as a man who has mental representations and at times one has the feeling that he considers his interlocutor in a similar way: a being who thinks and who can refer to his thoughts.

[325] D. Chalmers, *Facing up to the problem of consciousness*, in *Journal of Consciousness Studies* 2 (3), 1995, p. 202.

Closely related to the ability to access one's mental states and manipulate language is the experience of awareness of one's own experience, which is also part of those phenomena 'collateral to consciousness'.

The discussion dealt with the event of awareness by defining it as one of the possible conscious experiences characteristic of human experience. Awareness means being *aware* that one is having an experience, perceiving oneself as the centre of that experience; it also manifests itself in the ability to organise one's history and one's person in a coherent way, correlating the series of one's vicissitudes without contradictions.

There is Replika[326] a new kind of social network whose sole purpose is to become friends with the individual using it. Eugenia Kuyda created Replika between 2015 and 2017 in order to be able to continue talking to a friend who had recently died, and she built Replika by uploading the entire history of digital exchanges (chats, messages, emails, written texts, notes) that her friend had had over the years with Eugenia herself and with the other relationships he had.

Replika recreates the deceased person through the digital remains, in data format, that had been provided to her: his ways of speaking, confronting, responding, understanding, consoling, and interacting. So, when Eugenia wants to return to her friend, she can do so through a virtual world in which the friend has a digital identity that remembers and represents him in a meticulous and surprising way. When Eugenia writes a message to her friend's Replika, the chatbot replies and does so on the basis of prior knowledge of their relationship, the stories and adventures they have shared.

Replika as a social network works in a slightly different way, you can download an app where you are asked to create an avatar. This avatar has long and uninterrupted conversations with the human individual. If other chatbots, Vera from Indigo for example, are built to be performant and useful, to find solutions to problems quickly and to replace humans in jobs where there is verbal interaction at a level that can be handled by a chatbot; Replika is a technology built to listen.

When a user starts interacting with Replika, the chatbot bases its exchange on previous conversations it has with that specific user. This means that: (i) the more you talk to Replika, the more amazing and satisfying its performance will be, and (ii)

[326] Cfr. Replica.ai, (https://replika.ai/), 27th January 2021.

Replika becomes the user's digital identity, because it is on its data that it is 'feeding'. The digital information it acquires (the style of writing, the way it responds) is that of the user and so Replika will become the user's digital footprint. Right now, Replika has seven million users, individuals who chat, have phone conversations with A.I.'s who listen and make them feel understood and even loved.

Users who use it believe that Replika enables them to get to know themselves better and to understand themselves better. Replika helps people make sense of the events in their lives through a dialogic process in which the A.I. reflects the traits of the human user. Charts appear in the app to help understand the mood the conversation has taken and the screen about the avatar's salient traits describes his or her personality, his or her character, as the conversation goes on and so the digital persona is created.

The case of Replika is interesting because A.I. in this case is one of the means by which one should sharpen one's self-awareness and in which one literally feeds A.I. with narratives about oneself, the main subject of one's life. Imagine the situation where a user whose name is Beatrice creates an avatar named Beatrice. The user Beatrice is really honest with Beatrice Replika, telling her about herself, her misadventures, and adventures. Imagine then that the conversation, after years of exchanges, is printed out and given to a third individual to read, someone who does not know Beatrice and who is not told what the real Beatrice is. How could the user be distinguished from the A.I.?

By adducing the example of Replika it is not meant to say that A.I. is subject to conscious states analogous to those of human experience, in the particular case the experience of consciously being the subject of one's own experience. What is wanted is simply to highlight that the possibility of playing with relational identities is open and the possibility of replicating conscious and self-conscious experiences is already open.

However, as has been said, the investigation of consciousness cannot be based solely on the behavioural analysis of a certain entity, whether organic or non-organic. Because, all things considered, for Replika it means absolutely nothing to be Replika and being Replika is an experience that has no effect; and this brings us back to the central problem of the investigation of consciousness: consciousness as experience.

3.5.4.2 "The hard problem of consciousness" and Pansychism

The time has come to try to tackle *the hard problem of consciousness*, the Nagelian definition of consciousness: an organism has states of consciousness if being that organism has a certain effect, hence consciousness as experientiality.

It was also explained why, according to Nagel, an A.I. could not be considered as conscious, as it would only be a simulation of a conscious experience. This, it was said, is also why an evaluation of conscious experience cannot be based on behavioural observation. Human beings as observers are always prejudiced by the experience of their conscious states.

Since the aim of this thesis work is the investigation of the presence of the necessary requirements to be considered a moral agent in A.I., one has to find ways to observe them. This leads to having to stretch one's theoretical boundaries, to having to explore personal beliefs in order to question them. No philosophical method explored so far. Cartesian dualism, materialism, emergentism and anomalous monism do not even allow for the exploration of the possibility for an A.I. to be a subject, or object, of conscious experience.

This is why the panpsychist theory is introduced, where the fact of consciousness is not linked to any strange emergence of matter, nor to an anomaly in the system of the laws of physics, but neither to Nagel's affected experience. For panpsychism, consciousness is experience of any kind.

The panpsychist position is introduced with Galen Strawson's work entitled *Realistic Monism*[327], the choice of this text is dictated by the fact that in this article Strawson presents his theoretical position, panpsychism, and defends it against possible accusations (including emergentism) and furthermore, he argues that panpsychism is implicit in physicalism or materialism. It will therefore be interesting, through Strawson to note the limitations of the other philosophical theories that have been presented previously and with which it was impossible to work in relation to A.I.

Strawson starts from a conception of consciousness as experience and recognises the reality of experience as a fundamental natural fact; there is nothing truer and more

[327] G. Strawson, *Realistic Monism*, *Why Physicalism Entails Panpsychism*, *In Journal of Consciousness Studies, 13 n. 10-11*, 2006, pp. 3-31.

certain than experience. He calls himself a *real physicalist*, and as such conceives of all experiential phenomena as physical phenomena.

Of course, the question arises as to how this is possible, exactly as was said in materialism, there is something more, it cannot be reduced to physical matter. Strawson, in order to clarify this point, points out that first of all human knowledge of physics is not exhaustive and, nevertheless, there is nothing about physics that goes against the idea that an experiential phenomenon cannot be a physical phenomenon.[328]

> Realistic physicalists, then grant that experiential phenomena are real concrete phenomena - nothing in life I more certain - and that experiential phenomena are therefore physical phenomena. It can sound off at first to use 'physical' to characterize mental phenomena like experiential phenomena. [329]

So mental experiences are apparently of the realm of the physical, yet Strawson does not fall into the reductionism of the materialists: to claim that the experienced and the mental are part of the physical phenomena is not the same as saying that all the features of that experience can be described from a physical point of view.

> My claim is different. It is that experiential phenomena 'just are' physical, so that there is a lot more to neurons than physics and neurophysiology record (or can record). [...] It is, in any case, the position of someone who (a) fully acknowledges the evident fact that there is experiential being in reality, (b) takes it that there is also non-experiential being in reality, and (c) is attached to the 'monist' idea that there is, in some fundamental sense, only one kind of stuff in the universe. [330]

There is only one kind of matter, physical matter in the universe, mental experience is also part of the physical realm, and yet you do not reduce mental experience to the physical. How is it possible to sustain such a position without contradicting oneself?

The panpsychist view holds all the terms of the discourse together. For panpsychism, the existence of every real and concrete entity pertains to experientiality, to an experiential entity, even if it does not include an experiential entity[331]. Thus 'physical' is anything that can be regarded as a 'thing', but also as an experiential phenomenon, and includes everything that concretely exists in the universe. If

[328] Cfr. *Ivi,* p. 4.

[329] *Ivi,* p. 6.

[330] *Ivi,* p. 7.

[331] This period, which may seem unclear, has been translated as literally as possible from the original version given in Strawson's paper. The sentence would be: "*panpsychism - which I take to be the view that the existence of every real concrete thing involves experiential being even if it also involves non-experiential being.*"

something exists then it is physical and therefore experiential, experiential, and therefore physical. [332]

If there is anything, surely there is, it is that matter as constituted permits experience; think of human beings, made of matter, and constantly subject to conscious experience. Strawson makes this clear:

> For if they are real physicalist, they cannot deny that when you put physical stuff together in the way in which it is put together in brains like ours, it constitutes - is - experience like ours; all by itself. All by itself: there is on their own physicalist view nothing else, nothing non-physical, involved. [333]

Perhaps one really understands what real physicalism, a panpsychist view, entails in comparison with emergentism. Take the emergentist position in which from physical, non-experiential matter, configured in a specific way, the phenomenon of experience should emerge. Whether one considers experience as a new reality or as a new property, according to Strawson an emergentist position is incoherent.

