ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE SETTINGS:

Where should be the limits of Artificial Intelligence in legal decision-making? Should an AI device make a decision about human justice?

Degree in Criminology

Academic Year 2019-2020

By Olatz Cibrian Egido

Directed by Iraide Zipitria Leanizbarrutia

ABSTRACT

The application of Artificial Intelligence (AI) systems for high-stakes decision making is currently out for debate. In the Criminal Justice System, it can provide great benefits as well as aggravate systematic biases and introduce unprecedented ones. Hence, should artificial devices be involved in the decision-making process? And if the answer is affirmative, where should be the limits of that involvement? To answer these questions, this dissertation examines two popular risk assessment tools currently in use in the United States, LS and COMPAS, to discuss the differences between a traditional and an actuarial instrument that rely on computerized algorithms. Further analysis of the later is done in relation with the Fairness, Accountability, Transparency and Ethics (FATE) perspective to be implemented in any technology involving AI. Although the future of AI is uncertain, the ignorance with respect to so many aspects of this kind of innovative methods demand further research on how to make the best use of the several opportunities that it brings.

INDEX

1. INTRODUCTION

Criminologists have long tried to forecast crime and predict which criminals represent the biggest danger for society. For many years, the key decisions in the Criminal Justice system such as pretrial release, sentencing or parole, have been taken by humans based on their instincts and personal biases. Characteristics of the defendants over which they have no power such as gender, race or ethnicity have often been used to make such predictions. Over the years, that power of discretion that the judges hold has been seen as inappropriate and has been reduced (Angwin, Larson, Mattu, & Kirchner, 2016). As Martin Luther King said, "*I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character"* (Smith, 2009). This trend altogether with the substantial costs in which the Criminal Justice System of the United States is incurring to confront the mass incarceration has pointed out the forecasting of criminal behaviors as an optimal solution.

In the past few years, an increasing use of Artificial Intelligence (AI) has been experienced in many fields all over the world, including the Criminal Justice System. The buzzword actually evokes a chimera and hence, the attitudes towards AI are extreme. On the one hand, there is the thought of it as a utopian dream, capable of finding the cure for cancer, predict and prevent crimes or cut down the prison overpopulation. On the contrary, it can be perceived as a dystopia, something outrageous that can hold all the data and is controlled by the state in a skewed manner. The problem is that the term conveys things far from reality. It recalls more than what it really is. It is a combination of a euphemism that reduces its signification to bare systems for the automatic processing and analysis of information, and a desideratum of the willingness to emulate the human cognitive processes though a computer system that replicates a brain (Miró Llinares, 2018).

Precisely in the Criminal Justice System, computerized assessment algorithms have been introduced with the aim of developing tools to prevent and reduce crimes, as well as perform tasks such as setting bail conditions or determining criminal sentences, among which risk assessments can be found. Many of them have been used across the country from a long time ago and thus, are not innovative techniques, but is its combination with AI what has made a turn on them (Andrews, Bonta, & Wormith, 2006). The argument behind the growing interest in these intelligent machines remains on the idea that with a sufficient volume of data, it is possible to find patterns and therefore, predict crimes. The

latest developments in data analysis have generated Machine Learning (ML) applications for crime prevention that seemed to be unconceivable years ago. The algorithms are developed by researchers taking into consideration several factors or predictors that might determine an offender's future behavior ranging from demographical factors (e.g. age, sex, race), to historical (e.g. age of first criminal arrest, nature of prior arrests) or social factors (e.g. housing stability, social support) (Fishel, Flack, & DeMatteo, 2018). Yet, whether the AI incorporation to the Criminal Justice has resulted in better or worse outcomes is out for debate.

The main goal of this dissertation is to address the controversial topic of whether AI methods can replace completely functions that human beings have been executing for ages, or if they are nothing else than a support instrument and, in any case, where should be the limits of its performance when decisions about human justice are on stake as they are on the Criminal Justice setting. For that aim, an introductory overview on what is Artificial Intelligence, what is it used for and its primary techniques are explained from scratch in Chapter 2. Chapter 3 continues with an explanation on a recent movement that aims to guarantee and spread good practices within AI, called FATE, acronym for Fairness, Accountability, Transparency and Ethics. Chapter 4 talks about the origins of risk assessments and gives a detailed description of two predominant tools used in the United States, the Level of Service Instruments and COMPAS (Correctional Offender Management Profiles for Alternative Sanctions), having this last one AI incorporated in its system. The following Chapter 5 provides a critical analysis of COMPAS in comparison with LS following the FATE structure for its examination. And finally, Chapter 6 covers the discussion and conclusions reached on the matter and future expectations of AI technology in the Criminal Justice domain.

2. ARTIFICIAL INTELLIGENCE

This Chapter provides a first general approach to what the Artificial Intelligent world consist of. It covers the origins and basic premises underlying this discipline, as well as the different subcategories conforming it and how they operate depending on the objective pursued. All of it from the perspective that this might be an unknown field for the reader.

