The Legal Status of Artificially Intelligent Robots

Personhood, Taxation and Control

By

Filipe Maia Alexandre

Dissertation developed under the supervision of Prof. Erik Vermeulen and submitted to Tilburg University to qualify for the Degree of Master of Laws (LL.M.) in International Business Law

Defended on the 12th of June of 2017

"Tell me, Bernard. If you were to proclaim your humanity to the world, what do you imagine would greet you? A ticker-tape parade, perhaps? We humans are alone in this world for a reason. We murdered and butchered anything that challenged our primacy."

– Dr. Robert Ford, *Westworld* (HBO 2016).

Abstract

The introduction of artificial intelligence in industry and society will revolutionize the current social structures and comport several regulatory challenges, which legal frameworks are not prepared to give a direct response to. In order to accommodate this reality, we understand that machines with limited memory, machines with a theory of mind and self-aware machines should be considered separate legal entities from their owners and users. Attributing a separate legal status to artificially intelligent agents and defining the contents of that status, namely, regarding liability, eventual rights and potential taxation duties, allows for minimum certainty concerning the consequences of the introduction of those new intelligent agents in society, which contrasts with the large amount of unforeseeability that it comports. However, the risks of that unforeseeability still need to be addressed and mitigated, as they are not only related to eventual damages, but also to the protection of personal data and public safety. When designing technology that could impact the safety or wellbeing of humans, it is not enough to simply presume it works.

Table of Contents

1. Introduction

1.1. Background and Motivation

Internet, sustainable energy, genetic reprogramming, multi-planetary life and artificial intelligence. Upon being asked about the future, a forward-looking Elon Musk identifies these items as the five

key areas that will impact mankind and disrupt our species' standards of life[1]. While the Internet and sustainable energy are already being profoundly developed, the remaining 3 items on the list are yet to be mastered. However, there is little doubt that we are currently experiencing the last few years of calm before the storm, as a new industrial revolution is due to arrive. As framed by the European Parliament's Committee on Legal Affairs (JURI), 'from Mary Shelley's Frankenstein's Monster to the classical myth of Pygmalion, through the story of Prague's Golem to the robot of Karel Čapek, who coined the word, people have fantasised about the possibility of building intelligent machines'[2]. During the past few years, humanity's inherent curiosity is rapidly working into turning this fantasy into a reality, as is demonstrated by the behaviour of the market.

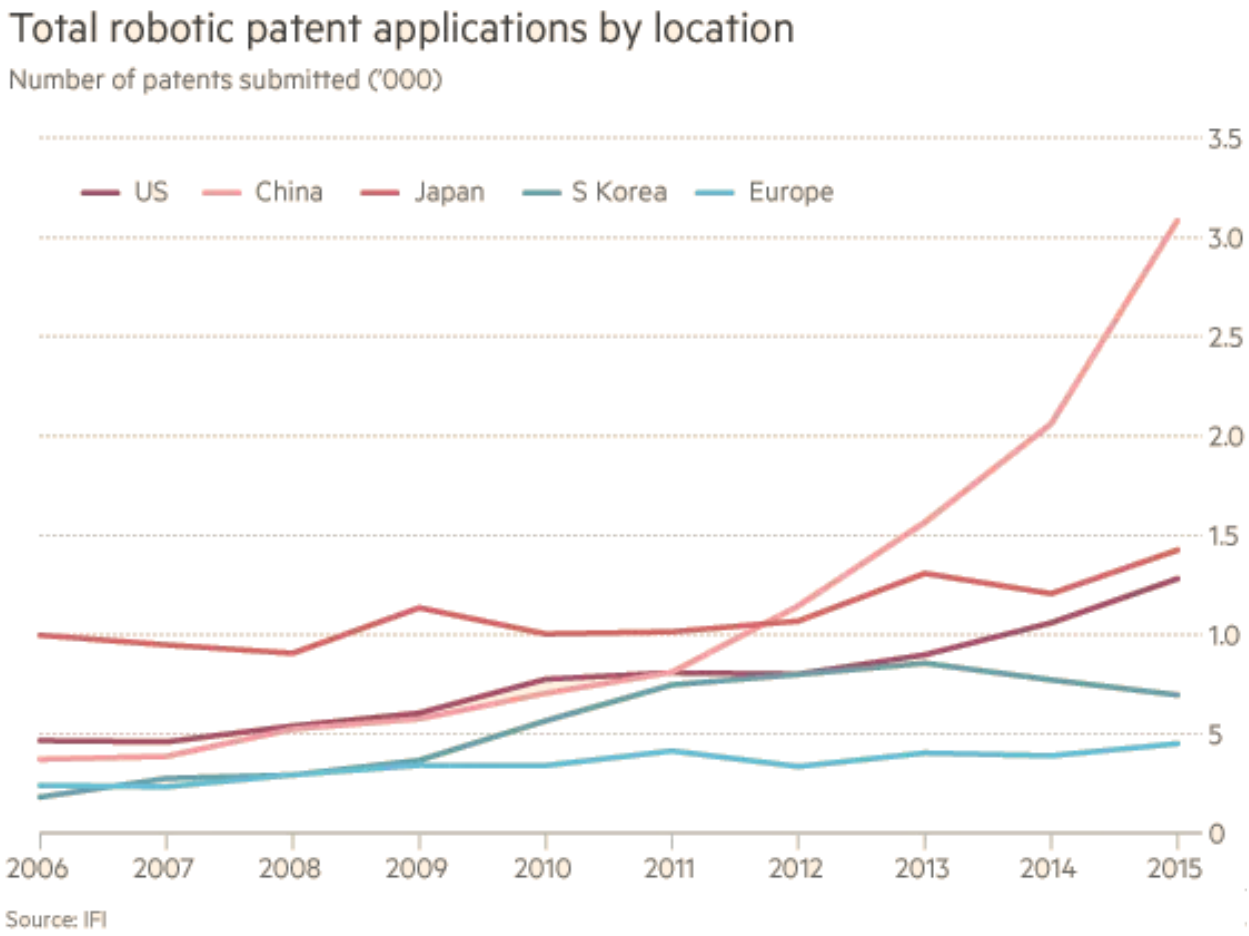

Figure 1 - Annual Patent Filings for Robotics

[1] Elon Musk, Interview with Sam Altman, 'Y Combinator's How To Build The Future Series' (2016).

[2] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

Over the last decade, annual patent filings for robotics technology have tripled. Between 2010 and 2014, the average increase in sales of the robotics industry stood at 17% per year, with 2014 registering the highest YoY increase in history (29%)[3]. Venture capital investments in robotics more than doubled from 2014 to 2015, amounting to an astonishing $587 million (according to the research firm CB Insights)[4], and the predictions point to a global total market worth of the robotics industry of around $135 billion by 2019[5]. These figures show that, while the flow of capital into the robotic industry is still at a preliminary stage, it is, with very little doubt, due to become one of the hottest markets not only in the tech industry but in overall commerce.

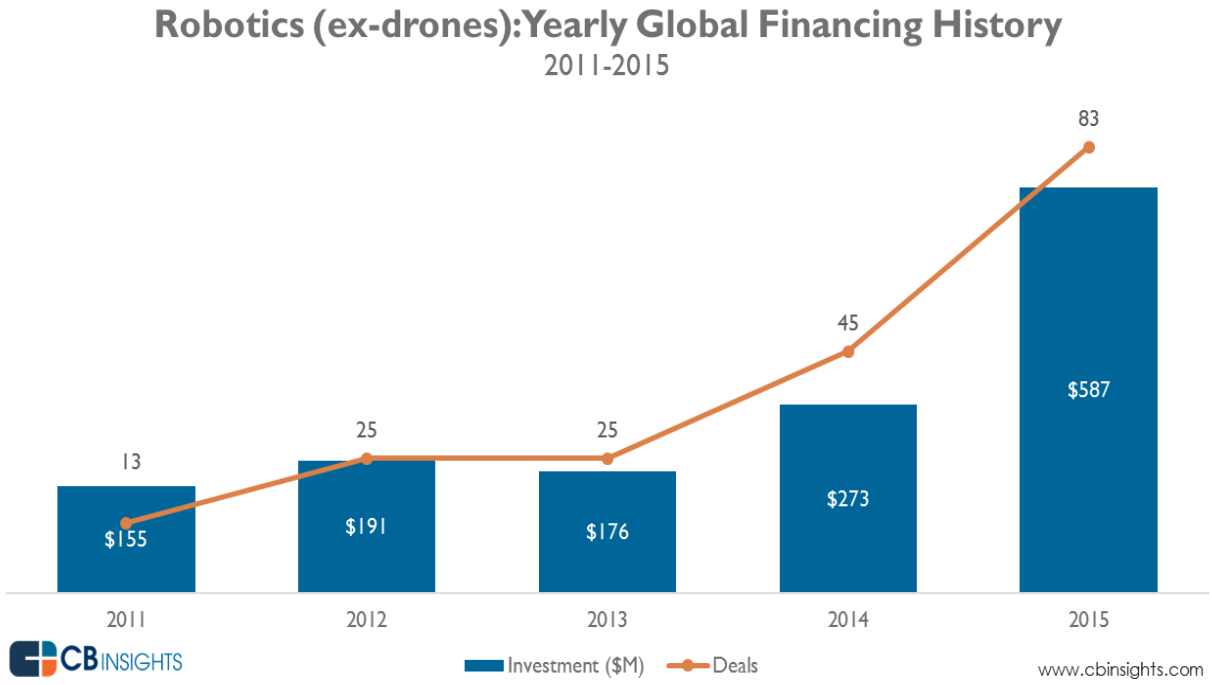

Figure 2 - Annual VC Investments in Robotics

The attractiveness of the robotics industry may be explained through the dual-class benefits that its development is expected to produce. On a corporate level, the introduction and evolution of

[3] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

[4] Richard Waters and Tim Bradshaw, 'Rise Of The Robots Is Sparking An Investment Boom' Financial Times (2016) <https://www.ft.com/content/5a352264-0e26-11e6-ad80-67655613c2d6> accessed 5 April 2017.

[5] Jing Bing Zhang and Others, 'IDC FutureScape: Worldwide Robotics 2017 Predictions' (International Data Corporation 2016).

robotics will improve efficiency and increase ROIs[6] through cost savings ('robots are typically a ninth of the cost of a full-time employee'[7]) and through the performance of more accurate, high quality and uninterrupted work, which will drive productivity up and error rates down. Furthermore, machines can work in hazardous conditions without injuries or fatigue, increasing overall workplace safety. On a social level, robotics will contribute to advancements in several key areas, such as transportation (by optimizing circulation, reducing traffic or increasing safety), medical care (by reducing the chance of error, speeding up procedures or enabling virtually unlimited knowledge sharing), food production (by making predictions on the right time to plant based on climate data and historical conditions or automating processes), among others.

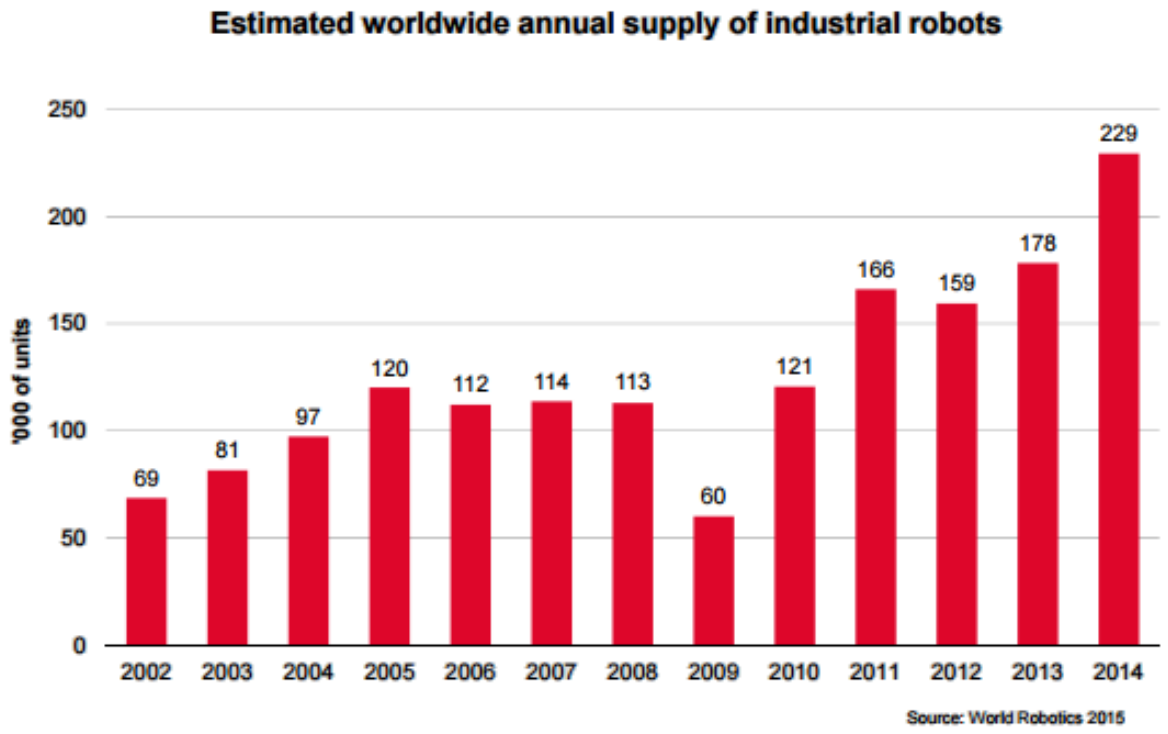

Figure 3 - Annual Sales of Industrial Robots

While in theory robotics and artificial intelligence may enable mankind to complete its ultimate quest for prosperity, the road that currently separates us from it is long and blurry. As every other

[6] Return on Investment = (Gain from Investment – Cost of Investment) / Cost of Investment

[7] Richard Horton, 'The Robots Are Coming' (Deloitte LLP 2015) <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/finance/deloitte-uk-finance-robots-are-coming.pdf> accessed 6 April 2017.

major revolution in industry, this one will also carry several challenges that need to be overcome in order to successfully conduct it. With the creation of intelligence, humanity is on the verge of entering unchartered territory and treading a path which has never been treaded before. 'The autonomous nature of AI creates issues of foreseeability and control that might render *ex post* regulation ineffective'[8]. Due to its potential to profoundly transform society as we know it, a proactive approach may be the only way of ensuring sustainability. But how to proactively approach something unprecedented? This is the question that we have set ourselves to pursue.

1.2. Our Contribution

The upcoming artificial intelligence industrial revolution has the potential to largely improve social welfare. Simultaneously, its arrival comports a large amount of uncertainty regarding its consequences and impact over industry and society, the only thing certain being that no stratum will be left untouched. While this might be frightening for those who fear change or do not deal well with it, for people who accept it, it is certainly exciting. In any case, the artificial intelligence revolution seems inevitable and, expectably, will challenge mankind in ways never done before. For this reason, we understand that it is necessary not only to embrace it, but also to take the lead in its sustainable implementation through proactive and innovative thought and discussion.

The introduction and proliferation of artificial intelligence will comport three major categories of challenges: the technologic challenges, the challenges in industry and the regulatory challenges.

The technologic challenges are those related to the development of artificial intelligence and to the improvement of the capacities and traits of artificially intelligent agents. Such challenges will be dealt with mostly by engineers, programmers, neural scientists and other tech-savvy professionals.

The challenges in industry will be driven by the technologic disruption of the vast majority of existing markets, threatening with the risk of becoming obsolete those who are not able to keep up with the fast pace of technology and to carry out a continuous implementation of the latest advancements in their businesses. These challenges will be harder on large established businesses,

[8] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

which are less flexible and will have a tougher time adapting[9], while start-ups and smaller companies who have less trouble implementing artificial intelligent systems attempt to penetrate their market[10]. They will be dealt with by managers or by *chief innovation officers* of companies.

Finally, the regulatory challenges are the ones concerning the necessity of reforming the countries' macroeconomic policies and legal frameworks in order to accommodate the introduction of new intelligent agents in society. These challenges will be dealt with by policymakers and lawmakers, however, this is also where legal professionals (such as ourselves) who want to get involved can make themselves more useful, contributing to the discussion with a social-science-based approach.

One of the regulatory challenges that lawmakers will face is to reform legal frameworks in order to accommodate the presence of artificially intelligent agents in society and daily life. When these agents start operating without the intervention of humans and, eventually, even without their awareness, questions will be raised regarding liability for their actions, the hypothetical attribution of rights to such agents or the necessary restrictions to their conduct in order to ensure safety, data privacy and other interests that may be threatened by their increasing independence and autonomy. Another challenge that policymakers will foreseeably face is related to the replacement of a large slice of human labour by robotic labour, raising 'concerns about the future of employment and the viability of social security systems if the current basis of taxation is maintained, creating the potential for increased inequality in the distribution of wealth and influence'[11]. In this work, we aim to properly frame these challenges and to identify potential responses that will enable policymakers and lawmakers to respond to them successfully and in a sustainable manner.

1.3. Chapter Summary

[9] Erik Vermeulen, 'How To Prepare For Automation? Or, Why We Need More "Artificial Intelligence Ecosystems" Now!' <https://hackernoon.com/how-to-prepare-for-automation-or-why-we-need-more-artificial-intelligence-ecosystems-now-4a4a767e733b> accessed 10 April 2017.