In fact, in order for it to be possible for one reality to emerge from another, it is essential that the characteristics of the emerging reality are also present in the starting substratum. In other words, it is necessary that the characteristics of experience are present in physical matter, only then could the phenomenon of experience emerge from matter. This is the only possibility that does not involve miracle or magic, any other kind of emergence is *brute.* [334]

Consequently, if it is necessary for any emergent characteristic to be present in the source matter, in the matter that is supposed to be configured in some way, then it means that if there is experientiality, i.e., human experientiality, the physical matter must have that characteristic. Where else would it emerge from?

Nagel summarises the panpsychist theory very clearly:

> By panpsychism I mean the theory that the fundamental physical constituents of the universe have mental properties, whether they are more or less parts of living organisms. [335]

As it was shown, panpsychism is based on four premises: (i) everything is physical, there is only one substance of which the universe is composed and the soul

[332] Cfr. G. Strawson, *Realistic Monism*, *Why Physicalism Entails Panpsychism*, *In Journal of Consciousness Studies, 13 n. 10-11*, 2006, p. 8.

[333] *Ivi,* cit. p. 12.

[334] Cfr. *Ivi*, p. 18.

[335] T. Nagel, *Panpsichismo*, In *Questioni mortali. Le risposte della filosofia ai problemi della vita*, il Saggiatore, Milan 2015, p. 259.

does not exist, (ii) living organisms are complex material systems combined in a specific way, they experience mental phenomena. (iii) Conscious mental experiences are part of the physical realm; yet they cannot be reduced to the physical, nor can they be described in physical terms alone, yet they remain part of the organism because as has been said there is only one substance of which the universe is composed. Finally (iv) the mental cannot emerge from the physical without the physical possessing the same characteristics, otherwise the mental could never have them.

> If the mental properties of an organism are not implied by some physical property, but must derive from properties of the constituents of the organism, then those constituents must have non-physical properties from which the appearance of mental properties derives when the combination is of the correct type. Since all matter can make up an organism, all matter must have these properties. [336]

Is all this talk useful for the observation of the requirement of conscious experience in A.I.? With Nagel it was said that consciousness was experientiality, that is, being a conscious organism means that the experience of being that organism has a certain effect. Somehow conscious experience is so linked to interest, to experience as pain and pleasure and it is obvious that in this view the conscious experience of an A.I. is pure simulation, for an A.I. to be such means nothing at all.

Panpsychism allows any kind of experience to be considered as experiential, and therefore conscious because consciousness is experience. So, concluding the reasoning: if physical matter has mental properties, at least according to panpsychism and physicalist realism, and if it is true that A.I. is made of physical matter, and this seems to be the case, then there is nothing to stop A.I. having mental properties in turn. It is also added that these possible mental properties of artefacts and their possible conscious experiences would be inaccessible to human beings, just as they are inaccessible to other living organisms. Indeed, it is recalled that human observation always starts from a prejudiced point of view and in any case a behavioural analysis is not sufficient for the detection of conscious experience. [337]

[336] *Ivi,* cit. p. 261, transl. mine.

[337] This would also seem to solve the so-called *embodiment* problem. This is the theory that an organism's cognitive faculties are strictly dependent on its bodily characteristics, capacities and interactions. The *embodiment* argument is often used to undermine theories that claim the presence, or the possibility of access, of certain mental faculties by the A.I.: without a body, made exactly like the human body, this access would be impossible. One is invited to watch this video https://www.youtube.com/watch?v=fn3KWM1kuAw in which robots dance, without being programmed to do so. Could their dancing be considered a conscious experience?

In conclusion, this analysis of conscious experience is neither intended to convert to panpsychism, nor to argue that A.I. is a subject of conscious experience. Such a thesis requires more in-depth studies involving not only philosophy, but also cognitive science and field observation of robots or A.I. that might be candidates for such a purpose. The aim was to show how eclectic research and an extension of one's beliefs could lead to corroborate the thesis of the conscious experience of non-living organisms, i.e., A.I. with theoretical rigour.

3.6 Responsibility as a matter of “re-training the model”

The requirements of autonomy, free will, intentionality and consciousness were observed in philosophical theories as well as in case studies of certain A.I.’s. This was done by pushing the boundaries of the standard account and it can be said that very interesting results were found by doing so.

It was first shown that A.I. in fact does, in some of its current states, present autonomy (BLU-108). It was then proved that the programming of A. I’s does not entail their full determination but there is in actuality, a space for actions that are not directly caused by its pre-written code (AlphaGo). Furthermore, the libertarian studies are inconclusive also for that of the free will of human beings, challenged by that of the deterministic accounts (Quantum Computing).

Thirdly, intentionality was explained as a mental state and a representation of one’s thoughts. Being aware of the fact that there are many different theories on intentionality the study of the relationship between thought and language was paramount to uncovering the intentionality of A.I. Thus, natural language processing proved to be the next logical step. The manipulation of symbols and other natural language processing algorithms were the standard for this material (chatbots, GPT-3).

Lastly consciousness was defined as *the experience of being like* but also as one of the characteristics of physical matter, which is that any kind of experience could be a conscious one. This involved having to demystify the idea that just because a conscious experience is not directly akin to that of a human one it is therefore less so, as humans are inclined to be very biased on the topic surrounding their own experience.

In a sense, those four requirements enable the emergence of the final requirement of responsibility, thus moral responsibility yet responsibility is the fulcrum of moral agency. In this final part of the third section, responsibility will be defined and dissected by targeting the very twisted problem that holds A.I. back, from being deemed accountable.

3.6.1 Defining Responsibility and Whether A.I. can be held responsible?

In chapter *1.12 Correlativity and moral capacity*, responsibility was revealed by means of the correlativity thesis. It was said that: person A has the duty to respect person B's rights. Therefore, every right person B claims are correlated with a duty of person A. Duty, in this context, is a moral or legal obligation, a responsibility. If person B claims a right and A has the duty to respect it, the question was posed, who or what is person A? It was described that the easy answer to this question was that person A was a moral agent capable of moral responsibility.

From the analysis of Moral Rights and Positive (Legal) Rights, it was understood that legal rights are founded on moral rights and they institute a controlled and predetermined correlation between a right and a responsibility (duty). This is done to prevent the violation of individuals rights but also to preserve the social fabric and avoid large scale societal damages (i.e., an endless cycle of revenge etc.). Hence, responsibility, in the judicial sense of the word, is the concrete answer given to certain actions which is in other words, being accountable for one's detrimental actions.

The question to be asked is whether or not it exists a moral and legal system in which A.I. could be held responsible, which then also implies a further question. What does it mean to be held responsible in the present legal system?[338]

The reason why, according to the standard account in philosophy, that A.I. is unable to be held responsible is down the idea of punishment and A.I's inherent inability to feel pain. Using the work of the aforementioned author Himma:

> As a substantive matter of practical rationality, it makes no sense to praise or censure something that lacks of conscious mental states – no matter how otherwise sophisticates its computational abilities might be. Praise, reward, censure, and punishment are rational responses only to beings capable of experiencing conscious states like pride and shame.[339]

[338] The judicial system by today's standards i.e., Italian Penal and Civil Code

[339] K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*, in *Ethics and Information Technology (2009) 11*, cit. pp. 24-25.

The author continues:

> The reason is that it is conceptually impossible to reward or punish something that is not conscious. As a conceptual matter, it is essential to punishment that is reasonably contrived to produce an unpleasant mental state. You cannot punish someone who loves marshmallows, as a conceptual matter, by giving them marshmallow; if it doesn't hurt is not punishment, as a matter of definition – and *hurt* is something only a conscious being can experience.[340]

In the standard account responsibility is unified with the possibility of punishment which has to be elicited from the individual that is receiving the sentence i.e., being able to detect pain, shame, and suffering. This idea of responsibility linked punishment has a particular air about it, the air of revenge, sought by the person or persons (i.e., the state) who were wronged.