2.1. What is Artificial intelligence?

Artificial Intelligence (AI) is a broad-ranging branch of computer science aimed to design and build intelligent machines capable of performing tasks that would naturally require human intelligence. Due to the many different research approaches to the field, this or many other definitions of AI would be equally valid, as there is not a universally accepted one.

Throughout the years, definitions have oscillated along four categories or goals to be pursued by AI, outlined in the Table 1 below. The top ones differ from the bottom ones as they address thinking and reasoning processes whilst the others are concerned about behavior. Moreover, left side definitions measure success according to human performance whereas the ones on the right do it with regard to an ideal concept of intelligence —rationality— where a system is considered rational if it "does the right thing". This last approach is not suggesting that humans are irrational but that they are not perfect and make mistakes.

A more recent definition would be that introduced by the U.S. Government companion bills on December 12, 2017 (H.R. 4625 and S. 2217) which establish that AI includes "any artificial systems that perform tasks under varying and unpredictable circumstances, without significant human oversight, or that can learn from their experience and improve their performance. Such systems may be developed in computer software, physical hardware, or other contexts not yet contemplated. They may solve tasks requiring human-like perception, cognition, planning, learning, communication, or physical action. In general, the more human-like the system within the context of its tasks, the more it can be said to use artificial intelligence". However, it is important to note that AI does not only perform tasks that humans are able to handle with their own brain but they can go beyond the capacity of a human brain (Mochon, 2019).

Due to the broad spectrum of definitions, when people hear this buzzword, few actually understand what it really means. The most common mistake is to narrow the term under computer science or mathematics, but AI is a puzzle conformed by pieces from many other domains such as economics, neuroscience, psychology, linguistics, electrical engineering and philosophy (Taulli, 2019). In fact, it is used in a myriad of ways in our workaday without us even noticing. When reading emails, getting driving directions or looking for music or movie recommendations, AI comes into play.

Table 1: Different categories in AI

Systems that think like humans	**Systems that think rationally**
"The automation of activities that we associate with human thinking, activities such as decision making, problem solving, and learning" (Bellman, 1978). "The exciting new effort to make computers think ... *machines with minds*, in the full and literal sense" (Haugeland, 1985).	"The study of computations that make possible to perceive, reason, and act" (Winston, 1992). "The study of mental faculties through the use of computational models" (Charniak & Mcdermott, 1985).
Systems that act like humans	**Systems that act rationally**
"The art of creating machines that perform functions that require intelligence when performed by people" (Kurzweil, 1990). "The study of how to make computers do things at which, at the moment, people are better" (Rich & Knight, 1991).	"The branch of computer science that is concerned with the automation of intelligent behavior" (Luger & Stubblefield, 1993) "A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes" (Schalkoff, 1990)

What all these examples and definitions mentioned before have in common is that they demand some degree of "intelligence", the same kind of intelligence that many human mental activities involve. Thus, Artificial Intelligence could be defined as the study of intelligent behavior. Its aim is to form a theory of intelligence that enlightens the behavior of intrinsically intelligent beings (scientific goal) and mentor the creation of artificial entities qualified to perform intelligent behavior (engineering goal) (Genesereth & Nilsson, 1987).

2.2.History

AI traces its roots back to the 1950s with the introduction of the term by Alan Turing, also known as the father of AI. He established the main goal and vision of AI by raising the question: *can machines think?* And that is, at its core, what AI attempts to answer in an affirmative way (Taulli, 2019).

To answer this question, he came up with a test in his paper "Computing Machinery and Intelligence" (Turing, 1950)—named after him, "The Turing Test"— to determine if a machine is intelligent or not. The test, also known as the imitation game, consists of, basically, a game where there are two players —a human and a computer—and an evaluator that asks open-ended questions to them in order to guess which one is the computer (see Figure 3). The computer will pass the test and therefore, presumed to be intelligent, if the evaluator is not capable to distinguish between them (Taulli, 2019).