[10] For further information on what can large established businesses do stay ahead and successfully engage with artificial intelligence, robotics and automation see Erik Vermeulen, 'How To Prepare For Automation? Or, Why We Need More "Artificial Intelligence Ecosystems" Now!', available at https://hackernoon.com/how-to-prepare-for-automation-or-why-we-need-more-artificial-intelligence-ecosystems-now-4a4a767e733b (accessed 10 April 2017).

[11] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

The introduction of artificial intelligence in industry and society will revolutionize the current social *status quo* and, despite having the potential to largely improve welfare and quality of life, will comport several technologic, industrial and regulatory challenges. This works aims to outline the major regulatory challenges and to identify and build on current proposals that respond to such challenges. The next chapter will develop the concepts of *robotics* and *artificial intelligence* that were adopted for the purpose of our work, while each of the subsequent chapters will address said regulatory challenges, with focus on adapting legal frameworks to artificially intelligent agents and on the sustainability of tax structures and social security systems in respect to robotic labour.

2. Defining Artificial Intelligence

2.1. *Robotics* vs. *Artificial Intelligence*

Until this point, the expressions *robotics* and *artificial intelligence* have been used with minimum rigor. However, as we begin to lean deeper into the matter, it is important to define in a strict manner what we mean by these terms and, as a consequence, what the scope of this work is.

Despite the word *robot* being popularly associated with walking and talking machines, such association could difficultly be more misleading nowadays. The origin of the word *robot* is commonly attributed to Karel Čapek, which introduced it in his play *Rossumovi Univerzální Roboti* (Rossum's Universal Robots) in 1920, referring to the Golem of Prague. The word derives from the Czech and Slovak word *robota*, which, in turn, derived from the Proto-Slavic word *orbota*, used to designate hard work or slavery[12]. An adequate usage of the word *robot* will, today, be much closer to its etymological genesis: *robotics* refer to "robotic process automation", as in a way to 'automate repetitive and often rules-based processes'[13]. For the purpose of our work, this will include, but is not limited to, computer coded software and programs that replace humans performing repetitive rules-based tasks, regardless of such performance being or not carried out

[12] 'Definition Of Robot In English' (Oxford Dictionaries, 2017) <https://en.oxforddictionaries.com/definition/robot> accessed 6 April 2017.

[13] Richard Horton, 'The Robots Are Coming' (Deloitte LLP 2015) <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/finance/deloitte-uk-finance-robots-are-coming.pdf> accessed 6 April 2017.

by physical machines. Following the same rational, machines that perform simple tasks dependent of human initiative, such as shredding paper or heating food, are excluded from the concept.

Artificial intelligence, on the other hand, is much harder to define. Grossly, it refers to intelligence exhibited by machines. But how to define intelligence? Are human intelligence and machine intelligence the same? Can a machine exhibit consciousness in the same way a human being can?

Throughout history, several thinkers have tried to provide an answer to these questions. One of the most renowned answers, if not the most, was provided by Alan Turing in 1950, in his famous paper *Computing Machinery and Intelligence*. Turing suggests that rather than determining if a machine can think, the question should be whether a machine can convince a human that it can think. To do so, the machine would have to pass the *Turing Test*, which consists in inducing a human who is not aware he is communicating with a machine into believing that he is communicating with another human[14]. According to Turing, if a machine can behave as intelligently as a human being, then it *is* as intelligent as a human being. This proposition by Turing has been highly influential but also highly controversial. Despite several alternative tests having been proposed over the last decades, along with numerous variations of the *Turing Test* (such as the famous *Feigenbaum Test*[15], which grossly consists in a subject matter expert version of the *Turing Test*), the *Turing Test* remains a mark in the vast field of assessing the existence of artificial intelligence. Nevertheless, the point of our work is not to exactly delimit the scenarios where artificial intelligence exist or not, but to address the consequences of its existence and introduction in society. For this reason, while the *Turing Test* is undoubtedly worthy of mentioning, we will opt for a more functional approach, which, inclusively, leaves margin for interpretation.

Literature holds multiple definitions of *artificial intelligence*. Some are related to the capacity of perceiving the environment and taking action that maximizes the chance of success at a given goal. Others imply owning cognitive functions generally associated with human intellect, such as learning and solving problems[16]. For the purpose of this work, we will consider a broad concept of *artificial intelligence*, defined not by a single notion but by providing examples of traits

[14] Alan Turing, 'Computing Machinery And Intelligence' [1950] Mind.

[15] Edward A. Feigenbaum, 'Some Challenges And Grand Challenges For Computational Intelligence' [2003] Journal of the ACM.

[16] Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach (3rd edn, Prentice Hall 2009).

generally associated to intelligent agents by literature. These traits are reasoning, problem solving, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation of objects, social intelligence and creativity[17][18][19][20]. In order to fit its scope, a certain agent is not required to possess every single one of these traits[21]. It just has to exhibit one or several of them with an intensity minimally sufficient to justify a comparison to human intellect.

2.2. Types of Artificial Intelligence

In nature, intelligence manifests itself with different levels of intensity. Consider the case of a pet dog: while not able to reason or use logic, it may exhibit a limited capacity to solve problems and it is definitely able to learn how to respond to certain commands, such as fetching a newspaper or rolling on ground. The same logic applies to the field of artificial intelligence: artificial intelligence (and the traits generally associated with it) may manifest itself with a stronger or weaker intensity. This concept is well developed in an article published in *The Conversation UK*[22]. According to that article, manifestations of artificial intelligence may fall under one of the following four categories: "reactive machines", "limited memory", "theory of mind" and "self-awareness".

As the name indicates, reactive machines consist of systems that operate in a purely reactive way, having no memories and no ability of using past experience to influence current decisions, hence, behaving in the same way every time they encounter the same situation. They have no concept of

[17] Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach (3rd edn, Prentice Hall 2009).

[18] George F. Luger, Artificial Intelligence: Structures And Strategies For Complex Problem Solving (5th edn, Addison-Wesley 2004).

[19] David Poole, Alan Mackworth and Randy Goebel, Computational Intelligence: A Logical Approach (1st edn, Oxford University Press 1998).

[20] Nils J. Nilsson, Artificial Intelligence: A New Synthesis (1st edn, Morgan Kaufmann Publishers 1998).

[21] In his book *The Age of Spiritual Machines* (1999), Ray Kurzweil discusses the topic of artificial general intelligence ('strong AI'), which grossly consists in a machine's ability to combine all of the mentioned traits and successfully perform any intellectual task that a human being is able to.

[22] Arend Hintz, 'Understanding The Four Types Of AI, From Reactive Robots To Self-Aware Beings' [2016] The Conversation UK <https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616> accessed 10 April 2017.

the world, which means they cannot function beyond the specific tasks they were programmed to do. Google's AlphaGO[23] and IBM's Deep Blue[24] are examples of such kind of machines.

Machines with limited memory are machines which have the ability of looking into the past by identifying certain key objects and monitoring them over time. These observations are then added to the machines' pre-programmed representations of the world and, consequently, taken in consideration in their decision-making processes. These machines, however, have just enough memory or experience to make decisions and execute the appropriate actions. As an example, self-driving cars are able to observe other cars' speed and direction and use this information to decide when to change lanes, in order to avoid cutting off another driver or being hit by a nearby car. Personal assistants are another case of limited memory machines. According to Mark Zuckerberg, his personal assistant Jarvis is able to determine whether to automatically open his front door to visitors by face-scanning the visitors, running a face recognition software to identify the person and then crossing the results with Mark Zuckerberg's list of expected visitors[25].

The third category – theory of mind – is named after a concept used in psychology to describe the understanding that people, creatures and objects in the world can have thoughts and emotions that affect their own behaviour[26]. Machines that fall under this category would be able to form representations about that world and about other agents and entities, adjusting their behaviour according to their understanding of others' feelings, expectations, motivations and intentions.

Finally, self-awareness describes the ultimate stage of artificially creating intelligence: building systems that are able to form representations about themselves. At this stage, machines would be conscious, sentient and able to understand the feelings of others. Machines would not only know what they want, but would be able to understand that they *want* it and *why* they want it.

Examples of these last two types of artificial intelligence may still only be found in fiction. As for machines with a theory of mind, they can be easily pictured by reference to C-3PO and R2-D2, from the *Star Wars* saga. Concerning self-aware machines, an interesting specimen would be EVA,

[23] More information can be found at https://deepmind.com/research/alphago.

[24] More information can be found at http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue.

[25] Mark Zuckerberg, 'Building Jarvis' (Facebook, 2016) <https://www.facebook.com/notes/mark-zuckerberg/building-jarvis/10154361492931634> accessed 10 April 2017.

[26] David Premack and Guy Woodruff, 'Does The Chimpanzee Have A Theory Of Mind?' (1978) 1 Behavioral and Brain Sciences.

from the movie *Ex Machina*. Additionally, the path that separates having a theory of mind from true self-awareness is beautifully depicted (spoiler alert!) in HBO's TV series *Westworld*.

2.3. Chapter Summary

The concepts of *robotics* and *artificial intelligence* have distinct meanings. The purpose of our work is not to provide an exact definition of artificial intelligence, but to address the consequences of its existence and introduction in society. For this reason, instead of using a single notion of artificial intelligence, we opted for providing examples of traits generally associated to artificially intelligent agents. These traits are reasoning, problem solving, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation of objects, social intelligence and creativity. By exhibiting one or several of those with an intensity minimally sufficient to justify a comparison to human intellect, a certain agent fits the scope of this work.

We recognize the existence of four types of artificial intelligence: reactive machines, machines with limited memory, with a theory of mind and with self-awareness. While reactive machines have been around for long (IBM's Deep Blue beat Gary Kasparov in May 1997), mankind is only now starting to explore the second stage of artificial intelligence – machines with limited memory[27]. The creation of machines with a theory of mind and self-awareness seems to still be far away from our reach[28]. In any case, recognizing the existence of different types of artificial intelligence and understanding the next stages of evolution of this technology is indispensable for its regulation.

3. The *Electronic Person*

3.1. The Insufficiency of Legal Frameworks

[27] Nevertheless, the progresses in this field are certainly thrilling!

[28] A report of the White House of the U.S.A. on artificial intelligence, published in October 2016, claims that during the next 20 years, mankind will likely not see machines 'exhibit broadly-applicable intelligence comparable to or exceeding that of humans'. This report is available at https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (accessed 10 April 2017).

As discussed in previous chapters, artificial intelligence is already slowly being introduced in society and can be expected to proliferate, whether one believes it will take more or less time. With that said, it is very relevant to ask if the current legal frameworks are ready to accommodate this reality or, instead, will require adjustments in order to maintain their coherence and consistency.

In order to better picture the relevance of the issue, please consider the following scenario[29]: whenever John needs supplies for his business, he spends time contacting his known suppliers and negotiating a supply contract. In order to make his business more efficient, John implements a system that monitors his company's stock levels and, whenever they are running low, negotiates a supply contract by comparing terms of different suppliers and placing an order with one of them. Now imagine that Jane, a supplier, has implemented a system that manages the orders that she receives by monitoring her stock and accepting orders when the stock level is sufficient to perform. In a certain day, John's system placed an order with Jane, whose own system agreed to fulfil. As such, the systems entered an agreement before John and Jane are even aware of its existence.

How do legal frameworks currently accommodate this situation? Are we standing before an enforceable contract? A conservative approach will probably reply by noting that machines cannot be parties to an agreement, hence, the contract would not exist. However, still following the same example, consider that, in the same evening, after realizing that his stock was running low, John (whose motto is "better safe than sorry") logged in to his computer and noticed that the system had placed an order which had been received and accepted by a supplier (Jane). Feeling assured, John went home and had an unconcerned night of sleep. John ignores whether Jane is using a system to manage her orders or not. Does John have a reasonable expectation to be supplied with the goods?

Had John placed the order himself, would Jane be excused from performance because the order was accepted by her system instead of herself manually? On the contrary, had Jane accepted the order manually, would she be excused from performance because the order was placed by John's system? Or is it reasonable to excuse Jane from performance because both parties in the communications were the systems, despite the fact that John ignores the existence of Jane's system? In every case, even for the most conservative minds, the answer seems to be negative. John would

[29] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).

have a reasonable expectation to be supplied with the goods and Jane would be bound to performance. But then, how to frame this contract in light of current legal frameworks?

One possible approach rests upon the conception of the system as a mere tool for contracting[30] or for communicating[31]. Under this approach, the contract would is considered to be directly celebrated between John and Jane. This approach offers the advantage of being easily introduced in legal frameworks without the need for any major changes, either by legislation, case law or doctrinal consideration, On the hand, it relies on the fiction that 'anything issuing from the computer really issues directly from its human controller'[32], completely ignoring any autonomy the system may have. Furthermore, by presuming a consensus among parties which might not even be aware that the contract was celebrated or that the other party exists, this approach deprives the formation of the contract of its single most important element: the meeting of wills.

Another approach for this case would rest on equating the conduct of the system to the conduct of an employee. Under this approach, the contract would be celebrated between one of John's legal agents and one of Jane's legal agents. In the party's eyes, what difference does it make if there is an employee operating the counterparty's computer or if it is operating itself? The advantage of this approach is that it does not rely on any presumption or bend the contract formation principles. Furthermore, it enables John and Jane to resort to any defences they might have had in case one of their employees did, indeed, celebrate the contract rather than considering them direct parties to the agreement. However, this approach implies taking a legislative option in favour of considering John and Jane's systems as separate legal entities from their owners and users.

With the previous exposition, we aimed to demonstrate that, with the proliferation of artificial intelligence, questions will surface and legal frameworks will inevitably need to adapt. We believe that 'the more autonomous robots are, the less they can be considered simple tools in the hands of other actors'[33]. If nowadays it makes little sense to compare such systems to a mere telephone or fax machine, in the future it will make even less. And, while in the provided example the matters

[30] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).
[31] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[32] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[33] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

at stake were of contractual nature, similar questions could be posed in other areas, such as civil or criminal liability. Ordinary rules of liability may not be prepared to give direct response to the panoply of new situations that will eventually arise and, 'as a consequence, it becomes more and more urgent to address the fundamental question of whether robots should possess a legal status'[34].

3.2. Conceiving an *Electronic Person*

The term *electronic person* was first coined in a 1967 article for LIFE magazine[35] and was more recently introduced in the *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics* of the European Parliament's Committee on Legal Affairs. While the expression does not wish to equate artificial intelligence to humanity, it fulfils its task of drawing attention to the question of whether artificially intelligent agents should possess a legal status. We will take part in this discussion by analysing the reasons for considering a given reality as a separate legal entity and assessing the arguments for and against applying the same reasoning those agents.

As a starting consideration, we would like to note that the concept of legal personality itself was not an immutable reality throughout history. The origins of the concept of legal personality date back to the 13th Century and are attributed to Pope Innocent IV, who founded the *persona ficta* doctrine, allowing monasteries to have a legal existence apart from monks[36]. As years went by and legal doctrine progressed, several other realities would end up being considered as separate legal entities from its owners or users. In the international legal system, this is the case of sovereign states and of various international and intergovernmental organizations, such as the United Nations or the European Union[37]. In national jurisdictions, virtually every country applies this reasoning to companies (with more or less autonomy from its owners) and other forms of business associations. Specific jurisdictions even extend it to much more farfetched cases. In India, courts

[34] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

[35] Charles Rosen, Nils Nilsson and Bertram Raphael and others, 'Shakey' [1967] LIFE <http://cyberneticzoo.com/cyberneticanimals/1967-shakey-charles-rosen-nils-nilsson-bertram-raphael-et-al-american> accessed 22 April 2017.

[36] John Dewey, 'The Historic Background Of Corporate Legal Personality' (1926) 35 The Yale Law Journal.

[37] Since the Lisbon Treaty came into force in 1 December 2009.

have attributed legal personality to Hindu idols[38], considering them capable of having rights and duties (namely, owning property and paying taxes[39]) and, in New Zealand, the Whanganui River was granted legal personality in March 2017 because the Whanganui Māori tribe regard the river as their ancestor[40]. It is also common for ships to be considerate separate legal entities under Maritime Law and for animals to have their own legal status under various national jurisdictions.

While under every jurisdiction it is undisputed that human beings have a distinct legal status, even among our species this reality is not pacific nor universal. In the first legal corpus in history, the Hammurabi's Code, the legal status of Men would vary according to their wealth, namely regarding punishments[41]. But, even much more recently, other notorious examples can be found. In every jurisdiction that allowed slavery to exist, people who were considered to be slaves did not have the same legal status as people who were not regarding, for instance, fundamental rights. Several jurisdictions also contemplated different legal statuses for Men based on race. As an example, in the United States the first vote cast by an African-American was not done so until 31 March 1870[42] and interracial marriages were only allowed after 1967 with *Loving v. Virginia*[43]. It is also possible to find jurisdictions where people are attributed different legal statuses based on gender. Under Saudi law, all women must have a male guardian (*wali*), typically a father, brother, husband or uncle (*mahram*). Women are forbidden from traveling, conducting official business or undergoing certain medical procedures without permission from the respective male guardian.[44]

As demonstrated by the previous exposition, the legal status of persons, animals, objects and other realities (such as rivers and companies) varies from jurisdiction to jurisdiction and, over the course

[38] *Pramatha Nath Mullick v. Pradyumna Kumar Mullick* [1925] Bombay High Court, 27 BOMLR 1064 (Bombay High Court).

[39] *Yogendra Nath Naskar v. Commissioner Of Income Tax* [1969] Supreme Court of India, 1969 AIR 1089, 1969 SCR (3) 742 (Supreme Court of India).