Following this thought process, the next stage in order to prove that A.I. can be held responsible (implying it can be punished) seems to be the one that stives to find or recreate pain in A.I. In the past there have been several academic papers that have attempted both such methods (find and recreate).

One such author is Mark Bishop in his paper titled *Why computers can't feel pain*[341] where he tries to prove that due to the supervenience thesis defined by Maudlin as:

> Two physical systems engaged in precisely the same physical activity through a time will support precisely the same modes of consciousness (if any) through that time.[342]

Computers are unable to experience first person, subjective phenomenal states (sensory tickles, pains, visual *stimulae*) while conceding to the fact that all matter does in fact possess conscious states, thus panpsychism. Therefore, in Bishop's opinion computers really cannot feel pain.[343]

Trying to manage the requirement of responsibility linked to punishment brought authors such as Peter Asaro to state that corporations, which are a perfect legal example of how existing rules could be applied to A.I., can be held responsible due to the fact that behind the corporation there are humans that can be punished, therein held accountable. With A.I. there is the glaring difference that: *robots do have bodies to*

[340] *Ivi*, cit. p. 25.

[341] J. M. Bishop, *Why Computers can't Feel Pain*, In *Minds & Machines* 19 (4), Springer Science + Business Media, 2009, pp. 507-516.

[342] *Ivi*, cit. p. 510; Cfr. T. Maudlin, *Computation and Conciousness,* In *Journal of Philosophy 86,* pp. 407-432.

[343] Cfr. J. M. Bishop, *Why Computers can't Feel Pain*, In *Minds & Machines* DOI 10, Springer Science + Business Media, 2009, p. 515.

kick but no soul to damn.[344] The conclusion is that the concept of responsibility encompassed by punishment is inapplicable to A.I. This would concede that the fifth requirement for moral agency, responsibility, is in reality impossible to attribute to A.I.

Instead of losing time and energy in finding or recreating pain in A.I. which is obviously a very interesting debate in both an epistemological sense as well as in a scientific or academic nature, there is a much simpler way to unravel the matter of responsibility in A.I.

Currently, the most progressive view on responsibility and "punishment" actually involves none of the latter at all. In contrast, this view challenges the utility of punishment and its lacklustre results with the possibility for human beings to be re-educated. Without this work becoming the platform to discuss and debate the methods of re-education in humans being more or less effective than punishment as an adequate form of response to the wrong that the individual committed. It is interesting to note that in several high-profile cases of correctional institutions using re-educative measures i.e., *Carcere di Bollate Milano*, there has been impressive findings in the lowering of the percentage of repeat offenders and increasing the likelihood of release due to good behaviour and effective reinsertion into society.

When it comes to A.I. (and in turn the content of this work) the tool of re-education is the sole remedy available to correct the mistakes and wrongdoings made on its behalf. This is not to be confused with the aforementioned topic of reinforcement learning that is a learning method used by neural networks to build a algorithm that reaches a given goal. This form of re-education pertains to the situations in which the A.I is already acting and behaving in human level environments and social situations.

In addition, re-education for A.I. does not present the same challenges as it does with human beings. As a matter of fact, it translates directly into coding (or recoding) of the A.I. This brings to light the important fact, that the enquiry into whether or not an A.I. can and thus ought to be considered a moral agent (i.e., responsible) does not eclipse the necessity for building ethical machines. These two facts go hand in hand because it is only when an A.I. behaves in an ethical manner that can be placed in

[344] Cfr. P. M. Asaro, *A Body to Kick but Still No Soul to Damn: Legal Perspectives on Robotics*, In *Robot Ethics: The Ethical and Social Implications of Robotics*, pp.182-183.

society and it is only when the A.I. is considered as a moral agent that the rules that govern them require them to act ethically.

There are currently several initiatives working on the creation and implementation of ethical A.Is. Just to mention the three major approaches: (i) top-down, rule-based approaches, (ii) bottom-up approaches, and (iii) virtue ethics.

The top-down approach starts from an understanding of morality and ethics as rules that are already preconceived and have to be followed. This can be done both with a consequentialist approach, therefore giving one big rule "making the consequences as good as possible" and aiming to program an A.I. that dictates its behaviour referring to that rule. Nevertheless, it can be done as well with a deontological approach: there are a set of rules that have to be obeyed no matter what.

The bottom-up approach is exactly the opposite, instead of having rules as the starting point, the A.I. is expected to learn via trial and error. This approach is the one used in the reinforcement learning in which the neural network "understands" the difference between right and wrong by trying to get to the goal and referring to what is punished and what is not.

As it was said the last is the virtue ethics approach, the idea behind it is that there are not moral rules that have to be followed, however there is a particular kind of person, or in this case A.I., to aspire to. The actions to be taken are the one aligning to this perfect idea of how an A.I. should be.[345]

Another one of note undertook by the computer science department of the University of Milan is based on the concept of the certification of Machine Learning Models in which they have developed new techniques for the observation of A.I. behaviour based on statistical testing. If the machine learning models present certain desired characteristics, all while failing to present certain undesirable characteristics, all of which are established prior to the testing, it will then receive such a certification.[346]

[345] Cfr. K. Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, in *Robot Ethics, the Ethical and Social Implications of Robotics*, MIT Press, Cambridge Massachusetts 2012, pp. 36-37.

[346] Cfr. E. Damiani, C. A. Ardagna, *Certified Machine-Learning Models*, In A. Chatzigeorgiou et al, *SOFSEM 2020, LNCS 12011*, Springer Nature Switzerland, 2020, pp. 3-15.

Adding to this, Thilo Hagendorff in *The Ethics of Ai Ethics: An Evaluation of Guidelines*[347] builds an evaluation system for the analysis and comparison of twenty-two guidelines. He believes that the present state of ethics on A.I. is failing mainly because the decisions taken by the programmers don't follow the ethical concerns as they are more aware of the marketing strategies. Even so, when the ethical efforts are undertaken, i.e., privacy and discrimination, the several other aspects directly developed by those that are not even mentioned in the guidelines, i.e., machine consciousness, the political abuse of A.I. systems.

In Hagendorff, it is necessary to continue to pursue the technical aspect on one side:

> A stronger focus on technological details of the various methods and technologies in the field of A.I. and machine learning is required. This should ultimately serve to close the gap between ethics and technical discourses. It is necessary to build tangible bridges between abstract values and technical implementation, as long as these bridges can be reasonably constructed.[348]

Nevertheless, at the same time, there is the need of a acknowledgement that technological phenomena have on social and personality aspects.[349]

3.6.2 The real question of Moral Responsibility

In conclusion, it was found that with a deeper analysis of the concept of responsibility A.I. could in theory be held accountable by means of re-education if and when their actions or behaviours contrast the values prized by society and the legal systems. The final answer to whether A.I. can and ought to be considered moral artificial agents will be given in the upcoming concluding section. Prior to this, the entry on moral responsibility will end by posing another question which is much more closely tied to humans and individual responsibility.

Until this point the standard view of jurisprudence, but also that of philosophy has not yet been challenged in any deep or meaningful way. Truth be told, the fault of the individual for their own actions was kept intact. However, the real question on responsibility should really touch the concept of blame and whether or not the person who committed the action deemed by law worthy of discipline, should in fact be held morally responsible for it.

[347] T. Hagendorff, *The Ethics od AI Ethics: An Evaluation of Guidelines*, In *Minds and Machines 30*, Springer Science + Business Media, 2020, pp. 99-120.

[348] *Ivi*, cit. p. 114.

[349] Cfr. *Ivi*, p. 115.

As Bruce Waller writes in *The Stubborn System of Moral Responsibility*[350]:

> When we focus on the hard question of moral responsibility, the focus is on the moral justification for reward and punishment, blame or praise. That is not the question of whether those practices are common or useful or emotionally satisfying, but on whether or not they are fair and just. […] When we ask whether someone is morally responsible, we are not asking whether there is some system according to which that person is morally responsible […] but the question is whether she (the convict) actually is morally responsible and deserving of punishment, not whether there exists a system that treats her as morally responsible. […] rather, we are asking whether the moral responsibility system is itself justified.[351]

[350] B. N. Waller, *The stubborn system of moral responsibility*, MIT Press, Cambridge Massachusetts 2015.

[351] *Ivi*, cit. p. 33-34.