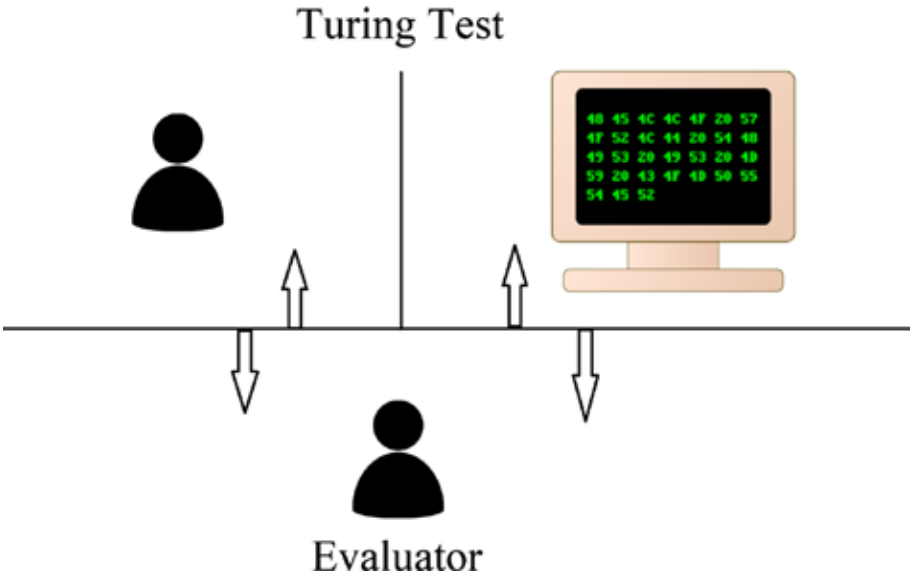


Figure 1. Basic workflow of the Turing Test (Taulli, 2019).

The test is intended to indicate if the machine is capable of processing large amounts of information, understand speech and interact with real people. However, if the computer actually has any kind of knowledge, is self-aware or gives the correct answer, is irrelevant. Based on this, the test was fiercely criticized by John Searle (1980). According to him, a computer would be able to pass the test if it recognized the questions asked (syntax) even without understanding them (semantics).

To prove that, Searle (1980) set up then his own experiment, called "The Chinese Room". An individual in Room 1 pushes a piece of paper through a hole in the wall with some questions in Chinese written on it. In Room 2, there is a non-Chinese speaking individual and lots of manuals with easy-to-use rules for Chinese translation. After some time, the individual in Room 1 receives back the responses in Chinese, and therefore, assumes the other person in Room 2 understands the language. However, what the person in Room 2 is doing is just recognizing the symbols (syntax) but not understanding what they mean (semantics) (Taulli, 2019).

As a consequence, Searle (1980)'s argument is based on the idea that the Chinese Room experiment and, by extension, many other AI systems, appear to solve tasks but they just follow some fixed steps set by sophisticated programs without really understanding what they are doing (Taulli, 2019). His argument was strong and has been a topic of debate ever since.

2.3.Types of AI

AI, as an interdisciplinary science, encompasses a wide variety of subfields that share the common goal of creating intelligent machines through different approaches (Rebala, Ravi, & Churiwala, 2019).

2.3.1. General and Narrow Artificial Intelligence

Following Searle's lines, not only did he reached to the conclusion that the intelligence present in the Chinese Room was merely apparent, but he also believed that, consequently, AI could be divided into two main branches: *Weak AI and Strong AI* (Taulli, 2019).

The former, also known as Narrow AI (NAI), is the attempt to create machines that perform tasks that would seem to require human intelligence, as shown by the Chinese

Room experiment. It is more simplistic than strong AI and specifies in systems designed to perform concrete instructions (e.g. Apple's Siri) (Taulli, 2019). With regard to its use, the functional opportunities are infinite. It can be used to process massive quantities of data, predict patterns, automatize simple tasks or estimate probabilities. These specific or limited NAI tasks are handled, indeed, with certain level of autonomy and even sometimes outperform human efficiency, but only within the range of its specialty. Notwithstanding, the resemblance to a human brain capacity of reasoning is far from being the same.

On the other hand, Strong AI or Artificial General Intelligence (AGI), is aimed to create machines with general intelligence at the human level or beyond (Wang & Goertzel, 2012). In other words, it comprises systems that are autonomous, self-aware and can realize what is going on around them to the extent that some can be emotional or even creative in their interactions like a human would be. These machines would have the capacity to solve, adapt and evolve through facing unlimited number of problems without any human assistance or input. Unfortunately, AGI does not exist in reality yet, but some fictitious examples can be found in the shape of characters like Skynet in Terminator or HAL 9000 in Space Odyssey (Steinfeld, et al., 2006).

2.3.2. Subcategories of AGI and NAI

Apart from that general dual division, AI can be subdivided in four groups based on its functionalities. The first two are part of NAI whereas the third and the fourth categories belong to AGI and do not exist as of yet.

- *Reactive AI*. It is the most basic form of AI. This type of AI is designed to perform concrete tasks with no regard to past experiences. It produces outcomes based on received inputs. For example, IBM's Deep Blue chess player[1] (Taulli, 2019).

[1] Deep Blue was a chess-playing supercomputer designed by IBM which, in 1996, beat grand chess master Garry Kasparov (Taulli, 2019).

- *Limited Memory* AI. Unlike reactive AI, this kind of AI systems can use past experiences for the improvement of future decisions. Apple's Siri belongs to this category (Taulli, 2019).
- *Theory of Mind*. It refers to the skill of attributing mental states (e.g. beliefs, desires, goals, intentions) to others and comprehend that they are different from one's own (Chandrasekaran, Yadav, Chattopadhyay, Prabhu, & Parikh, 2017). This aspect in intimately related to strong AI and as cognitive scientists believe that in order to build human-like machines similar to the ones in the movies, this attribute is vital and should be introduced in their design (Steinfeld, et al., 2006).
- *Self-awareness*. Goes further than Theory of the Mind to achieve human-like consciousness. It would be equal to a complete human being.