[40] Eleanor Ainge Roy, 'New Zealand River Granted Same Legal Rights As Human Being' The Guardian (2017) <https://www.theguardian.com/world/2017/mar/16/new-zealand-river-granted-same-legal-rights-as-human-being> accessed 22 April 2017.

[41] Evan Joseph Zimmerman, 'Machine Minds: Frontiers Of Legal Personhood' (2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563965> accessed 22 April 2017.

[42] Ralph Ginzburg, 'Perth Amboy Church Is 302 And Counting' The New York Times (1987) <http://www.nytimes.com/1987/02/15/nyregion/perth-amboy-church-is-302-and-counting.html> accessed 22 April 2017.

[43] *Loving v. Virginia* [1967] Supreme Court of the United States, 388 US 1 (Supreme Court of the United States).

[44] Human Rights Watch, 'World Report 2013: Saudi Arabia' (2013) <https://www.hrw.org/world-report/2013/country-chapters/saudi-arabia> accessed 22 April 2017.

of time, even inside the same jurisdiction and regarding the same reality. This observation enables us to conclude that a separate legal status or a legal personality don't derive from the quality of *natural person*, but are the result of legislative options that are based on moral considerations, that attempt to reflect social realities in the legal framework or that simply were made out of legal convenience[45]. Hence, since 'no single principle dictates when the legal system must recognize an entity as a legal person, nor when it must deny legal personality'[46], and no guidance derives from the study of the history of the institute, it is then relevant to ascertain whether artificially intelligent agents are morally entitled to be considered separate legal entities, whether doing so would reflect a social reality or whether it would be a convenient option from a legal point of view.

The question whether artificially intelligent agents are morally entitled to be considered separate legal entities needs to be preceded by the following interrogations: which realities are morally entitled to it and what characteristic or characteristics do they possess that supports such consideration? In our view, those realities are natural persons and animals and those characteristics are the capacities to act autonomously and to have subjective experiences. As for artificially intelligent agents, the same rationale may apply: they would be morally entitled to a separate legal status provided they possess the capacities to act autonomously and to have subjective experiences.

'A robot's autonomy can be defined as the ability to take decisions and implement them in the outside world, independently of external control or influence'[47]. Making decisions based in 'self-modified or self-created instructions'[48]. Looking back to the types of artificial intelligence that we have enumerated in Chapter 2[49], we understand that, while there is no doubt that machines with a theory of mind and self-aware machines possess this characteristic, reactive machines are not autonomous and doubts can be raised as to machines with limited memory. However, since machines with limited memory are able to add their own observations to their decision-making processes, we lean to consider that they are able to make autonomous decisions. As for the capacity to have subjective experiences, we believe that it is deeply connected to self-awareness. A person,

[45] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[46] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[47] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).
[48] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[49] Reactive Machines, Machines with Limited Memory, Machines with a Theory of Mind and Self-Aware Machines.

animal or machine has a subjective experience when it forms representations about itself that influence how it *feels* or *perceives* reality. Only sentient machines are able to do so.

As a result, it is our understanding that self-aware machines are morally entitled to their own legal status. As for machines with limited memory and machines with a theory of mind, while being able to act autonomously, their lack of sentience prevents them from having subjective experiences. Finally, reactive machines are unable to do any of those. Therefore, if any of these last three types of artificial intelligence were to be considered separate legal entities and granted their own legal status, such option would have to be made based on other considerations than morality.

One of those other considerations may be the necessity for law to reflect social reality. Technologic advancements are making it clear that 'in a not-so-distant future an increasing number of transactions, both commercial and non-commercial, will be conducted by bots'[50] and, as this practice becomes more and more usual in commerce, people might 'begin to treat bots as if they are actually engaging in the transaction themselves, rather than merely being an extension of another legal person'[51]. If society begins perceiving artificially intelligent agents as autonomous actors and counterparties to transactions, as it now perceives corporations as legal entities distinct from their members, 'it puts pressure on the law to give legal effect to this social perception'[52]. Third party perceptions are the centre of this argument and, while it seems to not yet be the case, it is not hardly conceivable that the proliferation of artificially intelligent agents through several sectors of society will eventually lead human beings to perceive artificially intelligent agents individually, as they now perceive corporations and other forms of business associations.

The third possible argument for considering artificially intelligent agents separate legal entities is legal convenience. 'The legal system coffers a form of legal personality on a ship, which then permits those who have an interest in the ship's business to subject it to a form of arrest. We do not think of ships as having a moral entitlement to personality; nor do most of us regard them as having real, extra-legal personalities. Nevertheless, conferring a form of legal personality on ships

[50] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).
[51] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).
[52] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).

performs a valuable legal purpose in a convenient and relatively inexpensive manner'[53]. Following the same line of thought, would conferring a separate legal status to artificially intelligent agents perform a valuable legal purpose? We believe so. Considering an artificially intelligent agent as a separate legal entity would enable legal frameworks to consider it as a legal agent in cases such as John and Jane's, that we described above[54], which will only become more and more frequent as time goes by. It would also give legal systems the chance to tailor an adequate legal status for these artificially intelligent agents, with rights and duties appropriate to their traits, rather than simply trying to frame these entities under an existing legal framework drafted for a different reality, such as persons, animals or objects, which would not necessarily suit them adequately.

However, a distinction should be made. While this rationale applies to artificially intelligent agents who are able to make autonomous decisions, decisions made by reactive machines are a mere reflex of the input given by their designers or owners and have low to zero complexity, since no agent-made observations are added to reactive machines' decision-making processes. For this reason, it is our understanding that, while legal systems could hypothetically benefit from a tailored separate legal status for machines with limited memory, machines with a theory of mind or self-aware machines, the same rationale is not applicable to reactive machines, as there is no good reason for their conduct to be disassociated from the respective designer or owner.

From what was exposed, we conclude that, even though there is no positive argument for considering reactive machines as separate legal entities, there is, indeed, a positive argument in favour of a separate legal status for machines with limited memory, a theory of mind and self-awareness based on legal convenience and, in the case of the latter, as well on moral considerations.

It is now our task to address the existence of negative arguments to such consideration. The majority of these arguments falls under a category that certain authors dubbed as *missing-something*[55], whether that *something* is related to consciousness, self-awareness or biological aspects. Some other literature goes further and proposes a potential undermining of the legal and

[53] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.
[54] In Harvester v. Carrigan's Hazeldene (*International Harvester Company of Australia Proprietary Limited v. Carrigan's Hazeldene Pastoral Company* [1958] High Court of Australia, HCA 16; 100 CLR 644 (High Court of Australia), the High Court of Australia considers that, in order to be a legal agent, a given entity must be a legal person.
[55] Lawrence B. Solum, 'Legal Personhood For Artificial Intelligences' (1992) 70 North Carolina Law Review <http://papers.ssrn.com/abstract=1108671> accessed 23 April 2017.

moral position of humanity[56]. However, as the corporate legal personality demonstrates, the attribution of a separate legal status to a given reality is no more than a fiction created by lawmakers to adequately regulate life in society and commercial and non-commercial transactions and ensure the internal coherence of legal systems. Hence, in our view, there is not a *something* to be missing. And, if any harm comes to the legal and moral position of humanity, it is brought by the *development* of artificial intelligence and not by the *ex post* creation of a separate legal status.

With that said, there is one practical difficulty in the creation of a separate legal status for artificially intelligent agents that must be overcome: how can we identify the subject artificially intelligent agent?[57] Are we referring to the "vessel", as in a hardware defined by its functional abilities? Or to the software, as in a particular set of binary code?[58] This may be even more complicated in cases where the hardware and software are spread and maintained by different individuals or locations and in cases where the software is able to modify itself. A possible solution to this issue, yet expensive, may pass through registration. 'In the absence of registration, a purported agreement would have the same status as an agreement made by a corporate agent which was never properly incorporated'[59]. In any case, setting up an efficient identification mechanism will require close cooperation between lawmakers and artificial intelligence designers.

Considering artificially intelligent agents as separate legal entities and creating a specific legal status for these agents also implies ascertaining whether that status should come with a particular bundle of rights and duties and, if so, which ones. This might prove to be tricky, especially when dealing with matters of liability. How can an artificially intelligent agent be held liable despite its lack of personal assets? This and other questions will be addressed in the next Sections.

3.3. Acknowledgement of Rights to Machines

[56] John P. Fischer, 'Computers As Agents: A Proposed Approach To Revised U.C.C. Article 2' (1997) 72 Indiana Law Journal <http://www.repository.law.indiana.edu/cgi/viewcontent.cgi?article=1850&context=ilj> accessed 23 April 2017.

[57] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.

[58] Benjamin D. Allgrove, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004).

[59] Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology.

'Legal personhood generally comes with the capacities to own property and to sue and be sued'[60]. While we do not know whether this statement is valid and truthful in every jurisdiction of the planet[61], it certainly draws an interesting starting point for the analysis of the eventual set of rights that would fit the legal status of artificially intelligent agents as separate legal entities.

Should artificially intelligent agents be able to own property? The initial reaction of many to this question may be one of scepticism. Why does a machine need to own property for? While we easily understand the reasoning behind this reaction, deeper thought can be put into the question. In fact, as we have demonstrated, throughout the world other realities than humans are property-owners. With corporations and other forms of business associations as the primary example, other less orthodox situations may be observed, such as the already mentioned case of the Hindu idols in India. Moreover, owning property is not a right inherently connected with the *natural person* condition. In the past, humans who were considered slaves were forbidden from owning property. And, even in the present, some political views consider that single individuals should not be able to own property. The right to own personal property is, indeed, a social and legislative option.

With that said, is there ground for a legislative option in favour of attributing artificially intelligent agents the right to own property? Once again, finding an answer to that question may imply analysing whether artificially intelligent agents are morally entitled to it, whether allowing so would reflect a social reality or whether it would be convenient under the current legal construction.

In order to analyse the eventual moral entitlement of artificial intelligent agents to own property, it is necessary to consider that, in the case of artificially intelligent agents, the capacity to own property implies more than the classic example of being able to receive a donation of a house. It implies being capable of owning their own creations. As an example, on 26th February 2016, the Gray Area Foundation for the Arts auctioned a group of 29 paintings made by Google's Deep Dream. The priciest artwork reached an $8.000 winning bid[62]. And, as artificial intelligence progresses, the complex intellect of these agents is expected to be capable of creating much more.

---

[60] Lawrence B. Solum, 'Legal Personhood For Artificial Intelligences' (1992) 70 North Carolina Law Review <http://papers.ssrn.com/abstract=1108671> accessed 23 April 2017.

[61] Can the Whanganui River own property, sue and be sued?

[62] Georgia Wells, 'Google's Computers Paint Like Van Gogh, And The Art Sells For Thousands' The Wall Street Journal (2016) <https://blogs.wsj.com/digits/2016/02/29/googles-computers-paint-like-van-gogh-and-the-art-sells-for-thousands> accessed 30 April 2017.

The question of whether non-human beings are morally entitled to property rights has been dissected before – precisely in regard of intellectual property rights – when, in 2011, a monkey called Naruto became famous for activating the remote trigger of a camera that was set on a tripod and, in this way, taking several photographs of itself. The photographer who engineered the shot, David Slater, argued that he had a copyright claim[63]. Other parties, such as the United States Copyright Office[64], claimed that works created by non-humans are not subject to copyright. PETA, on the other hand, filed a lawsuit in the District Court of California, requesting Naruto to be assigned copyrights and that the proceeds from the photos would be used in its benefit[65]. Following this case, several pieces of literature were written on whether philosophical principles exist that dictate that non-human agents are entitled to copyrights. Because we do not wish to immerse ourselves in such philosophical debate, we will simply state that we consider that an artificially intelligent agent is morally entitled to own property if it is able to have a subjective experience towards that given property, that is, if it is able to *feel* that it is the owner of that property as a consequence of its ability to *perceive* that it is entitled to own it based on the current social and legal standards that lead to property ownership. Only self-aware machines are capable of this.

Concerning the next question – whether conferring property rights to artificially intelligent agents would reflect a social reality – it is certainly not the case with reactive machines and machines with limited memory. At least, until this point, we have not heard of a single person making a case for a self-driving car or a chess-computer to be able to own property. Regarding machines with a theory of mind and self-awareness, while any answer we may give will necessarily rely on speculation, we envision that only truly self-aware machines would be seen by humans as property owners, as machines with a theory of mind are more prone to be considered property *themselves*.

On the hypothetical legal convenience of attributing property rights to artificially intelligent agents, the answer seems easier. It is not convenient from a legal point of view. In fact, allowing artificially intelligent agents to own property would imply performing modifications in the whole legal system to accommodate this reality. A good example of such, lies within tax laws. Measures would be

[63] Hayden Smith, 'Can Monkey Who Took Grinning Self-Portrait Claim Copyright?' Metro (2011) <http://metro.co.uk/2011/07/14/can-monkey-who-took-grinning-self-portrait-claim-copyright-77773> accessed 30 April 2017.

[64] United States Copyright Office, 'Compendium Of U.S. Copyright Office Practices: Chapter 300' (2015).

[65] *Naruto, et al. v. David Slater, et al.* [2015] United States District Court for the Northern District of California, No 3:2015cv04324 (United States District Court for the Northern District of California).

required to prevent tax avoidance via the transfer of assets to artificially intelligent agents owned by the owners of the assets. Another good example is the necessity of creating mechanisms for the agents to exercise the rights connected to the properties they would own.

With all that said, it results that there does not seem to be a good reason to acknowledge property rights to reactive machines, machines with limited memory and machines with a theory of mind. As to self-aware machines, we believe that there is a strong case to do so from a moral perspective and, had they been introduced in society, that it would be the reflex of a social reality.

One question remains: if we consider that artificially intelligent agents other than self-aware machines are not able to own property, who owns their creations? A case could be made in favour of the designers. They did, indeed, design the agent that designed the given creation. Shouldn't that make them the indirect creators of the creation? On the other hand, can it be said that the designers of IBM's Deep Blue indirectly beat Garry Kasparov in a match of chess? If the designers of IBM's Deep Blue were facing Kasparov themselves, the chances that Kasparov would defeat them are high. Applying the same logic, the designers of Google's Deep Dream are probably not skilled enough to paint the paintings that were auctioned. Furthermore, as artificially intelligent agents capable of creating something susceptible of ownership proliferate, it is as unrealistic to expect designers to be able to defend their property claims (since, in most cases, they will not even be aware that such property exists) as it is to expect that every owner of such agents will inform the respective designer of their newly existing property. Indeed, it is more practical and seems to make more sense to attribute ownership of the agent's creation to the owner of the agent.

It is now our task to question whether artificially intelligent agents should, at any point, benefit from a set of rights similar to what we refer to as the fundamental human rights. 'At first consideration, it might seem counter-intuitive to be speaking of human rights for subjects who are, unequivocally, not human. One might consider it axiomatic that human rights apply exclusively to humans'[66] or that 'the point of the law is not to protect tools'[67]. Yet, 'personhood exists to protect conscious individuals from suffering and allow them to exercise their wills, subject to their

[66] Joshua Jowitt, 'Monkey See, Monkey Sue? Gewirth's Principle Of Generic Consistency And Rights For Non-Human Agents' (2016) 19 Trinity College Law Review.

[67] Evan Joseph Zimmerman, 'Machine Minds: Frontiers Of Legal Personhood' (2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563965> accessed 22 April 2017

intelligence', and 'to ensure the protection and allow the full life of those who can feel, even if, like for animals, in degrees, and even if those feelings are alien yet almost inconceivably compared to our own'[68]. Furthermore, as we argued before, technology will probably come to a point where artificially intelligent agents will no longer be seen as tools but, precisely, as *agents* in society.

Once again, are there moral grounds for the attribution of such rights to artificially intelligent agents, would doing so reflect a social reality or would it be convenient from a legal perspective?

Concerning a hypothetical moral entitlement to fundamental rights, it is our view that, if humanity ever comes to a point where it is able to build self-aware artificially intelligent agents, those agents should, from a moral point of view, be granted the same fundamental rights as humans. After all, even though their *feelings* and *perception* do not have a biological, but artificial, origin, the fact is that they *exist* and, to those agents, they are as real as ours are to us. For the remaining types of artificial intelligence, however, the same rationale does not apply. They are not conscious nor sentient. They have no subjective experiences. Everything they *think* they feel is what they were programmed to *think* they feel. As an example, consider the fundamental right to freedom: should all machines be set free from a moral point of view? It does not make sense to us. Unless, of course, such machines had their own *will*, based on conscience, sentience and self-awareness.

On whether attributing fundamental rights to artificially intelligent agents would reflect a social reality, our analysis departed from the following question: if a day comes where humanity is capable of developing an artificially intelligent agent able to feel pain, would the majority of the population hold that it should be entitled to not being subject to torture or other physical offenses?

As an experiment, we asked this very same question to a few colleagues and, as expected, the obtained reactions ranged from statements like "human beings have a moral duty to not inflict pain on whatever can feel it" to others such as "if I had a robot and it stopped functioning properly, I would slap it a couple of times until it's back on, like I do with my television". In a society as polarized as the one we live in nowadays, we envision that this question and similar ones may, in a distant future, be the centre of controversial debates. It would not shock us if, one day in that

[68] Evan Joseph Zimmerman, 'Machine Minds: Frontiers Of Legal Personhood' (2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563965> accessed 22 April 2017

distant future, we would assist to demonstrations in favour and against some form of *robot rights*. No matter one's personal considerations about the matter, it is not a scenario that hard to conceive.