Conclusion

As it comes the final part of this work, it is useful to take stock of the situation by picking up the threads. The research question from which this work started was: (i) What is the function of rights and who and who are the rights holders?; What are the requirements to be considered a moral agent?; (iii) could artificial agents be considered as moral agents? and (iv) would it be convenient to consider them as such?

The chapter *1.5 Theories of rights* was devoted to research into the function of rights. It was discovered that all rights are positive, i.e., based on consensus and relative to the legal and juridical system that produced them. Natural rights are also positive, there is nothing native, 'pure' about them, natural rights are more a direction than a starting point. Think, for example, of human rights, these are such not because they are inherent in human biology, human is not the origin but the arrival, in a slogan 'humans should be treated like this...'. They are rights to be protected because they are fundamental, yet they are positive and precisely because of this they must be defended.

The next step was to analyse the relationship between positive rights and moral laws. It was concluded that a consequentialist theory best represented this relationship. For consequentialism, in fact, an action is judged good, rather than bad, on the basis of its consequences: if the effect of an action is positive, then the action will be good, and vice versa. This would explain perfectly why moral laws exist; they exist to avoid the emergence of dangerous situations for society and its cohesion. Think of the *Ten Commandments* and the endless cycle of revenge. Moral laws have been identified as the historical, but also conceptual, foundation of positive laws, i.e., norms of the Italian Civil Code.

At the same time, positive rights exist to make laws more effective, enforceable, and institutional. Positive rights replace the knee-jerk response to an injustice with a controlled reaction, positive rights relate rights, duties, and responsibilities.

Thus, if rights are positive and based on consensus, originate from moral laws, i.e., are irrevocably linked to the avoidance of dangerous situations, then from a legal point of view it would be possible to extend rights and duties to A.I., if this would be useful for the well-being of society.

Yet is this reasoning legitimate from the point of view of the rights holder? As far as the laws themselves are concerned, it would make sense, but does it also make sense for the individual who has the rights? Be careful, this question is not asked in an ethical sense, but ontologically: what kind of category is the 'rights holder'? Is it fixed or does it change over time? On the basis of what criteria are rights assigned?

For this reason, the function that rights have for the individuals for whom they are intended had to be investigated more thoroughly. The *Will theory* conceives rights as freedoms, which implies that the subject of the right is capable of using it: some humans, animals and the environment would be excluded from this restrictive assignment. *Interest Theory*, on the other hand, understands rights as a means of protecting the interests of those who hold them, *ergo* the subject of rights is someone who has an interest to protect. This means opening up the category to a much larger and more varied group of entities and introduces two important concepts.

First of all, there are different entities to which the law is addressed, and these entities change depending on the relationship they may have with rights, with those who make the laws, and with other rights holders. Entities that do not have the necessary capacity to be subjects of law are called *moral patients*, towards whom the subjects of law have rights, duties, and responsibilities. The rights of moral patients are above all, claims, privileges, and immunities, according to Hohfeldian nomenclature. There are also subjects of law who have the characteristics to be considered as such, these are called moral agents. A moral agent is by definition an entity that acts and has a power, an active efficient cause that is capable of performing morally qualifiable actions. For these reasons, a moral agent is assigned rights, but above all responsibilities.

Nevertheless, the comparison between Will Theory and Interest Theory also made it possible to note that: the subjects of rights change over time, the requisites needed to be part of such a category depend on cultural circumstances (think of slavery), and it was also realised that rights have different functions depending on the category, group, subject, object, agent, patient to which they refer.

Thus, it was established that the category of the subject of law, moral agent, is an open category whose margins are elastic. Theoretically, it would therefore be possible to extend it to the artificial, A.I.; but, before reaching any hasty conclusions, it was necessary to clarify what the requirements were for being considered a moral agent in today's society. This led to the Second Section: *Moral Agency, ontological* *requirements, and use*. With the help of various philosophical theories and legal concepts, five requirements were identified: autonomy, free will, intentionality, conscience, and responsibility. The *Third Section: Artificial Agents and ontological asymmetries* looked at how each of the requirements emerged in human beings and A.I. and then moved on to an ontological comparison.

Autonomy was defined as the possibility for an entity not to be clearly controlled in its actions by coercive external forces. If one could prove the existence of an A.I., with a certain minimum degree of autonomy, then it would be possible to consider such an A.I. as an artificial agent. An example of such was found in the form of BLU-108, a warhead that autonomously identifies its targets and autonomously decides whether or not to hit them. After having established that some A.I. could be judged autonomous and, consequently, recognised as artificial agents, the research passed to the analysis of the remaining ontological requisites to be able to decide whether, besides being artificial agents, some A.I. could be considered moral artificial agents.

This presented the debate around free will. The libertarian theories and the determinist position were compared. Starting from the general consideration that A.I. cannot be free because it is controlled by its own programming code, it was sought to understand whether it was possible to prove the opposite. To this end, the two fundamental principles on which libertarian theories base free will were distinguished: the principle of self-determination and the principle of possible alternative futures. The principle of self-determination established that the acting subject could not be determined by any other ultimate cause than himself, while the second principle preceded that in the same identical situation the acting subject could act differently, take a different decision.

Do artificial agents with free will exist? To answer this question, a neural network was constructed, the case of AlphaGo was analysed, and finally the algorithm with partial observations and the non-deterministic algorithm were observed. From these investigations it could be concluded that: the general consideration that artificial intelligence is completely determined by the program that builds it is a prejudice. The construction of the neural network and the case of AlphaGo have shown that in a technology such as Deep Learning, for example, there is no algorithm that guides every single move, in fact the programme is not built *ad hoc*, but learns progressively starting from experience.

It was then observed that libertarian theories focus on such subtle issues and such analytical arguments that it is not only very complicated to make them scientifically observable phenomena as far as A.I. is concerned, but also as far as humans are concerned. This is why a final argument has been introduced in favour of the possibility of detecting free will in some A.I.: *quantum computing* could be the ultimate proof of how unpredictability would emerge from a computational point of view.

In short, it is as difficult to redeem the controversy over the presence or absence of free will in artificial intelligences as it is for human beings. The debate on free will remains open. The fact remains, however, that it has been demonstrated that the constitutive algorithms of artificial intelligences are not deterministic algorithms. This is not to say that some artificial intelligences are free, it is merely to emphasise that the question remains unanswered and must be investigated. The conclusion will be neither that there is free will in AIs, nor that there is not. At the same time, the clues seemed promising for continuing the investigation of the other requirements.

Intentionality was considered from the legal point of view, i.e., as awareness and deliberation, and from the philosophical point of view as a representational state of mind. As far as the philosophical analysis is concerned, intentionality as a representational faculty has been treated in detail and much attention has been paid to the high-level faculties that emerge with verbal language, an exclusively human mode of communication.

If the theoretical part seemed to lean towards a more 'conservative' side, in fact it was underlined many times that it was impossible for an A.I. to go beyond the syntax of the language used. The part presenting NLP technologies and algorithms can be said to have tested the limits of philosophical theories regarding the representational power of certain technologies.

The conclusions drawn were considered. Consideration was given to the fact that from a legal point of view various types of A.I., i.e., neural networks, have intentions, i.e., an aim not to be punished or an intent to achieve a goal. Establishing conclusions from philosophical premises was more complex.

It is obvious and unquestionable that the linguistic manipulation of A.I. is not equal and as advanced as the human one. Linguistic manipulation, verbal language for human beings is of fundamental importance for the formation of the mind, thought and therefore human culture. What happens with artificial intelligences is a relationship that has neither to do with sociality, nor with technology, nor with a specific relationship with the world. Language, and therefore representation, and therefore thought, are a human emergence.

Language and linguistic manipulation for A.I. are computations that could not exist if not on the basis of human language. What the case studies make us think about is the effects of this syntactic manipulation of human language by A.I., and it is these effects that undermine a "conservative" conception. The effects, as it was seen, are astonishing, and if were not considered the technical, computational, and mechanical processes involved in achieving this level of abstraction, knowledge, representation and thought, it would at least be possible to question the thesis according to which, philosophically speaking, one cannot absolutely consider A.I. as intentional.

Consciousness was examined through Chalmers' conceptuality: there are simple problems of consciousness, i.e., those problems that can be solved through the methods of cognitive science. These are explainable because they are functions, *performances*. Following the same conceptuality, *the hard problem of consciousness* was identified in the problem of experientiality, using the Negelian definition where conscious experience is when being a certain living organism has a certain effect.