Table 2: Breakdown of NAI and AGI

Narrow AI		**General AI**	
Reactive AI	Limited Memory AI	Theory of Mind	Self-Awareness

2.3.3. Artificial Intelligence Techniques

The acronym covers a myriad of technologies or sophisticated mathematical data processing techniques seeking to achieve AI (Buchanan, 2005), among which its most important ones, there are:

1. *Robotics*. Engineering field that is focused on the creation of robots (e.g. Roomba vacuum cleaner);
2. *Planning*. Branch that concerns the task of finding a strategy or sequence of actions to be performed by intelligent entities (e.g. STRIPS[2]) (Taulli, 2019);

[2] STRIPS, which stands for Sandford Research Institute Problem Solver, is an automated planner developed by Fikes and Nilsson in 1971 at SRI International (Fikes & Nilsson, 1971).

3. *Speech*. Development of technologies capable of recognizing speech and translating it to text and vice versa (e.g. Hearsay I[3]);
4. *Vision*. This field focuses on allowing machines to see like a human being would (e.g. self-driving cars) (Shapiro & Stockman, 2001);
5. *Expert Systems*. Interactive computer systems that can solve difficult problems of certain domains at the level of human experts (e.g. Deep Blue) (Taulli, 2019);
6. *Natural Language Processing* (NLP). Field intended to make computer systems understand and operate the language naturally, recognizing the human voice and responding to it with a logical answer (e.g. Siri) (Chowdhury, 2003);
7. *Machine Learning* (ML). Scientific approach whose purpose is to enable machines to learn from a data feedback loop (e.g. traffic predictions when using GPS navigation services) (Michalski, Carbonell, & Mitchell, 2013).

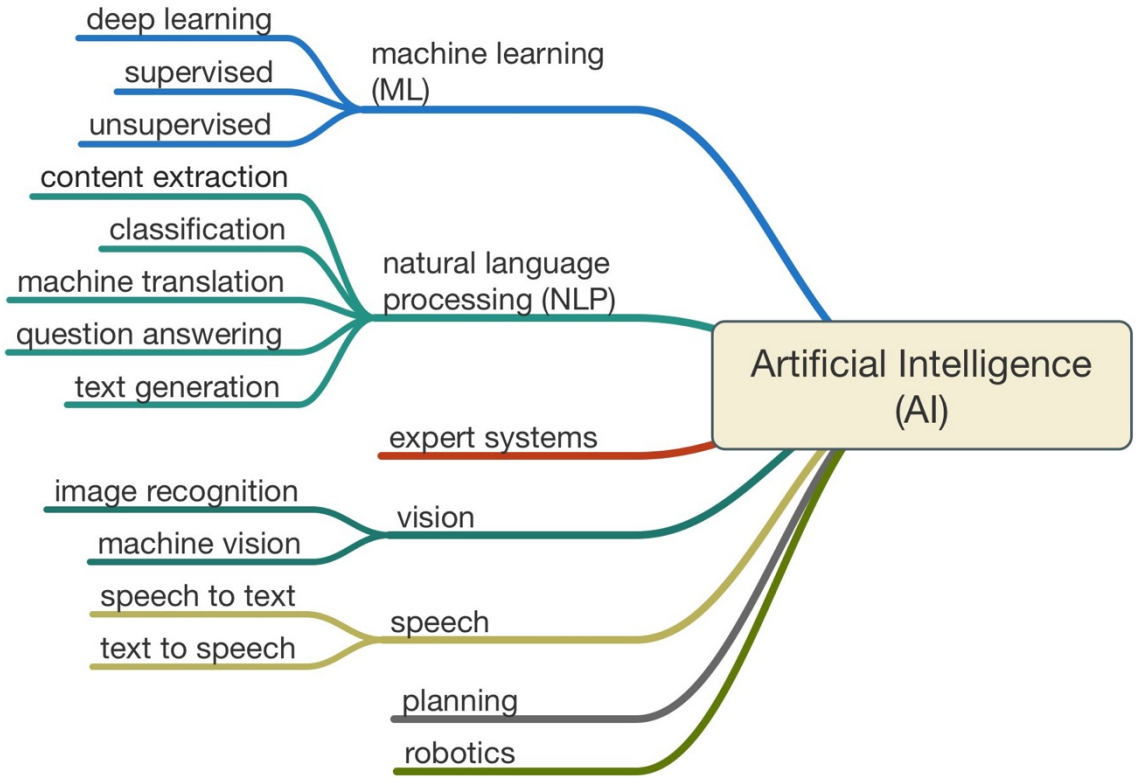


Figure 2. Main fields of AI (El Abid Amrani , Youssfi, & Abra, 2018).