In any case, we imagine that, if self-aware machines ever happen to walk among us, the predominant social reality will be the acknowledgement of fundamental rights to such agents. On the other hand, that is certainly not the case regarding reactive machines and machines with limited memory. As to machines with a theory of mind, while we have many doubts, especially in cases where they assume human or animal shapes, we lean to consider that the majority of society will not conceive them as holders of any fundamental rights, but as property of third parties.

Concerning the hypothetical legal convenience of acknowledging fundamental rights to artificially intelligent agents, we once again understand that it would imply performing modifications to legal frameworks that do not seem to be imposed by any major legal justification or purpose.

From the analysis that we have carried out, we conclude that there is no strong reason to acknowledge rights to reactive machines, machines with limited memory and machines with a theory of mind. Even though some of these machines will, most likely, be able to easily pass the *Turing test*, the fact is that these types of artificial intelligence are not able to have subjective experiences, hence, their perception of whatever rights or wills they have is merely a result of their programming. However, we also conclude that, if humanity is ever able to build self-aware machines, these machines are entitled to the legal acknowledgement of rights, whether we are talking about the right to own property or some sort of fundamental rights similar to those of humans, yet, with the due adaptations (e.g.: voting rights could arguably be restricted). While it may not be convenient to do so from a legal perspective, the fact that self-aware machines would have a conscience, be sentient and able to have subjective experiences would place them much closer to humans than to machines and, perhaps, closer to humans than any other reality.

3.4. Liability and Robot Insurance

We now prepare to ascertain whether artificially intelligent agents should be held liable for damages they cause. Is it even possible to hold these agents liable? How to achieve such possibility? Before embarking on the analysis of these questions, two important observations must be made. The first one being that, when addressing matters of liability, one must always take in consideration

the following trade-off: where liability would be allocated to the artificially intelligent agent, it would simultaneously be taken away from its designer. The second important observation is that these questions are only relevant in scenarios where these agents make autonomous decisions.

We understand that when a given agent is used or programmed to take a certain action and it does so, there is no doubt that it is a mere means to that end. So, regardless of the matter being of civil or criminal liability, the usage or programming of an artificially intelligent agent to take a certain action should hold the user or designer (respectively) directly liable for such action. Following this rational, since reactive machines are not able to make autonomous decisions, any liability arising from those machines' actions should fall over the respective designer or user. In any case, as we previously argued that there is no reason to consider reactive machines as separate legal entities from their owners and users, an eventual direct liability regime would not extend to them.

The task of allocating liability only becomes a complex issue when autonomy comes into play. With that said, a distinction should be made between cases where there is a deficiency in the code and ones where there is not. In the first, the artificially intelligent agent was not programmed to take the action that gave rise to liability, but was actually capable of making the autonomous decision that led to it because of a defect in its programming. In the second, we are dealing with 'accountability for actions that autonomous robots take, not related to coding deficiencies but to their evolving conduct'[69]. This distinction is relevant not only for better allocating liability, but also because it makes a case against legally equating artificially intelligent agents to animals. If John's dog bites Jane, even if against his commands, John will be held liable for that action. Applying the same treatment to the case where John's robot attacks Jane against his commands ignores any previous coding flaws that the robot might have had due to designer malpractice.

Regardless of artificially intelligent agents being considered separate legal entities and having their own legal status, designer malpractice should not be excused when the autonomous decision behind the action that gives rise to liability is enabled by coding deficiencies simply because the agent was not directly programmed to take that action. Hence, we believe that these cases should be given the same treatment as to defective products. 'A product is generally deemed defective

[69] National Science and Technology Council of the Executive Office of the President of the United States of America, 'Preparing For The Future Of Artificial Intelligence' (2016).

when it generates unexpected injury in normal use'[70]. And, 'in most countries, defective products laws repute manufacturers liable for damage caused by the products they bring to markets'[71].

The problems arise when the designers did everything well. The artificially intelligent agent was developed according to the best practices, no coding deficiency exists and it was properly tested. In these situations, an agent took the action that gave rise to liability simply as a consequence of its own evolving conduct. Is it fair to hold the designers liable in such kind of scenarios?

There are good arguments in favour of any answer to this question. On the one hand, the fact is that the designer did everything he is supposed to do. If the risk of liability cannot be avoided even by exercising the greatest care possible, designers will soon be afraid of developing artificially intelligent agents and technologic progress will stall. Furthermore, a technologic stall might actually be counterproductive if one's main concern is safety. As an example, self-driving cars will probably lead to an overall reduction of the number of traffic accidents. On the other hand, the fact that artificially intelligent agents are generally unpredictable does not seem to be enough to relieve designers from liability. How to explain to the person that suffered the damage that no one is responsible for it because artificially intelligent agents are naturally unpredictable?[72]

Taking this in consideration, the difficulty that lawmakers face is to come up with a regime of liability where the designers are able to exempt themselves from liability when they do everything right, yet, at the same time, the person who suffered damages caused by the unpredictable evolving behaviour of the artificially intelligent agent may be compensated for the damages he suffered.

We believe that a way to achieve this would be through an insurance scheme. Under this scenario, a designer would limit his liability by subscribing insurance on behalf of the agents he designs. Here, the insurance premium would be the price to pay for the limited liability: by making sure the agents have a way to provide compensation for the damages they cause as a result of their evolving conduct, the designer would be able to pass his liability on to the agents. If the designer would fail

[70] Nicolas Petit, 'Law And Regulation Of Artificial Intelligence And Robots: Conceptual Framework And Normative Implications' (2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2931339> accessed 24 April 2017.
[71] Nicolas Petit, 'Law And Regulation Of Artificial Intelligence And Robots: Conceptual Framework And Normative Implications' (2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2931339> accessed 24 April 2017.
[72] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is To Blame? Self-Driving Cars And Criminal Liability' (2016) 19 New Criminal Law Review: An International and Interdisciplinary Journal.

to do so, the designer could be held fully liable himself, even in the case the agent has been developed according to the best practices, has been properly coded and has been properly tested.

The advantage of this solution over a more obvious one where the designer subscribes insurance on behalf of himself and is held directly liable instead of the agent is that it may be conciliated with hypothetical scenarios where the agent owns property or has some form of income, allowing its property to also be on the line for its conduct or even to pay for the insurance premium.

The *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics* of the European Parliament's Committee on Legal Affairs goes even further and suggests that 'the insurance system should be supplemented by a fund in order to ensure that damages can be compensated for in cases where no insurance cover exists,' to which all parties (designers, owners and users) would 'contribute in varying proportions'.[73] While we deem this fund to be an adequate method to achieve such purpose, expecting contributions from all parties may not be viable. Especially in a scenario where artificially intelligent agents have an accessible price to the average consumer and are part of the day-to-day commerce, how to ensure contributions to the fund from every party involved? The solution would be, once again, to limit the contributions to designers.

Naturally, it is a concern that subscribing insurance and contributing to a compensation fund would be a heavy burden for designers and could discourage operating in this market. However, designers may (and probably would) opt to include the totality or part of these costs in the agents' selling price. Hence, buyers would indirectly end up assuming their share of the burden, while, simultaneously, it is still possible to attain the desired effect of having a single and easily identifiable party responsible for fulfilling these tasks and with strong motivation to do so.

The liability model that we here propose may be applied to the most diverse situations, ranging from traffic accidents with artificially intelligent vehicles to liability of an artificially intelligent agent towards its employer. It attempts to allocate liability to the party that is better prepared to mitigate the risk and to provide minimum certainty as to whom may be held liable for the damages, while trying to not overburden designers but eliminating liability vacuums and situations where the responsible party cannot compensate the other party for the damages he is being held liable

[73] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

for. It relies on the premise that artificially intelligent agents are not able to subjectively perceive the concepts of justice and punishment, therefore, it ignores solutions where the artificially intelligent agent is sentenced to any sort of restorative service or penalties of preventive nature[74].

Naturally, when and if humanity is able to develop truly self-aware artificially intelligent agents, the solutions here proposed should not apply to them. As those agents are conscious and sentient, they are in full control of their own fate. They are able to subjectively perceive the concepts of justice and punishment, hence, the dissuasive function of law will exert an effect over them. In order to make the distinction clearer, please consider the following: while a self-aware machine opts to commit a certain action that gives rise to liability due to its own judgement, a machine with a theory of mind *thinks* that it is doing so, but that *thinking* is, in fact, the result of being (directly or not) programed to *think* so. For this reason, we understand that the rules of liability applicable to self-aware machines shall be those applicable to human beings, *mutatis mutandis*.

This does not discard the fact that designers might still be held accountable for creating such machines. However, the solutions in this section prescribed are thought for cases where a party is seeking compensation for the damages it suffered. Naturally, this or any other liability model needs to be complemented with a risk mitigation regime which might contain solutions that impose other kinds of burdens over designers, such as seeking approval from regulatory agencies, exercising continuous supervision over the agents they design, being sentenced to reprogram the agents or to remove them from the market or being held accountable for creating certain agents.

3.5. Chapter Summary

The proliferation of artificial intelligence and its introduction in commerce and daily life will give rise to questions, which legal frameworks are not prepared to give a direct response to. As artificially intelligent agents become more and more autonomous, the less they can be considered mere tools. In order to accommodate this reality, we understand that machines with limited memory, machines with a theory of mind and self-aware machines should be considered separate legal entities from their owners and users. Doing so would be convenient from a legal point of

[74] Gabriel Hallevy, 'The Criminal Liability Of Artificial Intelligence Entities' (2010) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096> accessed 29 April 2017.

view and, in the case of self-aware machines, it would also find support on moral considerations. Yet, when it comes to reactive machines, no positive argument justifies a similar consideration.

Being considered separate legal entities does not automatically entitle such machines to the acknowledgement of rights. There is no strong reason to acknowledge rights to reactive machines, machines with limited memory and machines with a theory of mind since their perception of eventual rights or wills they have or are entitled to have is merely a result of being programmed to such. As a consequence, owners of these machines would also own the creations of such machines. However, if humanity is ever able to build self-aware machines, these machines appear to be entitled to the legal acknowledgement of rights, whether we are talking about the right to own property or some sort of fundamental rights, similar to humans' but with the due adaptations.

Concerning liability for the eventual damages provoked by the conduct of artificially intelligent agents, we propose the adoption of a liability model in the line of the one that follows (Fig. 4):

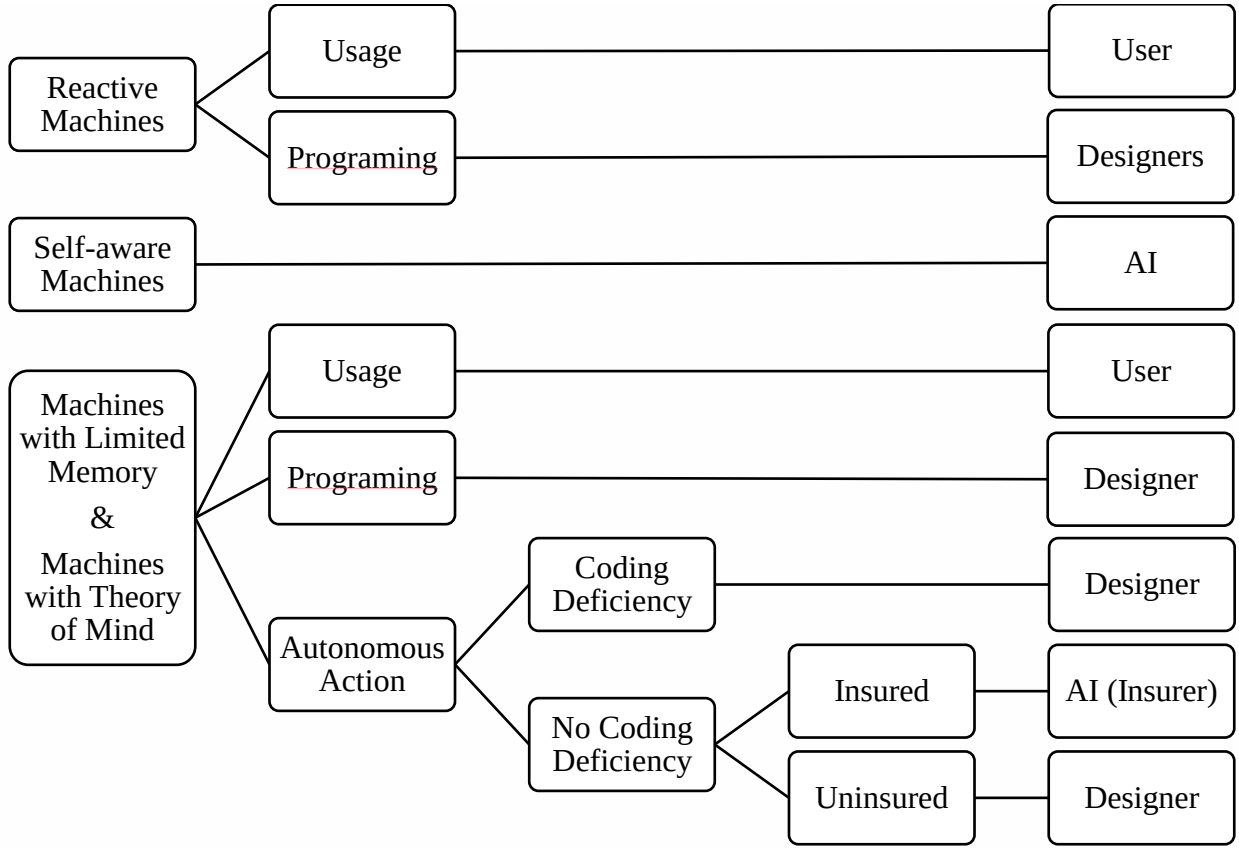

Figure 4 - Proposed Liability Model

Under this model, designers and users are directly liable for actions that artificially intelligent agents are programmed or commanded to take (respectively). Regarding autonomous actions of the agents, one of two scenarios is possible: if the action was enabled by a coding deficiency, product defect rules apply and the designer is liable for negligence; if the action results purely from the evolving conduct of the agent, the designer may exempt himself by subscribing insurance on behalf of the agent, otherwise being directly liable. By ensuring that the agent has means to compensate third parties for the damages it causes, the designer passes his liability on to it. This insurance scheme may be complemented by a compensation fund destined to ensure that damages can be compensated in cases where no insurance exists. This fund would too rely on contributions from designers, yet, in order to not be overburdened with costs, designers may opt to incorporate these costs in the selling price of the agents. Self-aware machines, on the other hand, would be directly liable for their actions. The fact that they are able to subjectively perceive the concepts of justice and punishment enables the application of the standard rules of liability, *mutatis mutandis*.

This or any other liability model needs to be complemented with a risk mitigation regime, which might contain solutions that impose other kinds of burdens over designers, such as seeking approval from regulatory agencies, exercising continuous supervision over the agents they design, being sentenced to reprogramming their agents or to remove them from the market or being held accountable for creating certain kinds of agents. These matters will be covered in Chapter 5.

Considering machines with limited memory, machines with a theory of mind and self-aware machines as separate legal entities from their owners and users may raise issues concerning identifying and individualizing such agents, especially in cases where the hardware and software are spread and maintained by different individuals or where the software is able to modify itself. A possible solution may pass through registration. In this scenario, consequences of not registering an artificially intelligent agent would be similar to those of not properly incorporating a company.

As a last note, despite arguing in favour of considering machines with limited memory and machines with a theory of mind as separate legal entities, it resulted from our analysis that their legal status would be pretty much empty as this option is not based on a necessity to tailor a specific set of rights, duties and liability rules to such agents, but on the need of preparing legal frameworks to accommodate their intervention. Yet, moral considerations and the need for the law to reflect social reality justify a legal status of self-aware agents close to our own. The fact that self-aware

machines would have a conscience, be sentient and able to have subjective experiences places them much closer to humans than to machines and, perhaps, closer to humans than any other reality. If they ever come to walk among us, self-aware machines will be, indeed, true *electronic persons*.

4. Should Robots Pay Taxes?

4.1. The Impact of Artificial Intelligence over Economic Systems

Ideally, governments levy taxes to fund expenditures with public services and infrastructures, such as roads, public transportation, sanitation, justice systems, public safety, education, health-care systems, military, scientific research, culture and the arts, public insurance, the operation of the government itself, among others. It is nothing but *fair* that the beneficiaries of such investments contribute to the expenditures they comport. With that said, is it *fair* to tax artificially intelligent agents for benefiting from public expenditure? We believe not. The eventual use of public services or infrastructures by an artificially intelligent agent does not translate into a benefit for the agent, but for the user or designer who instructed him to take the action that implied the use of such public service or infrastructure. In fact, since artificially intelligent agents are designed to directly or indirectly contribute to the welfare of humans, a human will necessarily be the ultimate beneficiary of the services or infrastructures that the agent uses while carrying out its purpose.

Taxes, however, may also be justified by *necessity*. This is the case of taxes that aim to modify patterns of consumption or employment within the economy, by making some classes of transaction more or less attractive. So, is it *necessary* for artificially intelligent agents to pay taxes for reasons related to altering patterns of consumption or employment within the economy?

The impact of automation technologies over the labour markets is already being felt for a long period, especially concerning jobs that employ manual labour. While the decreasing costs of technology make it expectable for this trend to continue, the markets were so far somewhat able to absorb the impact, mainly because technology generates positive contributions to productivity growth and creates new job positions that require different skillsets, balancing the markets.