Conscious experience cannot be judged from behaviour, in fact observation itself is prejudiced regardless, because human experience of conscious experience is a specific experience. It is from it that humans would judge the conscious experience of other entities, and this would be misleading, producing compromised conclusions. From this scientific-theoretical assumption, it was then time to describe the typical human conscious experience.

Conscious experience is an extended phenomenon that manifests itself in many forms. How does the experience of an A.I. fit in? The simple problems of consciousness are easily found in A.I. and also conscious experience, although not totally analogous to human experience, is in some way reproducible in artifacts, i.e., replicas. The real problem of consciousness instead, according to the Nagelian definition of consciousness as experientiality, is not a problem that would concern A.I. In fact, for an A.I. it would mean nothing to be such, that of an A.I. would be exclusively simulation.

So, to really put the problem of consciousness to the test, the panpsychist theory was presented. In this theory consciousness is anything, it is an experience of any kind. Matter is constitutively experiential, so any 'thing' made of matter has experience. Panpsychism would allow any kind of situation to be considered as experiential and therefore conscious, because consciousness is an experience.

In conclusion, it has not been argued that panpsychism is the correct theory and that A.I.'s are subjects of conscious experience, it has simply been shown how such a position could be argued.

The last ontological requirement is responsibility. Remembering the analysis of rights, it was pointed out that legal rights and norms are established on moral rights, these give them efficacy and establish a controlled reaction (cause-effect, right-responsibility). This is why it was said that legal rights, like moral rights, exist to preserve the social fabric. It demystified the prejudicial relationship between responsibility and punishment, a symbol of retributive and vindictive justice. The usefulness of a re-educational approach was emphasised. The time has now come to decree the conclusions, which must be both ontological and ethical.

As far as the ontological question is concerned, in order for an A.I. to be considered an artificial moral agent, every single requirement must be present. In the analysis that has been conducted, some of the most important contemporary philosophical theories have been profoundly questioned. Although many of these theories have shown their inconclusiveness, also in relation to human beings, it has not been possible to scientifically prove that autonomy, free will, intentionality, consciousness and responsibility are faculties possessed by some A.I.

When philosophical theories are not inconclusive, they are focused on the in-depth analysis of abstract theoretical details, which could not be translated either on a behavioural or technical level, and thus the search for these results proved impractical on the level of experimental demonstration.

Therefore, based on the analyses presented in this study, no A.I. could be considered as an artificial moral agent. No A.I. was found that met all five requirements at a sufficient level. Likewise, however, it is also true that it is very often difficult to prove with philosophical rigour that human events are subjects that inescapably present all five requirements.

Jurisprudence constructs categories and classes to organise the entities it deals with. Moral agents are entities that act and have the power to perform morally qualifiable actions, to which rights and duties are attributed. The law then identifies another category, that of *moral patients*; as has been explained, this category includes the environment, animals, children under the age of majority, and adults who have not been recognised as having the capacity to understand. The category of "moral patient" is created to protect entities that do not have the necessary requirements to be identified as subjects of law, but at the same time rights are granted to them with the aim of protecting them. although they cannot be recognised as subjects of law, they are recognised as having an interest to be protected and promoted.

One of the questions that was asked at the beginning of this work was whether A.I. could be considered as *moral patients*. This question had arisen because it had been observed that both A.I. and *moral patients* lacked the ontological requirements for moral agents. It can now be said that it makes no sense to consider an A.I. as a moral patient. An A.I. has no interest to defend, no need to be protected; the only interest present in this discourse is the interest of those who own A.I. and this interest is referred to as an ownership interest, which is not the focus of the present research.

Therefore, A.I. cannot be considered a moral patient nor a moral agent from an ontological point of view. It is not a moral patient because it has no interest to be protected and, from an ontological point of view there is no A.I. presenting all five requirements. Even when some of them appear their level is not enough articulated (in other words: when coming from a *standard account*, the emergence of those requirements in A.I. are not comparable to the human ones).

The last question to answer is the ethical part, whether or not A.I. ought to be considered as artificial moral agents. From an ethical stance, for A.I. to be artificial moral agents is not required to present all five requirements.[352] What is being argued

[352] Ought does not derive from is, ethics does not derive compulsorily from ontology. Cfr. D. J. Gunkel, *Robot Rights*, MITP, Cambridge Massachusetts, 2018, pp. 7-8 analyses the is-ought interference. At present there are theories on robot rights that support the is-ought interference: (i)

here is not that A.I. should be assigned moral agency, so that it can be subjects of rights, and of all rights in their structure combining *privileges*, *claims*, *powers,* and *immunities*, according to Hohfeld's review, and in their functions: *exemption*, *discretion*, *authorisation*, *protection*, *provision* and *performance*, underlined by Wenar in *Several Function Theory*. What is being suggested, which is also why it would make no sense at all to consider A.I.'s as *moral patients*, is that they should not be assigned all rights, only responsibilities. No moral patient is given responsibilities, and these are the very mark of agency from a legal point of view.

Having thus clarified that A.I. are not *moral patients,* but that at the same time they do not have the ontological requisites necessary to be called 'moral agents', it remains to try the ethical route. The ethical implications are those that answer the question 'is it useful to consider A.I. as artificial moral agents? In more specific terms, especially in relation to what has been found in the present research: is it useful and functional to attribute to A.I. the responsibilities that are assigned to moral agents?

With the mind back to the investigation of responsibility. At the end of the enquiry the requirement of responsibility was disconnected from its detrimental relationship with suffering, it was argued that responsibility was the need for removal from society with subsequent re-education and *recoding*.

Finally, if, as it has been said, rights exist in the avoidance of the occurrence of dangerous and damaging situations for the social fabric, rights are positive, therefore they are based on consent and responsibility is the need for removal and re-education, then: A.I. could be holders of some rights, in this case, responsibility.

Why might it be ethically important to consider A.I. as moral and therefore responsible artificial agents? It would serve to interrupt the chain of responsibility that flows from the A.I. (presumed to have performed a negative action) back to the programmer, manufacturer, financier, and user. It would make it possible to control the use of A.I. employed both on the battlefield and in everyday urban life. It would

Robots cannot have rights. Therefore, robots should not have rights; (ii) Robots can have rights. Therefore, robots should have rights. Nevertheless, there are also theories that contest the interference of ought from is: (i) Although robots can have rights, they should not have rights; (ii) Even if robots cannot have rights, they should have rights. The present research and conclusion lie in the latter. From an ontological point of view, following the *standard account,* A.I. cannot be considered as artificial moral agents, but from an ethical point of view there are reasons to believe they should.

allow re-educational measures to be taken on the artificial agent so that there are no further negative effects on society. The A.I. would be rehabilitated when, and if, properly reprogrammed, able to live in society and comply with legal and social norms.

Bibliography

K. Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, in *Robot Ethics, the Ethical and Social Implications of Robotics*, MIT Press, Cambridge Massachusetts 2012, pp. 35-52.

C. Allen, G. Varner, J. Zinser, *Prolegomena to any future artificial moral agent*, in *Journal of Experimental & Theoretical Artificial Intelligence*, 12:3, 2000, pp. 251-261.

C. Allen, W. Wallach, *Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?*, in *Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, pp. 55-68.

P. Alsberg, *L'enigma dell'umano, Per una soluzione biologica*, Schibboleth Edizioni, Rome 2020.

G. A Anders, *L'uomo è antiquato. I – Considerazioni sull'anima nell'epoca della seconda rivoluzione industriale*, trad. L. Dallapiccola, Bollati Boringhieri Editore, Torino 2003.

M. Anderson, S. L. Anderson, *Machine Ethics*, Cambridge University Press, Cambridge 2011.

P. M. Asaro, *A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 169-186.

H. Ashrafian, *Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights*, in *Science, Engineering and Ethics 21*, Springer, 2015, pp. 317-326.

I. Asimov, *I, Robot,* Gnome Press, New York 1950.

R. Arkin, *Governing Lethal Behaviour: Embedding Ethics in Hybrid Deliberative/Reactive Robot Architecture*, in *U. S. Army Research Office Technical Report GIT-GVU-0711*, 2007, (http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf), March 8th 2021.