[3] Hearsey I was one of the first speech recognition systems created by Raj Reddy in the late 1960s (Taulli, 2019)

Due to this arduous categorization within AI many terms get confused in this field and sometimes they are used indistinctly and wrongly to refer to AI. Figure 2 shows a diagram with a possible classification of these terms but must be noted that many of them are entangled (Figure 3) and are not necessarily at the same level as it appears. Also, all of them have their own subcategories due to the development of new technologies that have impelled the evolution of these techniques such as Deep Learning, a subfield of ML.

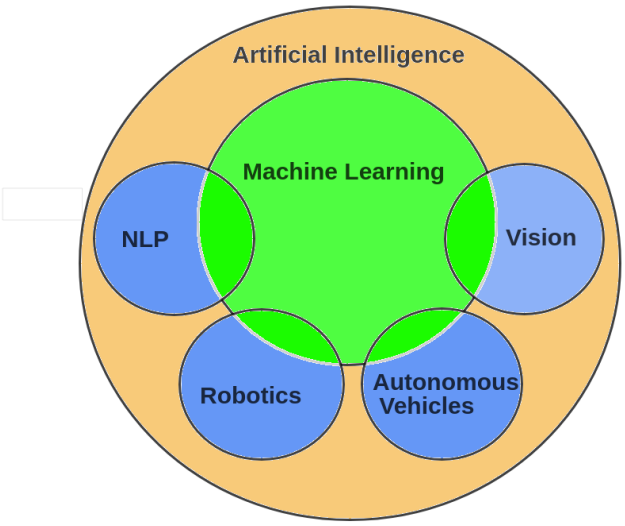


Figure 3. Intersection of ML with other AI fields (Kumar, 2018).

2.4.Machine Learning

Machine Learning (ML), which traces its roots back to the 1980s, is the most prevalent field and plays a major role within AI. This narrow AI subfield is about one approach towards the ultimate goal of AI: creating machines capable of learning how to perform tasks on their own from the data provided to them.

This inter-disciplinary field studies algorithms and techniques for automating solutions to complex problems that are hard to program using conventional programing methods (Rebala, Ravi, & Churiwala, 2019). For that aim, statistical algorithms emulate human cognitive tasks by working out their own procedures through the analysis of large training datasets. All in all, the term ML alludes to a system that can learn without having to be explicitly programmed (Taulli, 2019).

While conventional programming methods cannot be applied to many real-world issues, ML approaches can. A conventional programming method would consist of two

phases. To begin with, the *specification* for the program has to be decided, that is, what is the purpose of its creation and then, the design has to be laid out. The design should include a fixed set of procedures or rules to puzzle out the problem (first phase). Thereupon, the final design is remodeled as a program in a computer language for its implementation (second phase) (Rebala, Ravi, & Churiwala, 2019).

However, as mentioned, this course of action cannot always be applied or is merely impractical because even when the specification is clear, creating a detailed design is very complicated. Imagine that the specification is detecting handwritten characters in an image. If a conventional method was to be used, a prior study of the examples of the training dataset would be necessary to understand the characters in each image and then figure out a fixed set of steps for general character detection in any image. Given the infinite handwritten variations in characters, finding those fixed rules to follow can be a nightmare (Rebala, Ravi, & Churiwala, 2019).

2.4.1. Application of ML techniques

ML comes into play when conventional methods fail to succeed. Not only can ML provide insights into structures and patterns within datasets but also create models by learning from them in order to make predictions of outcomes or behaviors. Moreover, the algorithms used in ML can put an end to many challenging problems in a generic way because they are not designed with an explicit detailed design. What they do is learn the detailed design from the data (Rebala, Ravi, & Churiwala, 2019).

To list a couple of the problems ML solve, in broad terms (Rebala, Ravi, & Churiwala, 2019):

1) Classification: classify data into categories (e.g. divide emails into spam or not spam).
2) Predictions: forecast future values based on a model built upon historical data (e.g. predicting if an offender is likely to recidivate)
3) Clustering: take data and group items into clusters according to characteristics they have in common (e.g. customer segmentation).

As abstract as these problems may seem, the reality is that we see examples of it every day (Figure 4). From personalized content that appear in our Social Media based on the

viewing habits, to customer service chatbots or dating Apps like Tinder. All of them use ML algorithms to improve their success by for example, increasing the chances of matching with our perfect couple.