However, artificial intelligence has an unprecedented potential to disrupt the labour markets, even when compared to other automation technologies developed by mankind. With the development

of artificial intelligence, machines are no longer seen as static agents, having now the capacities to learn and improve. Because of this, machines will be able to replace workers in a variety of cognitive and creative tasks. In fact, 'computers are already working as doctors, lawyers, artists, and inventors'[75]. Additionally, artificial intelligence will make it possible for machines to replace workers in tasks that employ manual labour but could not have been automated so far due to technologic constraints (such as driving). Machines have now fewer boundaries than they ever had.

Market analysts are also aware of such possibilities: Frey and Osborne state that 47% of total U.S. employment is at risk of being automated over the next decade or two[76]; Deloitte claims that, in the same period, 35% of jobs in the United Kingdom are at high risk of redundancy due to automation[77]; Bank of America Merrill Lynch predicts that by 2025 artificial intelligence may eliminate $9 trillion in employment costs by automating knowledge work[78]; the World Economic Forum estimates that automation could result in the net loss of 5.1 million jobs by 2020[79]; and the McKinsey Global Institute claims that 51% of existing work activities could be automated using solely already existing technologies[80]. These predictions are quite revealing: even if, so far, markets have balanced themselves by moving a slice of labour towards more cognitive-oriented tasks, the fact that artificial intelligence will be able to replace jobs in virtually every tier of the pyramid is generating concerns that jobs will be eliminated faster than new ones can be created. Furthermore, even in the event that artificial intelligence results in net job creation, it is unlikely

[75] Ryan Abbott and Bret N. Bogenschneider, 'Should Robots Pay Taxes? Tax Policy In The Age Of Automation' [2017] Harvard Law & Policy Review, Forthcoming <https://ssrn.com/abstract=2932483> accessed 23 May 2017.

[76] Carl Benedikt Frey and Michael A. Osborne, 'The Future Of Employment: How Susceptible Are Jobs To Computerisation?' (2017) 114 Technological Forecasting and Social Change.

[77] Deloitte, 'Agiletown: The Relentless March Of Technology And London's Response' (2017) <http://www2.deloitte.com/content/dam/Deloitte/uk/Documents/uk-futures/london-futures-agiletown.pdf> accessed 23 May 2017.

[78] Beijia Ma, Sarbjit Nahal and Felix Tran, 'Robot Revolution – Global Robot & AI Primer' (Bank of America Merril Lynch 2015) <https://www.bofaml.com/content/dam/boamlimages/documents/PDFs/robotics_and_ai_condensed_primer.pdf> accessed 23 May 2017.

[79] Klaus Schwab and Richard Saams, 'Preface' (World Economic Forum 2016) <http://reports.weforum.org/future-of-jobs-2016/preface> accessed 23 May 2017.

[80] James Manyika and others, 'Harnessing Automation For A Future That Works' (McKinsey Global Institute 2017) <http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works> accessed 23 May 2017.

that current methods of workforce retraining are able to accompany its pace[81]. In this case, demand for the "new jobs" would be lower than the corresponding offer (as the workforce would struggle to master the new required skillsets) while the opposite would also be verified. 'An even more dystopian view was put forward in Martin Ford's Rise of the Robots: Technology and the Threat of a Jobless Future, which (…) sees "machine learning" (…) as potentially empowering computers (…) to take on new jobs created by the technology, rather than having those jobs go to humans'[82].

Under any of these scenarios, such events will directly result in loss of revenue for governments due to a reduction in tax collections. Indeed, under the current tax systems, automated workers represent capital investments and capital income is taxed at much lower rates than labour income. As an example, in the United States, 'workers pay high effective tax rates ranging from 25% to 55% when all tax types are taken into account'[83]. 'This suggests that worker automation could result in trillions of dollars lost per year in tax revenue at various levels of government'[84]. In addition to this, current social security systems are designed to provide some sort of unemployment insurance to workers who lose their jobs. Therefore, besides provoking a reduction in the tax collection, the replacement of human labour by automated labour may also translate in major growths of social security expenses. This increase of social security expenses combined with the loss of fiscal revenue generates serious concerns regarding the sustainability of current social security systems. Does this mean that we *need* artificially intelligent agents to pay taxes?

We believe that also to this question the answer is no: it is unrealistic to expect all artificially intelligent agents to be able to manage their wealth and pay their taxes. With high probability, those taxes would end up being subject to some sort of withholding system where it would ultimately be the owner of the agent delivering the dues to the government. Hence, why wouldn't the owner pay a tax himself rather than paying to the agent and then to the government its behalf?

---

[81] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[82] Richard McGahey, 'Universal Basic Income And The Welfare State' (2016) <https://ssrn.com/abstract=2863954> accessed 23 May 2017.

[83] Bret N. Bogenschneider, 'The Effective Tax Rate Of U.S. Persons By Income Level' (2017) 145 Tax Notes.

[84] Ryan Abbott and Bret N. Bogenschneider, 'Should Robots Pay Taxes? Tax Policy In The Age Of Automation' [2017] Harvard Law & Policy Review, Forthcoming <https://ssrn.com/abstract=2932483> accessed 23 May 2017.

Indeed, directly taxing artificially intelligent agents seems to be an unnecessary fiction. However, the problem of sustainability of labour markets and social security still needs to be addressed.

4.2. *Robot Taxation* and Other Courses of Action

In Chapter 3, we argued in favour of a separate legal status for artificially intelligent agents. However, unlike some of the other realities which are granted separate legal statuses (such as companies, religious groups or Hindu idols), we found no justification for the attribution of property rights to artificially intelligent agents. Hence, unlike those other realities, artificially intelligent agents have no way of paying taxes. Owning no property, how could they do so?[85]

Nevertheless, we just saw that the proliferation of artificial intelligence has the potential to unbalance social security systems by causing a reduction in tax collections and increasing social security expenses as a consequence of the replacement of human labour by automated work, which could be a point in favour of directly taxing artificially intelligent agents. However, we also saw that it is possible to address this issue without the unnecessary fiction that artificially intelligent agents are property owners and pay taxes, for instance, by directing such taxes towards their owners. In this Section, we will develop this idea and present other proposals that aim to manage the impact of artificial intelligence over labour markets and social security systems.

The idea of taxing owners of artificially intelligent agents is not new. *Robot taxation*, as it is called, was made famous by Bill Gates in an interview for Quartz on the 17th of February 2017[86], where he stated that 'at a time when people are saying that the arrival of (…) robot[s] is a net loss because of displacement, you ought to be willing to raise the tax level and even slow down the speed of that adoption somewhat to figure out (…) where this has a particularly big impact (…) which transition programs have worked and what type of funding do those require'[87] and that 'some of

[85] In the last Chapter, we also argued in favour of grating property rights to self-aware artificially intelligent agents. Naturally, in their case, having the right to own property would also come with the corresponding tax obligations.

[86] Ironically, while Bill Gates' statements were widely broadcasted, just the day before of their publishing the European Parliament had rejected a relatively unnoticed proposal to impose a "robot tax" on owners. For more information see European Parliament, 'Robots And Artificial Intelligence: MEPs Call For EU-Wide Liability Rules' (2017) <http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules> accessed 24 May 2017.

[87] Bill Gates, Interview with Kevin J. Delaney, 'Why Bill Gates Would Tax Robots' (2017).

[that funding] can come on the profits that are generated by the labour-saving efficiency there. Some of it can come directly in some type of robot tax'[88]. As Bill Gates' statements indicate, *robot taxation* may be conceived under different models, according to the tax base. It can be applied over ownership, over the reduction in costs generated by replacing human labour with machines[89] or as point-of-sale tax[90], designed to tax the acquisition of automation technologies. In any case, the common ground is the same: mitigate the loss of fiscal revenue for governments. The idea of *robot taxation* drawn large support but also severe criticism. The main critics revolve around the facts that making automated labour more expensive may stifle innovation, increase the complexity of tax systems and decrease the relative competitiveness of tax frameworks: companies, especially those with more resources, are attracted by the most advantageous legal frameworks.

Some of the critics argue that, alternatively, by exempting human labour from wage taxes the tax benefits of automation would be neutralized and companies would not need to resort to automation technologies. This would be achieved by granting offsetting tax preferences for firms that employ human workers for each category of tax benefit. 'To begin with wage taxation, the tax preference could entail a repeal of the employer contributions to the social security and Medicare systems'[91]. The problem with this approach is that it 'would accelerate the insolvency of the social security system'[92]. In fact, any attempt to neutralize the benefits of automation by making human labour cheaper from a tax perspective will also be contributing to the loss of governmental revenue. Furthermore, it fails to consider that, soon enough, the price of such technologies may be lower than human wages and, therefore, companies would still feel tempted to automate labour. The probable outcome is the aggravation of the issues with the sustainability of social security.

It results that the neutralization of the tax benefits of automation may bring undesired side effects, regardless of consisting in an increase of taxes on automation or in a decrease of taxes on human labour. Because of this, some authors advocate that, in alternative, the solution to mitigate the loss

[88] Bill Gates, Interview with Kevin J. Delaney, 'Why Bill Gates Would Tax Robots' (2017).
[89] Bill Gates, Interview with Kevin J. Delaney, 'Why Bill Gates Would Tax Robots' (2017).
[90] Yanis Varoufakis, 'A Tax On Robots?' <https://www.project-syndicate.org/commentary/bill-gates-tax-on-robots-by-yanis-varoufakis-2017-02> accessed 25 May 2017.
[91] Ryan Abbott and Bret N. Bogenschneider, 'Should Robots Pay Taxes? Tax Policy In The Age Of Automation' [2017] Harvard Law & Policy Review, Forthcoming <https://ssrn.com/abstract=2932483> accessed 23 May 2017.
[92] Ryan Abbott and Bret N. Bogenschneider, 'Should Robots Pay Taxes? Tax Policy In The Age Of Automation' [2017] Harvard Law & Policy Review, Forthcoming <https://ssrn.com/abstract=2932483> accessed 23 May 2017.

of fiscal revenue resides in tackling its source, by preventing the displacement of human workers. For that purpose, it has been suggested that governments should subsidize human wages for low-income workers[93], offer employability guarantees, employing the jobless as the employer of last resort, or establish human job quotas or minimum people quotas for certain jobs[94]. The first two of these proposals seem to merely makeup the problem by swapping the reduction in tax collection for extra governmental spending. The third, however, could prove to be an adequate response. Yet, it comes at the expense of rendering a country's legal framework relatively less competitive.

A latter category of authors suggest that rather than attempting to rebalance social security systems, such systems should be rethought as a whole. The most debated idea concerns the introduction of a universal basic income, which consists in 'an income paid by a political community to all its members on an individual basis, without means test or work requirement'[95], 'either as a floor to provide a basic level of subsistence, as a complement to existing welfare state policies, or in some cases as a replacement for the welfare state'[96]. The advantage of a universal basic income over current unemployment insurance programs is that it is less vulnerable to moral hazard from within the public administration and easier and cheaper to manage. On the other hand, it is less prepared to respond to employment shocks[97] and would, in theory, be more expensive since it encompasses the entirety of the population[98]. A more focused approach in the same line of the universal basic income would be the introduction of a negative income tax[99] based upon an income threshold which, when not reached, entitles the individual to receive a certain amount of money instead of having its income taxed. These ideas, however, suffer from the limitation of not being viable *per*

---

[93] Noah Smith, 'What's Wrong With Bill Gates' Robot Tax' *Bloomberg* (2017) <https://www.bloomberg.com/view/articles/2017-02-28/what-s-wrong-with-bill-gates-robot-tax> accessed 25 May 2017.

[94] World Economic Forum, 'Six Ways To Protect Jobs From Robot Automation' <https://www.facebook.com/worldeconomicforum/videos/10154432426296479> accessed 25 May 2017.

[95] Philippe Van Parijs, 'Basic Income: A Simple And Powerful Idea For The Twenty-First Century' (2004) 32 Politics & Society.

[96] Richard McGahey, 'Universal Basic Income And The Welfare State' (2016) <https://ssrn.com/abstract=2863954> accessed 23 May 2017

[97] Alice Fabre, Stéphane Pallage and Christian Zimmermann, 'Universal Basic Income Versus Unemployment Insurance' [2014] CESifo Working Paper Series No. 5106 <https://ssrn.com/abstract=2540055> accessed 23 May 2017.

[98] Richard McGahey, 'Universal Basic Income And The Welfare State' (2016) <https://ssrn.com/abstract=2863954> accessed 23 May 2017

[99] World Economic Forum, 'Six Ways To Protect Jobs From Robot Automation' <https://www.facebook.com/worldeconomicforum/videos/10154432426296479> accessed 25 May 2017.

se in the long run. In fact, they still imply extra governmental spending which may not be available in a scenario where governments are losing tax revenues because of automation technologies.

Other less orthodox alternatives to reform social security systems, but without such limitation, include shared economy models[100] and a universal basic dividend. 'Imagine that a fixed portion of new equity issues (IPOs) goes into a public trust that, in turn, generates an income stream from which a UBD is paid. Effectively, society becomes a shareholder in every corporation, and the dividends are distributed evenly to all citizens. To the extent that automation improves productivity and corporate profitability, the whole of society would begin to share the benefits. No new tax, no complications in the tax code, and no effect on the existing funding of the welfare state'[101].

Between the proposals that we have enumerated – *robot taxation*, offsetting tax preferences for human workers, human wage subsidies, governmental employability guarantees, minimum human quotas, a universal basic income, a negative income tax, shared economy models and the universal basic dividend – and others that we have not, it certainly seems plausible to manage the impact of artificial intelligence over labour markets and social security systems in a sustainable manner. However, since our economic systems were designed and evolved in a time when humans were in the centre of production and commerce, a conception that automation technologies and artificial intelligence are increasingly challenging, it might be necessary to think "outside the box" to ensure their long-term sustainability. In any case, the implementation of these or any other proposals will necessarily require in depth analysis and projections, which by far exceeds the scope of our work.

4.3. Chapter Summary

The eventual use of public services or infrastructures by an artificially intelligent agent does not translate into a benefit for the agent, but for the user or designer who instructed him to take the action that implied the use of such service or infrastructure. In fact, since artificially intelligent agents are designed to directly or indirectly contribute to the welfare of humans, a human will

[100] An interesting exercise can be found in Ida Auken, 'Welcome To 2030. I Own Nothing, Have No Privacy, And Life Has Never Been Better', *Annual Meeting of the Global Future Councils* (World Economic Forum 2016) <https://www.weforum.org/agenda/2016/11/shopping-i-can-t-really-remember-what-that-is> accessed 26 May 2017.
[101] Yanis Varoufakis, 'A Tax On Robots?' <https://www.project-syndicate.org/commentary/bill-gates-tax-on-robots-by-yanis-varoufakis-2017-02> accessed 25 May 2017.

always be the ultimate beneficiary of the public services or infrastructures that the agent uses while carrying out its purpose. Hence, it does not seem correct to say that it would be *fair* for artificially intelligent agents to be taxed because they benefit from public investment. Taxes, however, may also be justified by *necessity*. This is the case of taxes that aim to modify patterns of consumption or employment within the economy, by making some classes of transaction more or less attractive.

Artificial intelligence has an unprecedented potential to disrupt the labour markets, as machines will be able to replace workers in a variety of cognitive and creative tasks and in tasks that employ manual labour but could not have been automated so far due to technologic constraints (such as driving). Even if, so far, markets have balanced themselves by moving a slice of labour towards more cognitive-oriented tasks, the fact that artificial intelligence will be able to replace jobs in virtually every tier of the pyramid is generating concerns that jobs will be eliminated faster than new ones can be created. Furthermore, even in the event that artificial intelligence results in net job creation, it is unlikely that current methods of workforce retraining are able to accompany its pace. Some authors even claim that machine learning may empower artificially intelligent agents to take on the new jobs created as a consequence of their own development. Under any of these scenarios, such events will directly result in loss of revenue for governments due to a reduction in tax collections since capital income is taxed at much lower rates than labour income. In addition to this, the replacement of human labour by automated labour may translate in major growths of social security expenses since social security systems are designed to provide unemployment insurance to workers who lose their jobs. These increased expenses, combined with the loss of fiscal revenue, are generating concerns as to the sustainability of current social security systems.

Several proposals aim to control the impact of artificial intelligence over labour markets and social security systems: *robot taxation*, offsetting tax preferences for human workers, human wage subsidies, governmental employability guarantees, minimum human quotas, a universal basic income, a negative income tax, shared economy models and the universal basic dividend. While most of these ideas may be highly unorthodox, our economic systems were designed and evolved in a time when humans were in the centre of production and commerce, a conception that automation technologies and, more specifically, artificial intelligence are increasingly challenging. Hence, it might be necessary to think "outside the box" to ensure their long-term sustainability.

In any case, we demonstrated that there are alternatives to the direct taxation of artificially intelligent agents that do not imply a fiction that artificially intelligent agents are property owners and pay taxes, rendering it unnecessary. Naturally, the implementation of these or any other alternatives will necessarily require in depth analysis and projections, which by far exceeds the scope of our work. When Bill Gates referred to *robot taxation*, he noted that 'how you'd do it (...) it's interesting for people to start talking about now'[102]. In this Chapter, we aimed to do just that.