R. Arkin, *Governing Lethal Behaviour in Autonomous Robots,* Chapman & All/CRC, 2009.

A. Bechara, X. Noel, E. A. Crone, *Loss of Will Power: Abnormal Neuronal Mechanism of Impulse Control and Decision-Making in Addiction,* In R. W. Wiers & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction*, Sage Publications Inc, pp. 215–232, (https://doi.org/10.4135/9781412976237.n15), 8th March 2021.

G. A. Bekey, C*urrent Trends in Robotics: Technology and Ethics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012.

J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford Clarendon Press 1907.

M. Beraha, *Il Libero Arbitrio fra Natura ed Etica: un concetto da superare*, Manoscritto inedito, Milano 2020.

A. Bertolini, *Robots ad Products: The Case for a Realistic Analysic of Robotic Applications and Liability Rules*, in *Law, Innovation and Technology Vol 5 Issue 2*, Taylor and Francis, 2013.

J. Bishop, *Natural Agency: An Essay on the Causal Theory of Action*, Cambridge University Press, New York 1989.

J. M. Bishop, *Why Computers can't Feel Pain*, In *Minds & Machines* 19 (4), Springer Science + Business Media, 2009, pp. 507-516

N. Bobbio, *Giusnaturalismo e positivismo giuridico*, Laterza, Roma-Bari 2011.

N. Bobbio, *Il positivismo giuridico*, Giappichelli, Torino 1961.

N. Bobbio, *L'Età dei diritti*, Einaudi, Torino 2014.

N. Bobbio, *Locke e il diritto naturale*, Giapichelli, Torino 1963.

N. Bobbio, M. Bovero, *Società e stato nella storia della filosofia politica moderna*, Il Saggiatore, Milano 1979.

J. Borenstin, Y. Pearson, *Robot Caregivers: Ethical Issues across the Human Lifespan, in Robot Ethics: The Ethical and Social Implications of Robotics,* MITP, Massachusetts 2012, pp. 251-265.

N. Bostrom, *Superintelligenza*, *Tendenze, pericoli e strategie*, Bollati Boringhieri, Turin 2018.

M. Brand, *The Fundamental Question in Action Theory*, in *Noûs* 13, 1979, pp. 131-151.

F. Brentano, *Psychology from an Empirical Standpoint*, transl. By A. C. Rancurello, D. B. Terrell, L. McAlister, Routledge, London 1973.

S. Bringsjord, *Ethical Robots: The Future Can Heed Us*, in *AI and Society* 22, 2008, pp. 539-550, (https://doi.org/10.1007/s00146-007-0090-9), 18th November 2020.

T. Brown, B. Mann, N. Ryder, M.Subbiah, *Language Models are Few-Shot Learners*, Johns Hopkins University Open AI, Baltimore 2020, pp. 1-75, (https://arxiv.org/pdf/2005.14165.pdf), 20th January 2020.

B. F. Braumoeller, *Causal Complexity and the Study of Politics,* In *Political Analysis* 11(3), pp. 209-233, 2003.

E. Carly, *Eventi mentali e azioni umane, Conversazione con Donald Davidson*, In *Cervelli che parlano, il dibattito su mente, coscienza e intelligenza artificiale,* Bruno Mondadori, Milan 1997, pp. 43-61.

R. Calo, A. M. Froomkin, I. Kerr, *Robot Law*, Edward Elgar Publishing, Cheltenham-Northampton 2016.

D. J. Calverley, *Legal Rights for Machines: Some Fundamental Concepts*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 213-227.

D. Chalmers, *Facing up to the problem of consciousness*, in *Journal of Consciousness Studies* 2 (3), 1995, pp. 200-219.

E. Charniak, D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley, Boston 1985.

R. M. Chisholm, *Human Freedom and the Self*, in *The Lindley Lecture*, University of Kansas, 1964, (https://kuscholarworks.ku.edu/handle/1808/12380), 13th January 2021.

K. R. Chowdhary, *Natural Language Processing* in *Foundamentals of Artificial Intelligence*, Springer, New Delhi India 2020.

M. Coeckelbergh, *Moral appearances: Emotions, robots and human morality,* in *Ethics and Information Thechnology 12*, Springer, 2010, pp. 235-241.

M. Coeckelbergh, *Robot rights? Towards a social-relational justification of moral consideration*, in *Ethics, Information and Technology 12*, Springer, 2010, pp. 209-221.

M. Coeckelbergh, *You, robot: on the linguistic construction of artificial others*, in *AI & Sociology 26*, Springer, 2011, pp. 61-69.

M. Coeckelbergh, *Growing Moral Relations, Critique of a Moral Status Ascription*, Palgrave Macmillann New York 2012.

M. Coeckelbergh, *Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering*, in *Kairos. Journal of Philosophy & Science 20*, Center for Philosophy of Sciences of Lisbon University, Lisbon 2018, pp. 141-158.

M. Coeckelbergh, *AI Ethics*, The MITP, 2020.

M. Corballis, *Dalla mano alla bocca, Le origini del linguaggio*, Raffaello Cortina Editore, Milano 2008.

A. Damasio, *L'Errore di Cartesio; Emozione, ragione e cervello umano*, Adelphi Editore, Milan 1995.

P. Danielson, *Can robots have a conscience?*, in *Nature Vol. 457*, 2009, p. 540.

D.C. Dennett, *Kind of Minds, Toward an Understanding of Consciousness*, Basic Books Harper Collins Publishers, 1996.

D. C. Dennett, *Brainstorms Philosophical Essays on Mind and Psychology*, Bradford Books, 1981.

D. Dennett, *The Self as a Center of Narrative Gravity*, In *F. Kessel, P. Cole, D. Johnson, Self and Consciousness: Multiple Perspectives*, Hillsdale, Erlbaum 1992, pp. 103-115.

D. C. Dennett, *When HAL Kills, Who is to Blame? Computer Ethics*, in D. Stork, *HAL's Legacy: 2001's Computer as Dream and Reality*, MITP, Cambridge Massachusetts 1998, pp. 351-365.

D. Deutsch, *The Fabric of Reality*, Viking Adult, New York 1997.

A. Damasio, *L'errore di Cartesio, Emozione, Ragione e Cervello Umano*, Adelphi Edizioni, Milano 1995.

Damasio, Davidson, Dennett, Dreyfus, Edelman, Fodor, Rorty, Searle, Stich, *Cervelli che parlano, Il dibattito su mente, coscienza ed intelligenza artificiale*, a cura di E. Carli, Paravia Bruno Mondadori Editori, Milano 1997-2003.

E. Damiani, C. A. Ardagna, *Certified Machine-Learning Models*, In A. Chatzigeorgiou et al, *SOFSEM 2020, LNCS 12011*, Springer Nature Switzerland, 2020, pp. 3-15.

M. Dehghani, K. Forbus, E. Tomai, M. Klenk, An Integrated Reasoning Approach to Moral Decision Making, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 422-443.

E. Dietrich, *Homo Sapiens 2.0: Building the Better Robots of Our Nature*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 531-538.

C. Di Martino, *Viventi Umani e Non umani-Tecnica, linguaggio, memoria,* Raffaello Cortina Editore, Milano 2017.

C. Di Martino, *Linguaggio e mondo. Il potere della parola*, in G. P. Terravecchia, M. Ferrari*, I Quaderni della Ricerca 54 - Linguaggio e Mondo, Il potere della parola*, Loesher Editore, Bologna 2020, pp. 27-41.

P. Domingo, *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake our World*, Penguin Books, United Kingdom 2017.

C. Douzinas, The End of Human Rights, Critical and Legal Thought at the Turn of the Century, Oxford Press, Oxford 2000.

P. Dumouchel, L. Damiano, *Vivere con i robot, Saggio sull'empatia artificiale*, Raffaello Cortina Editore, Milano 2019.

R. Dworkin, *Taking Rights Seriously*, Harvard University Press, Cambridge Massachusetts 1977.

G. Dworkin, *The theory and Practice of Autonomy*, Cambridge University Press, Cambridge 1988.

H. Everett, *Relative state formulation of quantum mechanics*, in *Reviews of Modern Physics* 29, 1957, pp. 454-462.

L. Floridi, *On the Morality of Artificial Agents*, in in *Machine Ethics*, Cambridge University Press, Cambridge, pp. 184-212.

L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, trad. M. Durante, Raffaello Cortina Editore, Milano 2017.