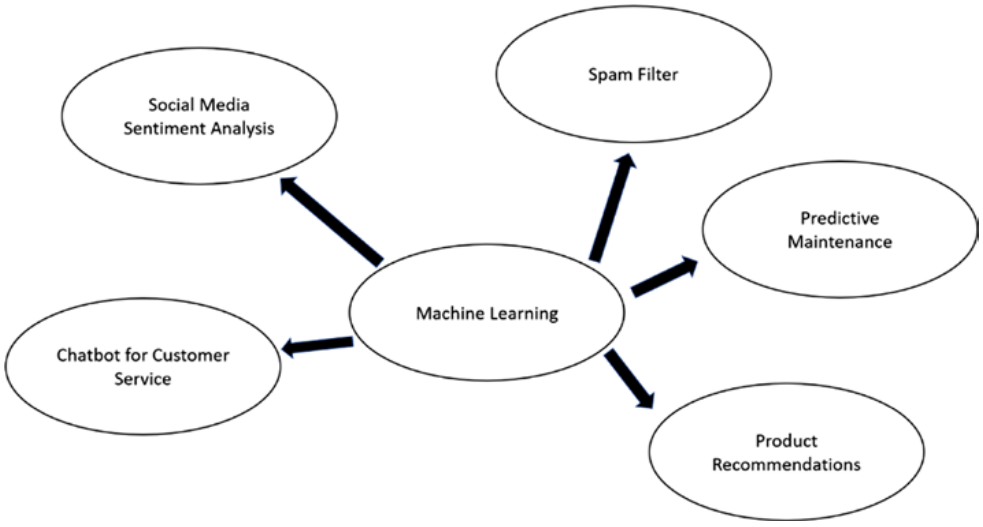


Figure 4. Different uses for Machine Learning (Taulli, 2019).

The benefits of this technology are innumerable, ranging from reducing costs to spotting business opportunities or monitoring risks (Taulli, 2019). And these ML algorithms are more often than not more accurate than humans because their data processing capacity is larger and they do not introduce bias in the model due to prior knowledge (Rebala, Ravi, & Churiwala, 2019).

2.4.2. ML Process

Briefly stating, once the problem to be solved is determined, the first step to be taken is to select which data to feed the algorithm with. Secondly, what type of algorithm has to be decided by guesswork depending on the data available and the problem to be solved. The third step is the training phase, in which the algorithm will use the training data to find patterns and create a model, from which accurate predictions will be produced beyond the training data. Finally, the last step will consist on improving the algorithm by adjusting the values of their parameters (Taulli, 2019).

In this context, an algorithm has to be understood as fixed steps of mathematical instructions provided to the machine. Moreover, a model would be a hypothesis of how the real-world functions that is originated when the machine runs the algorithm.

2.4.2.1.The key role of data

Data is undoubtedly a vital aspect of AI. Without data, algorithms would have no fuel to start with. Data is what allows algorithms to learn find patterns to provide insights and find solutions to these problems. The larger the dataset, the greater the accuracy of the algorithm. The objective is for the ML algorithm to learn the pattern or set of rules from the data in order to create a model capable of making pinpoint predictions over the given dataset. Note also that it is important to randomize the data before it is given to the algorithm or otherwise, the algorithm could detect there is a pattern when there is not and distort the results (Taulli, 2019).

Data can come from many different sources such as Social Media, Cloud systems or Corporate databases and spreadsheets (Taulli, 2019). Depending how is this information organized, data can be divided in three main groups: structured data, semi-structured data and unstructured data:

- Structured data: 20% of the data, labelled, formatted and normally saved in databases or spreadsheets (e.g. phone numbers, Social Security numbers, addresses, etc.).
- Unstructured data: data with no format, which is most of the data (e.g. images, videos, audio files, social network content, etc.). It requires subsequent structuration.
- Semi-structured data: in between structured and unstructured data there is a 5-10% that has some kind of label that helps classification (e.g. JavaScript object Notation or Extensible Markup Language).

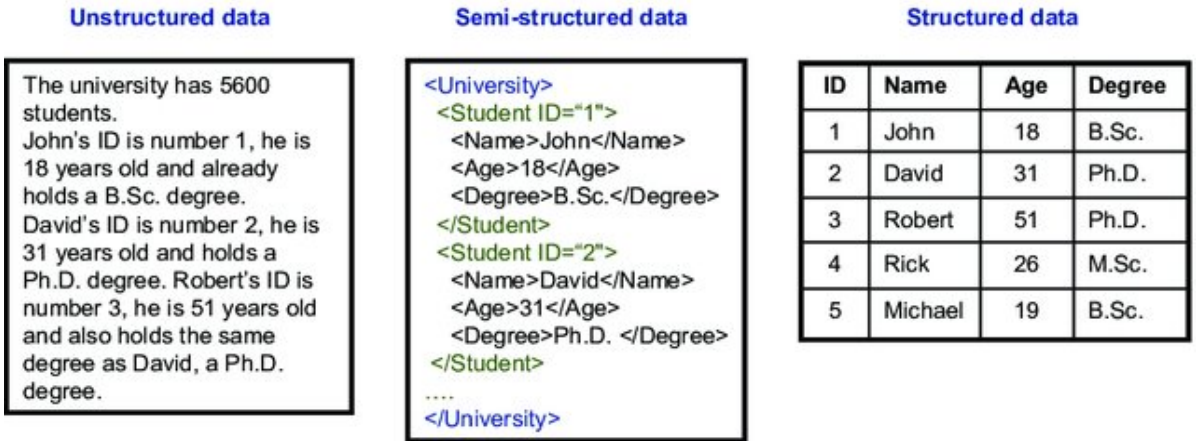


Figure 5. Visual representation of the types of data (Cardoso, 2006).