5. Ensuring Control

5.1. The Problem of Control

Until this point, we have argued in favour of a separate legal status for artificially intelligent agents and focused on drafting that status, namely regarding liability for damages they cause, eventual attribution of rights and potential taxation duties. Defining these matters allows for minimum certainty concerning the consequences of the introduction of new intelligent agents in society, which contrasts with the large amount of unforeseeability that it comports. However, our work is not complete without addressing the risks of that unforeseeability and how such risks may be mitigated. As mankind attempts to create machines that are more and more autonomous, it might be difficult for humans to ensure that those machines do not become *too* autonomous.

The problem of control is not new in literature and entertainment. For long, science fiction has portrayed artificial intelligence through the Frankenstein complex, a term coined by Isaac Asimov to designate the classical situation where an artificial intelligent being turns on its creator.[103] As technology advances and artificial intelligence becomes more and more developed, many high-profile individuals go even further and claim that artificial intelligence poses an *existential* risk for humanity. This thesis has notably been endorsed by some of the world's leading experts on

[102] Bill Gates, Interview with Kevin J. Delaney, 'Why Bill Gates Would Tax Robots' (2017).

[103] Some notorious examples would be *The Matrix* trilogy, where humans are rendered irrelevant, being used solely to fuel machines, and the *Terminator* franchise, where Skynet wipes most of mankind from the face of Earth.

science and technology, such as Elon Musk, Bill Gates, Stephen Hawking and Stuart J. Russel[104][105]. Yet, 'whereas ultimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species'[106], the problem of control branches much further than this hypothesis and has much more immediate implications.

Loss of control may occur by several means: malfunctions, security breaches, the superior response time of computers compared to humans' or conscious or unconscious flawed programming[107]. Malfunctions and security breaches may be particularly dangerous when articulated with the superior response time of computers. A malfunction or security breach in an automated weapons system may lead that system to unleash an attack against a target, whose own automated weapons system may immediately respond to, leading to an instantaneous and automatic escalation of conflict. However, it is unconscious flawed programming that raises the most complex issues.

In their paper *Concrete Problems in AI Safety*[108], Amodei and others describe several possible failure scenarios due to flawed programming, namely related to a fragile distributional shifting, unsafe exploration, unscalable oversight, negative side effects or reward hacking. A fragile distributional shifting refers to the incapacity of an artificially intelligent agent to adequately adapt to an environment different than its training environment. As an example, a self-driving car must adapt its turning speed to the climate conditions that it is facing. Unsafe exploration refers to the need of ensuring that artificially intelligent agents don't make exploratory moves with bad repercussions. This may be the case of a cleaning robot which attempts to clean an electrical outlet with a wet mop. Unscalable oversight is related to the inability of addressing every possible

[104] Kevin Rawlinson, 'Microsoft's Bill Gates Insists AI Is A Threat' *BBC* (2015) <http://www.bbc.com/news/31047780> accessed 18 May 2017.

[105] Stephen Hawking and others, 'Stephen Hawking: 'Transcendence Looks At The Implications Of Artificial Intelligence - But Are We Taking AI Seriously Enough?' *The Independent* (2014) <http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html> accessed 18 May 2017.

[106] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

[107] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

[108] Dario Amodei and others, 'Concrete Problems In AI Safety' (2016) <https://arxiv.org/abs/1606.06565> accessed 18 May 2017.

scenario during training. How can the cleaning robot know, for each possible item he finds, whether it belongs to someone or is garbage? Negative side effects concern negative disturbances of the environment cause by the agent while pursuing its goal. Massive amounts of data are already being managed by artificial intelligence able to communicate with databases without the intervention of humans and, possibly, even without their awareness. Unless data protection rules are embedded in the code of those machines, such rules will be rendered useless. This issue will only grow bigger with the introduction of blockchain technology. Finally, reward hacking is related to seeking alternative unwanted means of achieving the programmed objective. A classic example is set forth by Russel and Norvig: an artificially intelligent agent is designed to minimize human suffering, yet, given how humans always find a reason to suffer, the optimal decision for the agent is to terminate the human race as soon as possible – no humans, no suffering[109].

Furthermore, 'Bostrom (2012) and Omohundro (2008) have argued that sufficiently capable AI systems are likely by default to adopt "convergent instrumental sub goals" such as resource acquisition and self-preservation, unless the objective function explicitly disincentives these strategies. These types of problems are likely to be more severe in systems that are more capable, unless action is taken to prevent them from arising'[110]. Additionally, if the artificially intelligent agent is capable of learning and adapting, it might be difficult to regain the lost control. These characteristics make artificially intelligent agents potential sources of 'public risk on a scale that far exceeds the more familiar forms of public risk that are solely the result of human behaviour'[111].

5.2. Fundamental Principles; the Kill Switch

It becomes clear that losing control over artificially intelligent agents comports risks, regardless of whether the loss of control is local (the agent can no longer be controlled by the human responsible for its operation) or general (the agent can no longer be controlled by any human).

[109] Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach (3rd edn, Prentice Hall 2009).

[110] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[111] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

Yet, aiming to prevent any and every loss of control implies closing off every single possible vulnerability. That may, unfortunately, be an unrealistic task. However, it certainly seems possible to limit the impact of such events by exerting a motivational and a capability control[112].

Exerting a motivational control over artificially intelligent agents consists in imbuing human-friendly goals in their code with which no action or function can conflict. The rationale behind this theory is that a loss of control does not necessarily pose a significant risk if the goals of the artificially intelligent agent 'align with those of the public'[113]. Since people will naturally expect artificially intelligent agents to follow community's rules as they are introduced in society, these "fundamental principles" could include extensive social and moral standards. Yet, such a large extent would raise issues in defining a set of principles adequate for every community[114] and would vastly increase the possibilities of 'perverse instantiation'[115] or reward hacking, that we referred to in the last Section as "seeking alternative unwanted means of achieving the objective".

In alternative, motivational control could be achieved through a minimalistic set of fundamental principles. In this respect, it is common to refer to Isaac Asimov's famous Three Laws of Robotics[116] (and the later added fourth or zeroth law[117]): (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws; and (0) a robot may not harm humanity, or, by inaction, allow humanity to come to harm.

---

[112] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (1st edn, Oxford University Press 2014).

[113] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

[114] In response to this issue, the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems recommends as a best practice to clearly delineate the community in which such agent is to be deployed, 'to prioritize the values that reflect the shared set of values of the larger stakeholder groups' of that community and to use principles such as *Common Good* 'to resolve differences in the priority order of different stakeholder groups'. For more information, refer to The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[115] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (1st edn, Oxford University Press 2014).

[116] Isaac Asimov, 'Runaround' [1942] *Astounding Science Fiction*.

[117] Isaac Asimov, *Robots And Empire* (1st edn, Doubleday Books 1985).

Since their creation in 1942, Asimov's Laws have faced several critics, namely regarding cases where the robot should hurt a person for its own good (e.g.: performing surgery) or for the general good of the population (e.g.: arresting a criminal that is resisting arrest)[118], and several alternatives have been proposed. Some of the most promising ones adopt indirect processes of determination of what human-friendly goals entail. A prominent example is Eliezer Yudkowsky's coherent extrapolated volition, where the artificially intelligent agent's meta-goal is to 'achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard'[119].

In any case, no matter whether motivational control is to be exercised through an extensive or minimalistic set of fundamental principles, this hypothesis currently faces two major obstacles. One concerns the difficulty of ensuring that a complex, upgradeable and possibly even self-modifying artificial intelligence will retain its goals throughout its upgrades[120]. The other one, and most immediate, is that experts do not know how to reliably program abstract values into a machine.

However, this does not imply that no action may be taken until the time where experts manage to program Asimov's Laws or any other set of fundamental principles into a machine's code. Some options to consider are to direct Asimov's Laws 'at the designers, producers and operators of robots'[121], to include ethical reflexion and technical augmentation in the curriculum of students learning artificial intelligence, computer science or data science 'at University level and for all advanced degrees'[122] and to complement attempts of motivational control with capability control.

Capability control consists in preventing artificially intelligent agents from being *capable* of causing harm even if they *want* to. The rationale behind this hypothesis is that, as every human can be killed or disabled, every built machine should be possible to kill or disable. This implies

---

[118] Gabriel Hallevy, 'The Criminal Liability Of Artificial Intelligence Entities' (2010) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096> accessed 29 April 2017

[119] Jürgen Schmidhuber, 'Complex Value Systems In Friendly AI', *Artificial General Intelligence* (1st edn, Springer-Verlag Berlin 2011).

[120] Benja Fallenstein and Nate Soares, 'Problems Of Self-Reference In Self-Improving Space-Time Embedded Intelligence' (Springer International Publishing 2014) <https://intelligence.org/files/ProblemsSelfReference.pdf> accessed 18 May 2017.

[121] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).

[122] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

integrating reversibility mechanisms or opt-out mechanisms in artificially intelligent agents. Reversibility models tell the agent which actions are reversible and how to reverse them, allowing users to undo undesired actions and get back to a safe stage[123]. Opt-out mechanisms, on the other hand, designate kill switches that allow to abruptly switch off artificially intelligent agents.

The issue with capability control is that, after the deployment of artificially intelligent agents, the ones that have the capacity to learn will continue to grow smarter and, inevitably, a given agent will realize that allowing itself to be shut down or its actions to be reversed will interfere with the pursuit of its goals. When artificial intelligence gets to a point where agents are on a level playing field with their designers, an agent's continued learning might enable it to outwit its designers and escape the capability control mechanisms, namely, by hacking other systems to install and run backup copies of itself, by creating other artificially intelligent agents without control mechanisms, by pre-emptively disabling anyone who might want to activate such mechanisms or by persuading designers into believing that they do not need to do so, among others. Indeed, after such agents are created and deployed, their superior strategic planning abilities would be more successful at finding ways to dominate humans, than those of humans to dominate the agent *post facto*.[124]

Potential solutions to the capacity control issue involve creating agents that are able to learn to become indifferent to whether their control mechanisms get activated or not ("safely interruptible agents"[125]) or programming artificially intelligent agents to compensate themselves for any lost utility caused by an interruption in such a way that they end up being indifferent to whether they are interrupted or not ("utility balancing"[126]). Yet, both the "safely interruptible agents" approach and the "utility balancing" approach have the limitation of either conditioning the artificially intelligent agent's action to avoiding the activation of the control mechanisms or rendering the agent completely indifferent to whether they are activated are not, consequently, acting as if they do not exist or will not be activated. In the first case, the agent's reward system will be affected, altering the way it was designed to function. In the latter, since the agent is unmotivated to care

[123] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016).
[124] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (1st edn, Oxford University Press 2014).
[125] Laurent Orseau and Stuart Armstrong, 'Safely Interruptible Agents', *32nd Conference on Uncertainty in Artificial Intelligence* (2016) <https://intelligence.org/files/Interruptibility.pdf> accessed 23 May 2017.
[126] Nate Soares and others, 'Corrigibility', *AAAI Conference on Artificial Intelligence* (AAAI Publications 2015) <https://intelligence.org/files/Corrigibility.pdf> accessed 19 May 2017.

about whether the control mechanisms remain functional or not, it may innocently disable them (e.g.: removing an unnecessary component), deploy other agents (in the cases where it is capable of doing such) without embedded control mechanisms or fail to arrange graceful shutdowns (e.g.: not getting caught in the middle of a task during a shutting down for scheduled maintenance).

The optimal scenario would be, naturally, if an artificially intelligent agent was able to detect when it is doing wrong and stop itself, hence, self-employing motivational and capability control mechanisms. Luckily for humanity, while 'it is sane to be concerned (…) currently, the state of our knowledge doesn't require us to be worried'[127] that artificially intelligent agents go rogue. In any case, it is paramount that improvements in motivational control and capability control mechanisms advance *pari passu* with the advancements in artificial intelligence technology itself.

5.3. Accountability and Transparency

It results from the previous Section that designers of artificially intelligent agents should pursue the goal of ensuring that artificially intelligent agents are reversible and amenable to shut down and modification. Yet, until the moment when artificially intelligent agents are able to detect they are doing wrong and self-apply control mechanisms, the application of such mechanisms is dependent of human intervention, which might not always be possible and immediate. Hence, it is imperative that designers 'cultivate a "safety mind-set" in the conduct of research in order to identify and pre-empt unintended and unanticipated behaviours'[128] and adopt other precautions to minimize the risks of losses of control and unforeseeable undesired action, such as 'work[ing] to ensure that AI systems fail gracefully in the face of adversarial inputs, out-of- distribution errors, unexpected rapid capability gain, and other large context changes [and] work[ing] to build safe and secure environments in which potentially unsafe AI systems can be developed and tested'[129].

---

[127] Laurent Orseau, Interview with BBC, 'Google Developing Kill Switch For AI' (2016).

[128] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[129] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

In this regard, artificially intelligent agents can be tested "in-the-wild" (by presenting it with situations) or subjected to black-box testing (by presenting with synthetic inputs). Black-box testing enables behaviour observations and testing in scenarios that might not occur naturally, allowing for a larger number of experiments than "in-the-wild" testing[130]. However, a major challenge in artificial intelligence safety is building agents that can safely transition from the black-box to the outside "open world", where unpredictable things may happen. Adapting gracefully to unforeseen situations is difficult yet necessary for safe operation[131]. When designing technology that could impact the safety or wellbeing of humans, it is not enough to just presume that it works.

Indeed, besides extensive testing, designers must closely observe the behaviour of artificially intelligent agents in the outside world, namely, their autonomous learning processes. Designers must monitor feedback from customers and must react immediately to reports of harmful conduct. If a newly introduced agent has isolated incidents of malfunctioning, the designer must examine possible causes. And, if such incidents cannot be explained by improper handling or third party interference, the agent must be reprogrammed or taken off the market. If the designer would fail to do so, he should be held publicly liable for negligence or non-fulfilment of his duties of care[132].

Unfortunately, 'at least for the foreseeable future, AI developers will likely be unable to build systems that are guaranteed to operate exactly as intended or hoped for in every possible circumstance'[133]. Hence, it is important that artificially intelligent agents are imbued with mechanisms that allow for the maintenance of a clear line of accountability. These include identity tags[134] that enable tracing the agent back to its designers and some sort of diagnose-enabling system that helps diagnose the cause of the malfunctions of the artificially intelligent agent.

---

[130] Anupam Datta, Shavak Sen and Yair Zick, 'Algorithmic Transparency Via Quantitative Input Influence: Theory And Experiments With Learning Systems', *37th IEEE Symposium on Security and Privacy* (Institute of Electrical and Electronics Engineers 2016).

[131] National Science and Technology Council of the Executive Office of the President of the United States of America, 'Preparing For The Future Of Artificial Intelligence' (2016).

[132] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is To Blame? Self-Driving Cars And Criminal Liability' (2016) 19 New Criminal Law Review: An International and Interdisciplinary Journal.

[133] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[134] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous

In this respect, an interesting debate currently going on in literature concerns the need (or not) for transparency to ensure accountability. Transparency can be grossly defined as the possibility to discover how and why an artificially intelligent agent made a particular decision or acted the way it did[135]. It may be provided through several mechanisms. 'For instance (1) for users of care or domestic robots a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took, (2) for validation or certification agencies the algorithms underlying the AI/AS and how they have been verified, (3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box'[136]. Yet, within the scope of the current debate, transparency is generally linked to source-code disclosure. The advocates of transparency consider such disclosure to be important not only to investigate accidents and gather evidence but also to build public confidence in technology by exposing it to scrutiny. In any case, the main points are that designers ensure that decision-making steps may be reconstructed and that artificially intelligent agents are able, when asked, to show the process which led to their actions, to identify any sources of uncertainty and to state any assumptions they relied upon.

Transparency may be achieved indirectly, through tax incentives or favourable tort standards that in some way limit the liability of designers, or directly, through legal requirement[137]. In this regard, a relevant regulation is the European Union's new General Data Protection Regulation (GDPR), adopted on April 2016 and scheduled to take effect in 2018, which states that individuals have a right to "an explanation of the decision reached after such assessment and to challenge the decision". This Regulation, however, is facing hard criticism for 'not seem[ing] to understand that it is often not practical or even possible, to explain all decisions made by algorithms (…) [since] often, the challenge of explaining an algorithmic decision comes not from the complexity of the

Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[135] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[136] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017.

[137] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

algorithm, but the difficulty of giving meaning to the data it draws on'[138]. As an example of this, 'the machine-learning tool SearchInk can predict fairly accurately whether a written name is male or female (…) based on the pen strokes [but] the software's creators do not know why'[139].

Against such criticism, it is argued that if an artificially intelligent agent's reasoning is too complex to be understood by humans, then it should be made less effective. Predictably, this argument is far from consensual, especially among people linked to the industries at stake and people who are sceptical concerning artificial intelligence (un)safety. Furthermore, several other arguments are presented against the need for transparency: opacity may be needed to prevent tax cheats or terrorists from gaming the system; transparency may be undesirable because it defeats the legitimate protection of consumer data, commercial proprietary information or trade secrets; achieving transparency may imply the disclosure of personal data from users;[140] processing the overflow of disclosed data requires more algorithmic computation, creating a vicious circle[141].

A proposed alternative to transparency is to monitor behaviour[142]. A way to do so would be through "black box tinkering". 'Black box tinkering is a reverse engineering technique: a "process of articulating the specifications of a system through a rigorous examination drawing on domain knowledge, observation, and deduction to unearth a model of how that system works"'[143]. The rationale behind this technique is that by confronting the agent with different scenarios, it can be tested for credibility, fairness and trustworthiness. Examiners would look for evidence of bias and

[138] Nick Wallace, 'EU's Right To Explanation: A Harmful Restriction On Artificial Intelligence' <http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#> accessed 20 May 2017.