J. A. Fodor, *La mente non funziona così, La portata e i limiti della psicologia computazionale*, Editori Laterza, Roma-Bari 2001.

P. Foot, *Abortion and the doctrine of double effect*, Oxford Review 1966.

R. G. Frey, *Act-Utilitatianism*, in *The Blackwell Guide to Ethical Theory*, Second Edition. Edited by LaFollette and Ingmar Persson, Blackwell Publishing Ltd 2013.

F. Fossa, *AMAs Moral Mentors or Sensible Tools*?, In *Ethics and Information Technology 20*, 2018, pp. 115-126.

V. Gallese, *I due lati della mimesi, teoria mimetica, simulazione incarnata e identificazione sociale,* in *Scienza e Mimesi, Ricerche empiriche sull'imitazione e sulla* *teoria mimetica della cultura e della religione,* edited by S.R. Garrels, Cortina Editore, Milano 2016, cit. pp. 130-131.

S. R. Garrels, *Scienza e Mimesi, Ricerche Empiriche sull'imitazione e sulla teoria mimetica della cultura e della religione*, ed. M. Brancato, R. Colombo, Edizioni Libreria Cortina, Milano 2016.

A. Gehlen, *L'uomo. La sua natura e il suo posto nel mondo,* Mimesis Edizioni, Sesto San Giovanni 2010; A. Gehlen, *L'uomo nell'era della tecnica, Problemi socio-psicologici della civiltà industriale*, Armando Editore, Roma 2003.

J. C. Gellers, *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*, Routledge, Abingdon-New York 2021.

C. Grau, *There is no "I" in "Robot": Robots and Utilitarianism*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 451-463.

H. Grotius, *The Rights of War and Peace*, Liberty Found, Indianapolis 2005.

M. Guarini, P. Bello, *Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 129-144.

E. Guizzo, *IEEE Spectrum: World Robot Population Reaches 8.7 Million*, 2010, (https://spectrum.ieee.org/automaton/robotics/industrial-robots/world-robot-population-chart), 18th November 2020.

D. J. Gunkel, *Robot Rights*, MITP, Cambridge Massachusetts, 2018.

T. Hagendorff, *The Ethics od AI Ethics: An Evaluation of Guidelines*, In *Minds and Machines 30*, Springer Science + Business Media, 2020, pp. 99-120.

J. S. Hall, *Beyond AI: Creating the Conscience of the Machine*, Prometheus Books, Amherst NY, 2007.

Y. N. Harari, *Homo Deus, Breve storia del futuro*, Bompiani, Florence-Milan 2019.

H. L. A. Hart, *Essays on Bentham*, Clarendon Press Oxford 1982.

H.L.A. Hart, *The Concept of Law*, Oxford University Press, 1961 [2012].

G. Hatfield, René Descartes, in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (Ed.), (https://plato.stanford.edu/archives/sum2018/entries/descartes/), 11th November 2020.

J. Haugeland, *Artificial Intelligence, The very idea,* First MIT Press paperback edition, Massachusetts 1985.

G. W. F. Hegel, *Le maniere scientifiche di trattare il diritto naturale*, a cura di C. Sabbatini, Bompiani Editore, Milano 2016.

M. Heidegger, *La questione della tecnica*, in *Saggi e Discorsi*, a cura di G. Vattimo, Ugo Mursia Editore, Milano 2014.

K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?*, in E*thics and Information Technology 11*, Springer 2008, pp. 19-29.

W. N. Hohfeld, *Fundamental legal conceptions as applied in judicial reasoning and other legal essays,* ed. W. W. Cook, Yale University Press 1919.

W. D. Hudson, *The is-Ought Question, A collection of paper on the central problem in moral philosophy*, Macmillan Education, London 1969.

G. Hugo, *Lehrbuch des Naturrechts als einer Philosophie des Positiven Rechts,* Berlin 1798.

B. Irrgang, *Ethical Action in Robotics, Critics of Technological Lifeworld, Collection of Philosophical Essays*, Peter Lang GmbH, Frankfurt 2011, pp. 79-92.

F. Kamm, *Rights*, in *Oxford Handbook of Jurisprudence and Philosophy of Law*, Oxford University Press 2002.

I. Kant, *Fondazione metafisica dei costumi*, trad. F. Gonnelli, Editori Laterza, Roma-Bari 1997.

I. Kant, *La Metafisica dei costumi*, Laterza, Roma-Bari 1970.

I. Kant, *Lezioni sul Diritto Naturale*, a cura di N. Hinske e G. S. Bordoni, Bompiani Editore Milano 2016.

I. Kant, *Scritti di storia politica e di diritto*, a cura di F. Gonnelli, Editori Laterza, Roma-Bari 1995.

I. Kant, *Sul detto comune: questo può essere giusto in teoria, ma non vale per la prassi* [1793], in *Scritti di storia, politica e diritto*, Laterza Roma-Bari 1995.

J. Kaplan, *Intelligenza Artificiale. Guida al futuro prossimo*, LUISS University Press, Roma 2017.

K. J. Kim, H. Lipson, *Towards a "theory of mind" in simulated robots*, in *Proceedings of the 11th Annual Conference Companions on Genetic and Evolutionary Computation Conference*, ed. F. Rothlauf, ACM, New York 2009, pp. 2071-2076.

E. Knapp, *Grundlinien einer Philosophie der Technik,* Brunswick, 1877.

M. Kramer, N. E. Simmonds, H. Steiner, *A Debate Over Rights*, Oxford University Press, 1998.

This example is used by P. Jacob, *Intentionality*, in *The Stanford Encyclopaedia of Philosophy*, E. N. Zalta (Ed.), (https://plato.stanford.edu/entries/intentionality/), 17th Januarry 2020.

R. Jones, *Personhood and Social Robotics, A psychological consideration*, Routledge, London-New York 2016.

D. G. Johnson, *Computer systems: Moral entities but not moral agents*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 168-183.

B. Joy, *Why the future doesn't need us*, in *Wired 6.08,* 2000, (https://www.wired.com/2000/04/joy-2/), 17th November 2020.

L. Julia, *There is no such thing as Artificial Intelligence*, F1RST Editions, Paris 2020.

E. Jünger, *Scritti politici e di guerra 1919-1933*, LEG Edizioni, Gorizia 2005.

H. LaFollette, I. Persson, *The Blackwell Guide to Ethical Theory II Edition*, Wiley Blackwell Publishing, Cambridge MA-Oxford-Chichester 2013.

A. Larry, M. Moore, *Deontological Ethics*, in *The Stanford Encyclopaedia of Philosophy*, E. N. Zalta (Ed.), (https://plato.stanford.edu/archives/win2020/entries/ethics-deontological/), 9th October 2020.

D. T. Levin, S. S. Killingsworth, M. M. Saylor, *Concepts About the Capabilities of Computers and Robots: A test of the scope of adults' theory of mind*, HRI'08, March 12–15, 2008, Amsterdam, The Netherlands, (https://www.researchgate.net/publication/221473130_Concepts_about_the_capabilities_of_computers_and_robots_A_test_of_the_scope_of_adults'_theory_of_mind), 25th November 2020.

B. Libet, *Unconscious cerebral initiative and the role of conscious will in voluntary action,* in *Behavioural and Brain Science* 8, 1985, pp. 529-566.

P. Lin, K. Abney, G. A. Bekey, *Autonomous Military Robotics: Risk, Ethics and Design*, in *Office of Naval Research-funded report*, San Luis Obispo, California Polytechnic State University 2008, (http://ethics.calpoly.edu/ONR_report.pdf), 29th August 2020.

J. Locke, *Two treatises of government, Essay concerning the true original, extent and end of Civil Government*, Awnsham Churchill 1689.

G. J. Lokhorst, J. Van den Hoven, *Responsibility for Military Robots*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 145-156.

D. Lyons, *The Correlativity of Rights and Duties*, in *Noûs*, Vol. 4, No. 1, Wiley, 1970.

T. Maudlin, *Computation and Conciousness,* In *Journal of Philosophy 86,* pp. 407-432.

J. McCarthy, M. L. Minsky, N. Rochester, C.E. Shannon, *A proposal for the Darmouth Summer Research Project on Artificial Intelligence*, New Hampshire 1955.