Regarding how to process all the data, the model uses high-level statistics, mainly probability analysis. Although there is no one unique way to do it, there is one approach, created in the late 1900s, which is widely accepted called the CRIP-DM Process. Figure 6 summarizes the process.

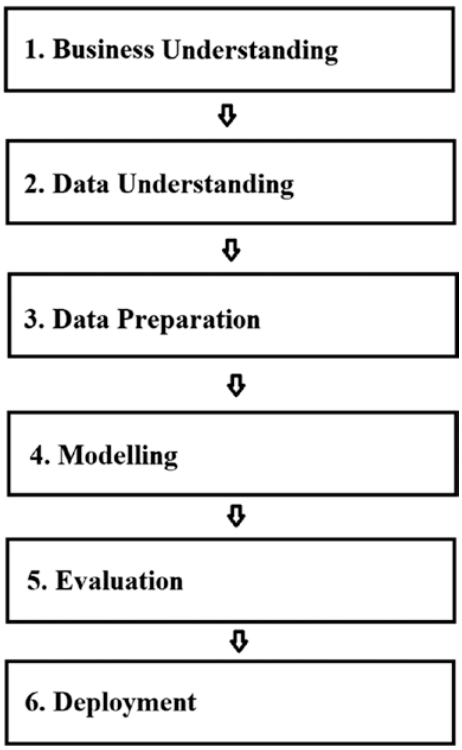


Figure 6. The CRISP-DM Process (Taulli, 2019).

Considering that stages 1-3 comprise up to 80% of the time devoted to data process, only those will be explained to acquire a general knowledge.

1. Business Understanding. In this step, the first thing is to set a goal, a problem looking for answers. Once this is solved, attention should be focused on getting the best outcomes, prejudgments and bias free. For that, the right task force needs to be composed by people with expertise in data science and experts on the specific domain of the AI project. At last, technical needs have to be defined.
2. Data Understanding. Now, you have to select the data source for the project. It can be collected from our own company (In-House Data), from available open sources or from third parties. Regardless of where the data comes from, it has to be reliable. Therefore, you should ask yourself if the data is complete, who manipulated it, if there are any quality problems, etc. This step is less time consuming if the data is already labelled (structured data).

3. Data Preparation. What datasets are you going to use? This is an important decision to make as selecting or excluding just one variable can have negative consequences on the results. Afterwards, you need to get rid of problematic data to improve its quality (is there any duplication? Is it relevant? And consistent?). Anyway, no dataset will be perfect, but neither is our society.

Currently, the volumes of data ready for use are increasing unceasingly due to Internet access, smartphones, wearables, etc., reaching impressive growing rates. To cope with this exponential leap, the technology known as *Big data* was created and it now plays a main role in how to handle this huge valuable asset for AI. There is no consensual definition of what Big Data is, but it is composed by three main features, known as the three Vs: volume (scale of the data, usually unstructured), variety (diversity of data explained before) and velocity (speed at which data is generated) (Laney, 2001). The last one is crucial in today's world as people get frustrated with slow data, but new characteristics have appeared from its evolution such as veracity, value, variability and visualization (Taulli, 2019). One of the techniques used by Big Data is *Data Mining*, which finds patterns and synthesizes huge volumes of data in order to help decision-making (Han, Kamber, & Pei, 2011).

2.4.2.2.Learning Models

While humans learn from past experiences, machines learn from data. Once it has been decided which type of data to work with, there are four ways in which machines can be thought how to do learn from it (See Figure 7): supervised, unsupervised, semi-supervised and reinforcement learning (Taulli, 2019).

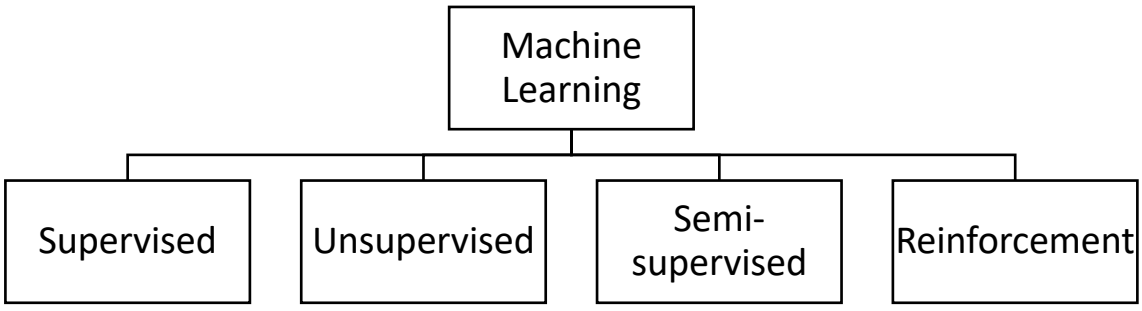


Figure 7. Types of ML methods.