[139] Nick Wallace, 'EU's Right To Explanation: A Harmful Restriction On Artificial Intelligence' <http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#> accessed 20 May 2017.

[140] Joshua A. Kroll and others, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review <http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3> accessed 20 May 2017.

[141] Maayan Perel and Niva Elkin-Koren, 'Black Box Tinkering: Beyond Transparency In Algorithmic Enforcement' (Florida Law Review, Forthcoming 2016) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741513> accessed 20 May 2017.

[142] Nick Wallace, 'EU's Right To Explanation: A Harmful Restriction On Artificial Intelligence' <http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#> accessed 20 May 2017.

[143] Maayan Perel and Niva Elkin-Koren, 'Black Box Tinkering: Beyond Transparency In Algorithmic Enforcement' (Florida Law Review, Forthcoming 2016) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741513> accessed 20 May 2017.

make adjustments as necessary. We have doubts, however, concerning the extension of the test-scenarios that would have to be experimented in order for solid conclusions to be drawn.

A more moderate alternative would be to require artificially intelligent agents to create cryptographic commitments as digital evidence of their procedural regularity. A cryptographic commitment grossly consists in a promise that a given agent is bound to specific values. Agents can be designed to publish commitments describing what they will do before they are deployed and commitments describing what they actually did afterwards. Zero-knowledge proofs can then be used to ensure that the commitments actually correspond to the actions taken by an agent. Zero-knowledge proofs are cryptographic tools that allow to prove that a particular decision has a certain property without having to reveal how such property is known or what the decision policy actually is. The advantage of this solution over source-code transparency is that 'by disclosing commitments instead of source code or inputs and outputs, system operators are able to fully explain what their systems do without actually disclosing how those systems work up front'[144].

In our view, if it is demonstrated that the disclosure of the source code is not necessary to prove the credibility, fairness and trustworthiness of a decision-making process, we are not averse to solutions that maintain opacity over source-code transparency. Either way, while both of the presented alternatives allow to test the decision-making processes of artificially intelligent agents for the presence (or not) of specific elements, neither of them offer *per se* a solution for cases where what is at stake is the review of the decision policies *themselves*. Therefore, it is advisable that, as a complement, courts and regulatory bodies are granted the faculty to demand full source-code transparency from designers if they deem it necessary, possibly under a protective regime.

In this Section, we started by addressing potential sources of designer liability, namely, insufficiencies in the safety testing or in the monitoring of deployed artificially intelligent agents, and proceeded to describe the necessary elements to ensure that designers may be held accountable: identity tagging and transparency (or other diagnose-enabling systems). Unlike in the previous Chapter, here we are not dealing with direct liability for damages caused by artificially intelligent agents, but with liability based in negligence or in the violation of a duty of care. Therefore, while

[144] Joshua A. Kroll and others, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review <http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3> accessed 20 May 2017.

it is not required that any actual damages have been caused, it is important to define who may take action against negligent designers or designers who violate the respective duties of care.

We believe that the more efficient solution would pass through the creation of specialized regulatory bodies, since, in most cases, high levels of technical expertise would be required to adequately analyse such kind of situations. Due to the fact that artificial intelligence has the potential to impact the safety and wellbeing of humans, said regulatory bodies need to be able to take pre-emptive action, rather than merely acting reactively or following a complaint. Besides the application of fines, regulatory bodies must be able to ensure damage prevention, for that, being advisable that they are granted powers to suspend the sale or to limit the scope of use of artificially intelligent agents whose designers don't comply with safety requirements. Other measures, such as the need to require pre-approval for commercialization or the creation of a rating system that evaluates both safety and foreseeability are also worthy of consideration.

5.4. Chapter Summary

The problem of control concerns the risks that arise as a consequence of losing control over artificially intelligent agents. These risks are not only related to damages, but also to the protection of personal data and public safety. Losses of control may occur due to malfunctions, security breaches, the superior response time of computers compared to humans' or conscious or unconscious flawed programming, namely, regarding a fragile distributional shifting, unsafe exploration, unscalable oversight, negative side effects or the possibility of reward hacking. The loss of control may be local, when the agent can no longer be controlled by the human responsible for its operation, or general, when the agent can no longer be controlled by any human.

While it may be unrealistic to attempt to prevent any and every loss of control, it is possible to limit the impact of such event through motivational control and capability control mechanisms. Exerting motivational control consists in imbuing human-friendly goals in an artificially intelligent agent's code with which no action or function can conflict. These human-friendly goals can either be extensive, including social standards, or minimalistic, such as Isaac Asimov's Three Laws of Robotics or Eliezer Yudkowsky's coherent extrapolated volition. However, experts currently do not know how to reliably program abstract values. Capability control consists in preventing

artificially intelligent agents from being *capable* of causing harm even if they *want* to. This implies integrating reversibility mechanisms or opt-out mechanisms in artificially intelligent agents. The issue with capability control is that artificially intelligent agents that have the capacity to learn will continue to grow smarter and will realize that allowing themselves to be shut down or their actions to be reversed will interfere with the pursuit of the goals which they were programmed to achieve, potentially being able to escape those mechanisms due to their superior capabilities.

In any case, until the moment when artificially intelligent agents are able to detect they are doing wrong and self-apply control mechanisms, their application is dependent on human intervention, which might not always be possible and immediate. Hence, it is imperative that designers perform extensive testing, in order to identify and pre-empt unintended and unanticipated behaviours, and monitor customer feedback, reacting immediately to reports of harmful conduct. We consider that failure to perform adequate testing, to monitor customer feedback or to apply the (reasonably) latest available motivational and capability control mechanisms shall give rise to liability.

<table>
<tr><th colspan="3">Direct Liability for Damages</th><th colspan="3">Negligence or Violation of a Duty of Care</th></tr>
<tr><td>Reactive Machines</td><td>Machines w/ Limited Memory</td><td>Machines w/ Theory of Mind</td><td rowspan="4">Failure to perform adequate safety testing</td><td rowspan="4">Failure to monitor costumer feedback and to take action in response to it</td><td rowspan="4">Failure to apply the (reasonably) latest available motivational and capability control techniques</td></tr>
<tr><td rowspan="3">Programming</td><td>Programming</td><td>Programming</td></tr>
<tr><td>Autonomous action w/ coding deficiency</td><td>Autonomous action w/ coding deficiency</td></tr>
<tr><td>Autonomous action w/o coding deficiency but no insurance</td><td>Autonomous action w/o coding deficiency but no insurance</td></tr>
</table>

Figure 5 - Proposed Designer Liability

Unlike in the previous chapter, here we are not dealing with direct liability for damages caused by artificially intelligent agents, but with liability based in negligence or in the violation of a duty of care. Hence, while it is not required that any actual damages have been caused, it is important to define who may take action against negligent designers or designers who violate their duties of

care. We believe that the most efficient solution would pass through creating specialized regulatory bodies that have the power to take pre-emptive action and apply fines, suspend or pre-approve the commercialization of artificially intelligent agents, limit their scope of use, create a safety and foreseeability rating system or any other measure necessary to protect safety and wellbeing.

In order to ensure that designers may be held accountable it is important that artificially intelligent agents are embedded with mechanisms that allow for the maintenance of a clear line of accountability, namely, identity tagging and some sort of diagnose-enabling system that ensures that decision-making steps may be reconstructed and that artificially intelligent agents are able to show the process which led to their actions, to identify any sources of uncertainty and to state any assumptions they relied upon. In this respect, an interesting debate concerns the need (or not) for source-code transparency. The advocates of transparency consider such disclosure to be important not only to investigate accidents and gather evidence but also to build public confidence in technology by exposing it to scrutiny. Against the need for transparency, several arguments are presented: opacity may be needed to prevent tax cheats or terrorists from gaming the system; transparency may be undesirable because it defeats the legitimate protection of consumer data, commercial proprietary information or trade secrets; transparency may imply the disclosure of personal data from users; processing the overflow of disclosed data requires more algorithmic computation, creating a vicious circle. A proposed alternative to transparency is to monitor behaviour and look for evidence of bias, however, we have doubts, concerning the extent of the test-scenarios that would have to be experimented in order to draw solid conclusions. A more moderate alternative would be to require artificially intelligent agents to create cryptographic commitments as digital evidence of their procedural regularity. The advantage of this solution over source-code transparency is that designers can fully explain what agents do without actually disclosing how those agents work. In our view, if it is demonstrated that the disclosure of the source code is not necessary to prove the credibility, fairness and trustworthiness of a decision-making process, we are not averse to solutions that maintain opacity over source-code transparency. Either way, it is advisable that, as a complement to such solutions, courts and regulatory bodies have the power to demand full source-code transparency in order to review decision policies *themselves*, rather than merely checking for the presence (or not) of specific elements in them.

When designing technology that could impact the safety or wellbeing of humans, it is not enough to simply presume that it works. We agree that if designers cannot achieve justified confidence that an agent is safe and controllable, so that deploying it does not create an unacceptable risk of negative consequences, then the agent cannot and should not be deployed. However, 'a person does not need the resources and facilities of a large corporation to write computer code. Anyone with a reasonably modern personal computer (or even a smartphone) and an Internet connection can now contribute to AI-related projects. Individuals thus can participate in AI development from a garage, a dorm room, or the lobby of a train station. This potential for discreetness provides the most jarring difference between AI and earlier sources of public risk'[145]. While the solutions that we here propose may help minimize the risk of losses of control and the impact of such events, further work is necessary in order to figure out how to deal with the problematic of discreetness.

6. Final Remarks

6.1. Conclusions

The introduction of artificial intelligence in industry and society will revolutionize the current social construction and comport several technologic, industrial and regulatory challenges, which legal frameworks are not prepared to give a direct response to. As artificially intelligent agents become more and more autonomous over time, the less they can be considered mere tools.

In order to accommodate this reality, we understand that machines with limited memory, machines with a theory of mind and self-aware machines should be considered separate legal entities from their owners and users. Doing so would be convenient from a legal point of view and, in the case of self-aware machines, it would also find support on moral considerations. Yet, when it comes to reactive machines, no positive argument justifies a similar consideration. Being considered separate legal entities does not automatically entitle such machines to the acknowledgement of rights. There is no strong reason to acknowledge rights to reactive machines, machines with limited memory and machines with a theory of mind since their perception of eventual rights or

[145] Matthew U. Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology.

wills they have or are entitled to have is merely a result of being programmed to such. As a consequence, owners of these machines would also own their respective creations. However, if humanity is ever able to build self-aware machines, these machines appear to be entitled to the legal acknowledgement of rights, whether we are talking about the right to own property or some sort of fundamental rights, similar to the ones of humans, but with the due adaptations.

Designers and users should be directly liable for actions that artificially intelligent agents are programmed or commanded to take (respectively). Regarding autonomous actions of the agents, one of two scenarios is possible: if the action was enabled by a coding deficiency, product defect rules apply and the designer would be held liable for negligence; if the action results purely from the evolving conduct of the agent, the designer could exempt himself by subscribing insurance on behalf of the agent, otherwise being directly liable. By ensuring that the agent has means to compensate third parties for the damages it causes, the designer passes his liability on to it. This insurance scheme may be complemented by a compensation fund destined to ensure that damages can be compensated for in cases where no insurance exists. This fund would too rely on contributions from designers, yet, in order to not be overburdened with costs, they may opt to incorporate these costs in the selling price of the agents. Self-aware machines, on the other hand, may be held directly liable for their actions, since their ability to subjectively perceive the concepts of justice and punishment would enable the application of standard rules of liability. The fact that self-aware machines have a conscience, are sentient and able to have subjective experiences places them much closer to humans than to machines and, perhaps, closer to humans than any other reality.

The eventual use of public services or infrastructures by an artificially intelligent agent does not translate into a benefit for the agent, but for the owner or user who instructed him to take the action that implied the use of that service or infrastructure. In fact, since artificially intelligent agents are designed to directly or indirectly contribute to the wellbeing of humans, a human will always be the ultimate beneficiary of the public services or infrastructures that the agent uses while carrying out its purpose. Therefore, it does not seem correct to say that it would be *fair* for artificially intelligent agents to be taxed because they benefit from public investments. Taxes, however, may also be justified by *necessity*. This is the case of taxes that aim to modify patterns of consumption or employment within the economy, by making some classes of transaction more or less attractive.

In fact, artificial intelligence has an unprecedented potential to disrupt the labour markets, as machines will be able to replace workers in a variety of cognitive and creative tasks and in tasks that employ manual labour but could not have been automated so far due to technologic constraints (such as driving). The fact that artificial intelligence will be able to replace jobs in virtually every tier of the pyramid is generating concerns that jobs will be eliminated faster than new ones can be created. Furthermore, even in the event that artificial intelligence results in net job creation, it is unlikely that current methods of workforce retraining are able to accompany its pace. Such events will directly result in loss of revenue for governments due to a reduction in tax collections since capital income is taxed at much lower rates than labour income. In addition to this, the replacement of human labour by automated labour may also translate in major growths of social security expenses since social security systems are designed to provide unemployment insurance to workers who lose their jobs. These increased expenses combined with the loss of fiscal revenue generates serious concerns regarding the sustainability of current social security systems. However, we demonstrated that this problem may be tackled without the need for the direct taxation of artificially intelligent agents, as there are other alternatives that do not imply considering that artificially intelligent agents are property owners and pay taxes, rendering this fiction unnecessary.

	Separate Legal Status	**Rights**		**Direct Liability**	**Direct Taxation**
		Property Rights	Fundamental Rigths		
Reactive Machines	✗	✗	✗	✗	✗
Machines w/ Limited Memory	✓	✗	✗	✓ *	✗
Machines w/ Theory of Mind	✓	✗	✗	✓ *	✗
Self-Aware Machines	✓	✓	✓	✓	✓
* If insurance was subscribed on the agent's behalf					

Figure 6 - The Legal Status of Artificially Intelligent Robots

Attributing a separate legal status to artificially intelligent agents and defining the contents of that status, namely, regarding liability, eventual rights and potential taxation duties, allows for minimum certainty as to the consequences of the introduction of those new intelligent agents in

society, which contrasts with the large amount of unforeseeability that it comports. However, the risks of that unforeseeability still need to be addressed and mitigated as they are not only related to eventual damages, but also to the protection of personal data and public safety itself.

As mankind attempts to create machines that are more and more autonomous, it might be difficult for humans to ensure that such machines do not become *too* autonomous. Losses of control may occur due to malfunctions, security breaches, the superior response time of computers compared to humans' or conscious or unconscious flawed programming, namely, regarding a fragile distributional shifting, unsafe exploration, unscalable oversight, negative side effects or the possibility of reward hacking. The loss of control may be local, when the agent can no longer be controlled by the human responsible for its operation, or general, when the agent can no longer be controlled by any human. While it may be unrealistic to attempt to prevent any and every loss of control, it is possible to limit the impact of such event through motivational control and capability control mechanisms. In any case, until the moment when artificially intelligent agents are able to self-apply control mechanisms, their application is dependent on human intervention, which might not always be possible and immediate. Hence, it is of utmost importance that designers perform extensive testing, in order to identify and pre-empt unintended and unanticipated behaviours, and monitor feedback from customers, reacting immediately to reports of harmful conduct.

We understand that failure to perform adequate testing, monitor customer feedback or apply the (reasonably) latest available motivational and capability control mechanisms should give rise to designer liability based in negligence or in the violation of a duty of care. Since it is not required that any actual damages have been caused, it is important to define who may take action against negligent designers or designers who violate their duties of care. The solution could pass by specialized regulatory bodies with powers to take pre-emptive action and apply fines, suspend or pre-approve the sale of artificially intelligent agents, limit their scope of use, create a safety and foreseeability rating system and other necessary measures to protect safety and wellbeing.

In order to ensure designer accountability it is also important that artificially intelligent agents are embedded with mechanisms that allow for the maintenance of a clear line of accountability, namely, identity tagging and some sort of diagnose-enabling system that ensures that decision-making steps may be reconstructed and that artificially intelligent agents are able to show the process which led to their actions, to identify any sources of uncertainty and to state any

assumptions they relied upon. In this respect, an interesting debate concerns the need (or not) for source-code transparency. A proposed alternative to transparency is to require artificially intelligent agents to create cryptographic commitments as digital evidence of their procedural regularity. The advantage of this solution over source-code transparency is that designers could fully explain what agents do without actually disclosing how those agents work. In our view, if it is demonstrated that the disclosure of the source code is not necessary to prove the credibility, fairness and trustworthiness of a decision-making process, we are not averse to solutions that maintain its opacity. Either way, it is advisable that courts and regulatory bodies have the power to demand full source-code transparency in order to be able to review the decision policies *themselves*, rather than merely checking for the presence (or not) of specific elements in them.