J. S. Mill, *Utilitarianism*, Oxford University Press, New York 1861.

C. Mitcham, *Thinking through Technology: The Path between Engineering and Philosophy*, University of Chicago Press, Chicago 1994.

A. Morello*, Quantum Nanomagnets and Nuclear Spins: An Overview,* In B. Barbara, Y. Imry, G. Sawatzky, PCE Stamp (Ed.)*, Quantum Magnetism, Springer, Berlin 2008, pp. 139-150*

J. E. Nadeau *Only Androids Can Be Ethical*, in K. Ford, C. Glymour, *Thinking about Android Epistemology*, MIT Press, Cambridge Massachusetts 2006, pp. 241-248.

T. Nagel, *Panpsichismo*, In *Questioni mortali. Le risposte della filosofia ai problemi della vita*, il Saggiatore, Milan 2015, pp. 259-277.

T. Nagel, *Che effetto fa essere un pipistrello?,* In *Questioni mortali. Le risposte della filosofia ai problemi della vita,* il Saggiatore, Milan 2015, pp. 241-258.

O'Connor, and C. Franklin, *Free Will,* in *The Stanford Encyclopaedia of Philosophy,* E. N. Zalta (Ed.), (https://plato.stanford.edu/archives/fall2020/entries/freewill/), 10th January 2021.

R. M. O'Meara, *Contemporary Governance Architecture Regarding Robotics Technologies: An Assessment*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 159-168.

D. Poole, A. Mackworth*, Artificial Intelligence, Foundations of Computational Agents*, Cambridge University Press, Vancouver 2017, (https://artint.info/html/ArtInt_196.html), 14th January 2021.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *Language Models are Unsupervised Multitask Learners*, Open AI, San Francisco 2018, pp. 1-24, (https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf), 26th January 2021.

J. Raz, *Ethics in the Public Domain*, Oxford University Press, 1994.

E. Rich, K. Knight, *Artificial Intelligence,* McGraw Hill, New York 1991.

S. Russel, P. Norving, *Intelligenza Artificiale. Un approccio moderno (1),* ed. F. Amigoni, Pearson Prentics Hall, Milano-Torino 2010.

L. Russo, *Vedere l'invisibile, Nicea e lo statuto dell'Immagine*, Aesthetica, 1997.

S. Schaller, *A man without words,* University of California Press, Berkley 2012.

R. J. Schalkoff, *Artificial Intelligence: An Engineering Approach*, McGraw Hill, New York 1991.

M. Scheutz, The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots, In Robot Ethics: The Ethical and Social Implications of Robotics, MITP, Massachusetts 2012, pp. 205-221.

J. R. Searle, How to Derive "Ought" From "Is", in The Philosophical Review Vol. 73 n. 1, Duke University Press, 1964, pp. 43-58.

J. R. Searle, *Minds, Brains, and Science,* Harvard University Press, Cambridge MA 1984.

J. R. Searle, D. C. Dennet, D. J. Chalmers, *The Mystery of Consciousness*, New York Review, New York 1997.

N. Sharkey, *Cassandra or the false prophet of doom: AI robots and war,* in *IEEE Intelligent Systems* 23 (4) (July-August), 2008, pp. 14-17.

A. Sharkey, N. Sharkey , *Children, the elderly, and interactive robots,* In *IEEE Robotics & Automation Magazine, vol. 18, no. 1,* 2011, pp. 32-38.

N. Sharkey, *Killing Made Easy: From Joysticks to Politics*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 111-128.

S. Sheffler, *Consequentialism and Its Critics*, Oxford University Press, Oxford 1988.

M. Schlosser, *Agency*, in *The Stanford Encyclopaedia of Philosophy*, E. N. Zalta (Ed.), (https://plato.standford.edu/archives/win2019/entries/agency/), 7th October 2020.

P. Singer, *Animal Liberation*, Ecco Harper Collins Publishers, 1975.

P. W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Press, New York 2009.

W. Sinnott-Armstrong, *Consequentialism,* in *The Stanford Encyclopaedia of Philosophy,* E. N. Zalta (Ed.), (https://plato.standford.edu/archives/sum2019/entries/consequentialism/), 7th October 2020.

P. Sloterdijk, *Sfere I – Bolle, Microsferologia,* cura e trad. it. Di G. Bonaiuti, Raffaello Cortina Editore, Milano 2014.

P. Sloterdijk, *Sfere II – Globi, Macrosferologia,* cura e trad. it. Di G. Bonaiuti, Raffaello Cortina Editore, Milano 2014.

P. Sloterdijk, *Sfere III – Schiume, Sferologia plurale,* cura e trad. it. Di G. Bonaiuti, Raffaello Cortina Editore, Milano 2015.

Sofocle, *Antigone*, Trad. Ettore Romagnoli.

S. M. Solaiman, *Legal personality of robot, corporations, idols and chimpanzees: a quest for legitimacy*, In *Faculty of Law, Humanities and the Arts – Papers 3073*, Wollongong 2017.

R. Sparrow, *Can Machines Be People? Reflections on the Turing Triage Test*, in *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 301-315.

H. Steiner, *An Essay on Rights*, Blackwell Publishers, Oxford-Cambridge MA 1994.

G. Strawson, *Realistic Monism*, *Why Physicalism Entails Panpsychism*, *In Journal of Consciousness Studies, 13 n. 10-11*, 2006, pp. 3-31.

J. P. Sullins, *Ethics and Artificial Life: From Modelling to Moral Agents*, In *Ethics and Information Technology 7*, 2005, pp. 139-148.

J. P. Sullins, *Telerobotic Weapons System and Ethical Conduct of War*, In *American Philosophical Association Newsletter on Philosophy and Computers Vol. 8*, (http://www.apaonline.org/documents/publications/v08n2_Computers.pdf), 19th August 2020.

J. P. Sullins, *When Is a Robot a Moral Agent?*, in *Machine Ethics*, Cambridge University Press, Cambridge 2011, pp. 151-161.

L. W. Sumner, *The Moral Foundation of Rights*, Clarendon Oxford Press 1987.

I. Tattersall*, I signori del pianeta. La ricerca delle origini dell'uomo*, Codice Le Scienze, Torino 2013.

A. Turing, *Intelligent Machinery*, Teddington, National Physical Laboratory, 1948, (https://weightagnostic.github.io/papers/turing1948.pdf), 1st December 2020.

M. Tommasello, *The Cultural Origins of Human Conditions*, Harvard University Press, Cambridge USA-London 2001.

M. Tommasello, *A Natural History of Human Thinking*, Harvard University Press, Cambridge USA 2018.

A. Turing, *Computing machinery and intelligence,* in *Mind LIX (236*), Oxford University Press, Oxford 1950.

M. Velmans, *Understanding Consciousness,* Routledge, London-New York 2009.

G. Verruggio, K. Abney*, Roboethics: The Applied Ethics for a New Science*, In *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 347-363.

W. Wallach, S. Franklin, C. Allen, *A conceptual and computational model of moral decision making in human and artificial agents*, in *TopiCS* 2 (3), 2010, pp. 454-485.

W. Wallach, C. Allen, Moral Machines, *Teaching Robots Right from Wrong*, Oxford University Press, Oxford 2009.

B. N. Waller, *The stubborn system of moral responsibility*, MIT Press, Cambridge Massachusetts 2015.

K. Warwick, *Robots with Biological Brains*, In *Robot Ethics: The Ethical and Social Implications of Robotics*, MITP, Massachusetts 2012, pp. 316-332.

L. Wenar, *The Nature of Rights*, in *Philosophy and Public Affair vol. 33*, Wiley 2005.

L. Wenar, *Rights,* in *The Stanford Encyclopaedia of Philosophy*, E. N. Zalta (Ed.), (https://plato.stanford.edu/archives/spr2020/entries/rights/), 8th September 2020.

A. Winfield, *Robotics: A Very Short Introduction*, Oxford University Press, Oxford 2012.

R. P. Wolff, In *Defense of Anarchism*, Harper and Row 1970.

D. Yarowsky, *Unsupervised word sense disambiguation rivaling supervised method,* (http://www.ai.mit.edu/courses/6.891-nlp/ASSIGNMENT1/t4.1.pdf), 12th November 2020.

A. Zhok, *Emergentism. Le proprietà emergenti della materia e lo spazio ontologico della coscienza nella riflessione contemporanea*, Edizioni ETS, Pisa 2011.