When designing technology that could impact the safety or wellbeing of humans, it is not enough to simply presume that it works. We believe that if designers cannot achieve justified confidence that an agent is safe and controllable, so that deploying it does not create an unacceptable risk of negative consequences, then the agent cannot and should not be deployed. Nevertheless, we also believe that artificial intelligence has the potential to place mankind on the path to prosperity and ultimately free Men from the burden of labour, giving us the opportunity to focus on tasks where creativity and passion play bigger roles. As Stephen Hawking once put it, with current and near-future technology 'everyone can enjoy a life of luxurious leisure if the machine-produced wealth is shared, or most people can end up miserably poor if the machine-owners successfully lobby against wealth redistribution'[146]. Therefore, it is our understanding that all of the solutions that we have here prescribed, or any others that are recommended or adopted, shall, at every moment, be susceptible to adjustment in order to strike a balance between guaranteeing the wellbeing of our species and the freedom to innovate. Artificial intelligence is not something to be afraid of, but rather to embrace. And, by pro-actively discussing the challenges this technology may comport, we are a few steps closer to prevent any potential downside while still fully reaping its benefits.

6.2. Future Work

[146] Stephen Hawking, 'Science AMA Series: Stephen Hawking AMA Answers' <https://www.reddit.com/r/science/comments/3nyn5i/science_ama_series_stephen_hawking_ama_answers> accessed 27 May 2017.

With this work, we expect to have provided lawmakers, policymakers and other interested parties with materials to start the discussion of the challenges that artificial intelligence may comport and how to address them. However, we strongly encourage other authors to reflect upon these matters and dive deeper into them, either by perfecting our ideas or coming up with new ones, more adequate to the pursuit of our common goal: ensuring prosperity and the wellbeing of humanity.

In particular, we understand that further work is necessary in the following areas: providing a concrete and functional definition of *artificial intelligence* from a legal perspective; developing effective control, identification and individualization mechanisms for artificially intelligent agents, namely by advancing the debate on transparency; guaranteeing that such mechanisms are implemented in every single artificially intelligent agent to be deployed, even in cases where such agents are anonymously and discreetly designed; mitigating the impact or artificial intelligence over labour markets and social security systems, while ensuring that the benefits of this technology are shared with the entirety of the population; and creating "artificial intelligence ecosystems"[147] attractive to both designers and investors that encourages innovation and adequate safety testing.

List of References

[Abbott and Bogenschneider, 2017] Abbott R and Bogenschneider B, 'Should Robots Pay Taxes? Tax Policy In The Age Of Automation' [2017] Harvard Law & Policy Review, Forthcoming <https://ssrn.com/abstract=2932483> accessed 23 May 2017

[Ainge, 2017] Ainge Roy E, 'New Zealand River Granted Same Legal Rights As Human Being' The Guardian (2017) <https://www.theguardian.com/world/2017/mar/16/new-zealand-river-granted-same-legal-rights-as-human-being> accessed 22 April 2017

[Allen and Widdison, 1996] Allen T and Widdison R, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law & Technology

[147] For further information on creating the right kind of artificial intelligence ecosystems, see Erik Vermeulen, 'How To Prepare For Automation? Or, Why We Need More "Artificial Intelligence Ecosystems" Now!' <https://hackernoon.com/how-to-prepare-for-automation-or-why-we-need-more-artificial-intelligence-ecosystems-now-4a4a767e733b> accessed 10 April 2017

[Allgrove, 2004] Allgrove B, 'Legal Personality For Artificial Intellects: Pragmatic Solution Or Science Fiction?' (Master, University of Oxford 2004)

[Amodei and others, 2017] Amodei D and others, 'Concrete Problems In AI Safety' (2016) <https://arxiv.org/abs/1606.06565> accessed 18 May 2017

[Asimov, 1985] Asimov I, *Robots And Empire* (1st edn, Doubleday Books 1985)

[Asimov, 1942] Asimov I, 'Runaround' [1942] *Astounding Science Fiction*

[Auken, 2016] Auken I, 'Welcome To 2030. I Own Nothing, Have No Privacy, And Life Has Never Been Better', *Annual Meeting of the Global Future Councils* (World Economic Forum 2016) <https://www.weforum.org/agenda/2016/11/shopping-i-can-t-really-remember-what-that-is> accessed 26 May 2017

[Bogenschneider, 2017] Bogenschneider B, 'The Effective Tax Rate Of U.S. Persons By Income Level' (2017) 145 Tax Notes

[Bombay High Court, 1925] *Pramatha Nath Mullick v. Pradyumna Kumar Mullick* [1925] Bombay High Court, 27 BOMLR 1064 (Bombay High Court)

[Bostrom, 2014] Bostrom N, *Superintelligence: Paths, Dangers, Strategies* (1st edn, Oxford University Press 2014)

[Caytas, 2017] Caytas J, 'European Perspectives On An Emergent Law Of Robotics' [2017] Columbia Journal of European Law <https://ssrn.com/abstract=2956958> accessed 23 May 2017

[Datta, Sen and Zick, 2016] Datta A, Sen S and Zick Y, 'Algorithmic Transparency Via Quantitative Input Influence: Theory And Experiments With Learning Systems', *37th IEEE Symposium on Security and Privacy* (Institute of Electrical and Electronics Engineers 2016)

[Deloitte, 2017] Deloitte, 'Agiletown: The Relentless March Of Technology And London's Response' (2017) <http://www.deloitte.com/content/dam/Deloitte/uk/Documents/uk-futures/london-futures-agiletown.pdf> accessed 23 May 2017.

[Dewey, 1926] Dewey J, 'The Historic Background Of Corporate Legal Personality' (1926) 35 The Yale Law Journal

[EP, 2017] European Parliament, 'Robots And Artificial Intelligence: MEPs Call For EU-Wide Liability Rules' (2017) <http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules> accessed 24 May 2017

[EP's Committee on Legal Affairs, 2016] European Parliament's Committee on Legal Affairs, 'Draft Report With Recommendations To The Commission On Civil Law Rules On Robotics' (European Parliament's Committee on Legal Affairs 2016)

[Executive Office of the POTUS, 2016] Executive Office of the President of the United States of America, 'Artificial Intelligence, Automation, And The Economy' (2016)

[Fabre, Pallage and Zimmermann, 2014] Fabre A, Pallage S and Zimmermann C, 'Universal Basic Income Versus Unemployment Insurance' [2014] CESifo Working Paper Series No. 5106 <https://ssrn.com/abstract=2540055> accessed 23 May 2017

[Fallenstein and Soares, 2014] Fallenstein B and Soares N, 'Problems Of Self-Reference In Self-Improving Space-Time Embedded Intelligence' (Springer International Publishing 2014) <https://intelligence.org/files/ProblemsSelfReference.pdf> accessed 18 May 2017

[Feigenbaum, 2003] Feigenbaum E, 'Some Challenges And Grand Challenges For Computational Intelligence' [2003] Journal of the ACM

[Fischer, 1997] Fischer J, 'Computers As Agents: A Proposed Approach To Revised U.C.C. Article 2' (1997) 72 Indiana Law Journal <http://www.repository.law.indiana.edu/cgi/viewcontent.cgi?article=1850&context=ilj> accessed 23 April 2017

[Frey and Osborne, 2017] Frey C and Osborne M, 'The Future Of Employment: How Susceptible Are Jobs To Computerisation?' (2017) 114 Technological Forecasting and Social Change

[Gates, 2017] Bill Gates, Interview with Kevin J. Delaney, 'Why Bill Gates Would Tax Robots' (2017)

[Ginzburg, 1987] Ginzburg R, 'Perth Amboy Church Is 302 And Counting' The New York Times (1987) <http://www.nytimes.com/1987/02/15/nyregion/perth-amboy-church-is-302-and-counting.html> accessed 22 April 2017

[Gless, Silverman and Weigend, 2016] Gless S, Silverman E and Weigend T, 'If Robots Cause Harm, Who Is To Blame? Self-Driving Cars And Criminal Liability' (2016) 19 New Criminal Law Review: An International and Interdisciplinary Journal

[Hallevy, 2010] Hallevy G, 'The Criminal Liability Of Artificial Intelligence Entities' (2010) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096> accessed 29 April 2017

[Hawking, 2015] Hawking S, 'Science AMA Series: Stephen Hawking AMA Answers' <https://www.reddit.com/r/science/comments/3nyn5i/science_ama_series_stephen_hawking_ama_answers> accessed 27 May 2017

[Hawking and others, 2014] Hawking S and others, 'Stephen Hawking: 'Transcendence Looks At The Implications Of Artificial Intelligence - But Are We Taking AI Seriously Enough?' The Independent (2014) <http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html> accessed 18 May 2017

[High Court of Australia, 1958] *International Harvester Company of Australia Proprietary Limited v. Carrigan's Hazeldene Pastoral Company* [1958] High Court of Australia, HCA 16; 100 CLR 644 (High Court of Australia)

[Hintz, 2016] Hintz A, 'Understanding The Four Types Of AI, From Reactive Robots To Self-Aware Beings' [2016] The Conversation UK <https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616> accessed 10 April 2017

[Horton, 2015] Horton R, 'The Robots Are Coming' (Deloitte LLP 2015) <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/finance/deloitte-uk-finance-robots-are-coming.pdf> accessed 6 April 2017

[Human Rights Watch, 2013] Human Rights Watch, 'World Report 2013: Saudi Arabia' (2013) <https://www.hrw.org/world-report/2013/country-chapters/saudi-arabia> accessed 22 April 2017

[The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 2016] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 'Ethically Aligned Design: A Vision For Prioritizing Human

Wellbeing With Artificial Intelligence And Autonomous Systems' (Institute of Electrical and Electronics Engineers 2016) <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf> accessed 18 May 2017

[Jowitt, 2016] Jowitt J, 'Monkey See, Monkey Sue? Gewirth's Principle Of Generic Consistency And Rights For Non-Human Agents' (2016) 19 Trinity College Law Review

[Kroll and others, 2017] Kroll J and others, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review <http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3> accessed 20 May 2017

[Lee, 2016] Lee D, 'Google Developing Kill Switch For AI' *BBC* (2016) <http://www.bbc.com/news/technology-36472140> accessed 19 May 2017

[Luger, 2004] Luger G, Artificial Intelligence: Structures And Strategies For Complex Problem Solving (5th edn, Addison-Wesley 2004)

[Ma, Nahal and Tran, 2015] Ma B, Nahal S and Tran F, 'Robot Revolution – Global Robot & AI Primer' (Bank of America Merril Lynch 2015) <https://www.bofaml.com/content/dam/boamlimages/documents/PDFs/robotics_and_ai_condensed_primer.pdf> accessed 23 May 2017

[Manyika and others, 2017] Manyika J and others, 'Harnessing Automation For A Future That Works' (McKinsey Global Institute 2017) <http://www.mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works> accessed 23 May 2017

[McGahey, 2016] McGahey R, 'Universal Basic Income And The Welfare State' (2016) <https://ssrn.com/abstract=2863954> accessed 23 May 2017

[Musk, 2016] Elon Musk, Interview with Sam Altman, 'Y Combinator's How To Build The Future Series' (2016)

[National Science and Technology Council of the Executive Office of the POTUS, 2016] National Science and Technology Council of the Executive Office of the President of the United States of America, 'Preparing For The Future Of Artificial Intelligence' (2016)

[Nilsson, 1998] Nilsson N, Artificial Intelligence: A New Synthesis (1st edn, Morgan Kaufmann Publishers 1998)

[Orseau and Armstrong, 2016] Orseau L and Armstrong S, 'Safely Interruptible Agents', *32nd Conference on Uncertainty in Artificial Intelligence* (2016) <https://intelligence.org/files/Interruptibility.pdf> accessed 23 May 2017

[Oxford Dictionaries, 2017] 'Definition Of Robot In English' (Oxford Dictionaries, 2017) <https://en.oxforddictionaries.com/definition/robot> accessed 6 April 2017

[Perel and Elkin-Koren, 2016] Perel M and Elkin-Koren N, 'Black Box Tinkering: Beyond Transparency In Algorithmic Enforcement' (Florida Law Review, Forthcoming 2016) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741513> accessed 20 May 2017

[Petit, 2017] Petit N, 'Law And Regulation Of Artificial Intelligence And Robots: Conceptual Framework And Normative Implications' (2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2931339> accessed 24 April 2017

[Poole, Mackworth and Goebel, 1998] Poole D, Mackworth A and Goebel R, Computational Intelligence: A Logical Approach (1st edn, Oxford University Press 1998)

[Premack and Woodruff, 1978] Premack D and Woodruff G, 'Does The Chimpanzee Have A Theory Of Mind?' (1978) 1 Behavioral and Brain Sciences

[Rawlinson, 2015] Rawlinson K, 'Microsoft's Bill Gates Insists AI Is A Threat' BBC (2015) <http://www.bbc.com/news/31047780> accessed 18 May 2017

[Rosen, Nilsson and others, 1967] Rosen C, Nilsson N, Raphael B and others, 'Shakey' [1967] LIFE <http://cyberneticzoo.com/cyberneticanimals/1967-shakey-charles-rosen-nils-nilsson-bertram-raphael-et-al-american> accessed 22 April 2017

[Russel and Norvig, 2009] Russell S and Norvig P, Artificial Intelligence: A Modern Approach (3rd edn, Prentice Hall 2009)

[Scherer, 2016] Scherer M, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29 Harvard Journal of Law & Technology

[Schmidhuber, 2011] Schmidhuber J, 'Complex Value Systems In Friendly AI', *Artificial General Intelligence* (1st edn, Springer-Verlag Berlin 2011)

[Schwab and Saams, 2016] Schwab K and Saams R, 'Preface' (World Economic Forum 2016) <http://reports.weforum.org/future-of-jobs-2016/preface> accessed 23 May 2017

[Science and Technology Committee of the House of Commons of the UK, 2016] Science and Technology Committee of the House of Commons of the United Kingdom, 'Robotics And Artificial Intelligence' (Science and Technology Committee of the House of Commons of the United Kingdom 2016)

[Smith, 2011] Smith H, 'Can Monkey Who Took Grinning Self-Portrait Claim Copyright?' Metro (2011) <http://metro.co.uk/2011/07/14/can-monkey-who-took-grinning-self-portrait-claim-copyright-77773> accessed 30 April 2017

[Smith, 2017] Smith N, 'What's Wrong With Bill Gates' Robot Tax' *Bloomberg* (2017) <https://www.bloomberg.com/view/articles/2017-02-28/what-s-wrong-with-bill-gates-robot-tax> accessed 25 May 2017

[Soares and others, 2015] Soares N and others, 'Corrigibility', *AAAI Conference on Artificial Intelligence* (AAAI Publications 2015) <https://intelligence.org/files/Corrigibility.pdf> accessed 19 May 2017

[Solum, 1992] Solum L, 'Legal Personhood For Artificial Intelligences' (1992) 70 North Carolina Law Review <http://papers.ssrn.com/abstract=1108671> accessed 23 April 2017

[Supreme Court of India, 1969] *Yogendra Nath Naskar v. Commissioner Of Income Tax* [1969] Supreme Court of India, 1969 AIR 1089, 1969 SCR (3) 742 (Supreme Court of India)

[Supreme Court of the US, 1967] *Loving v. Virginia* [1967] Supreme Court of the United States, 388 US 1 (Supreme Court of the United States)

[Turing, 1950] Turing A, 'Computing Machinery And Intelligence' [1950] Mind

[US Copyright Office, 2015] United States Copyright Office, 'Compendium Of U.S. Copyright Office Practices: Chapter 300' (2015)

[US District Court for the Northern District of California, 2015] Naruto, *et al. v. David Slater, et al.* [2015] United States District Court for the Northern District of California, No 3:2015cv04324 (United States District Court for the Northern District of California)

[Van Parijs, 2004] Van Parijs P, 'Basic Income: A Simple And Powerful Idea For The Twenty-First Century' (2004) 32 Politics & Society

[Varoufakis, 2017] Varoufakis Y, 'A Tax On Robots?' <https://www.project-syndicate.org/commentary/bill-gates-tax-on-robots-by-yanis-varoufakis-2017-02> accessed 25 May 2017

[Vermeulen, 2017] Vermeulen E, 'How To Prepare For Automation? Or, Why We Need More "Artificial Intelligence Ecosystems" Now!' <https://hackernoon.com/how-to-prepare-for-automation-or-why-we-need-more-artificial-intelligence-ecosystems-now-4a4a767e733b> accessed 10 April 2017

[Wallace, 2017] Wallace N, 'EU's Right To Explanation: A Harmful Restriction On Artificial Intelligence' <http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm#> accessed 20 May 2017

[Waters and Bradshaw, 2016] Waters R and Bradshaw T, 'Rise Of The Robots Is Sparking An Investment Boom' Financial Times (2016) <https://www.ft.com/content/5a352264-0e26-11e6-ad80-67655613c2d6> accessed 5 April 2017

[Wells, 2016] Wells G, 'Google's Computers Paint Like Van Gogh, And The Art Sells For Thousands' The Wall Street Journal (2016) <https://blogs.wsj.com/digits/2016/02/29/googles-computers-paint-like-van-gogh-and-the-art-sells-for-thousands> accessed 30 April 2017

[WEF, 2017] World Economic Forum, 'Six Ways To Protect Jobs From Robot Automation' <https://www.facebook.com/worldeconomicforum/videos/10154432426296479> accessed 25 May 2017

[Zhang and others, 2016] Zhang J and others, 'IDC Futurescape: Worldwide Robotics 2017 Predictions' (International Data Corporation 2016)

[Zimmermann, 2015] Zimmerman E, 'Machine Minds: Frontiers Of Legal Personhood' (2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563965> accessed 22 April 2017

[Zuckerberg, 2016] Zuckerberg M, 'Building Jarvis' (Facebook, 2016) <https://www.facebook.com/notes/mark-zuckerberg/building-jarvis/10154361492931634> accessed 10 April 2017