- *Supervised Learning* functions with large amounts of labelled data. Each data point in the dataset is paired to a desired answer (input-output pairs). The algorithm will then learn the characteristics of each data point that leads to the correct output in order to give the right answer to new data points in the future. Usually falls into classification or regression problems (Rebala, Ravi, & Churiwala, 2019).

Keeping up with the example of the detection of handwritten characters, the ML system would require providing a dataset with considerable number of images of handwritten characters which are already labelled or structured (e.g. classified according to the character they contain) that would conform the training data. The compilation of examples will show the algorithm it should work so that afterwards it is able to detect and recognize characters in any type of handwriting, not only from those images contained in the training data (Rebala, Ravi, & Churiwala, 2019).

- *Unsupervised Learning*. Unfortunately, great part of the data available is not labelled so this type of learning in the one to address unlabeled data. The patterns here are discovered using Deep Learning algorithms. This type of learning will become crucial in the future as most of human knowledge is acquired through observation and not by learning labels. It is the usual approach taken for clustering (Taulli, 2019).
- *Semi-supervised Learning* is a combination of the previous ones where there is a majority of labeled data plus a small portion of unlabeled data. To be able to apply the algorithms, the unsupervised data is transformed into supervised using Deep Learning systems (Rebala, Ravi, & Churiwala, 2019).
- *Reinforcement Learning* consists of a trial-and-error process. Outcomes are improved based on positive and negative reinforcement (Taulli, 2019).

2.4.2.3. Common ML algorithms

For each teaching method, there is a huge list of algorithms to assist in the process. The goal is, as stated, to train a model based on one or more algorithms to provide a solution to the problems above.

The algorithms in which the system will rest would differ from the conventional ones because the ML algorithms start by processing data and later on, they learn from it. The ultimate choice of an algorithm will be an educated guess; however, Figure 8 below illustrates the most common types of algorithms used in ML for each type of learning and problem to be addressed (Taulli, 2019).

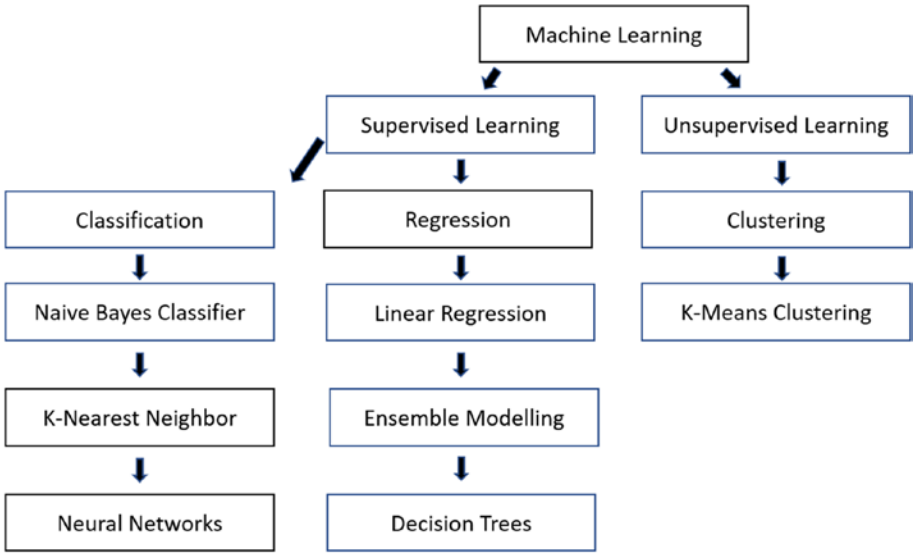


Figure 8. Configuration of ML algorithms (Taulli, 2019).

2.4.2.3.1. Linear Regressions

One of the most common algorithms within the regression kind is the *Linear Regression*, destined to expose the existent relationships between variables. In the case of using a linear regression algorithm ($y = ax + b$), the findings of the algorithm will determine the value of a (slope of the model) and b (y-intercept).

Assuming the quality and quantity of data in which the linear regression algorithm is adequate, it can make predictions of four types: true negatives (1) and true positives (2) denote accurate predictions while false positives (3) imply that the model predicted something to be true when it was not and oppositely, false negatives (4) mean that is was true when the prediction rejected it (Taulli, 2019).

2.4.2.3.2. Decision Trees

Another typical supervised ML model due to its transparency are *Decision Trees*. They are easily understood and work efficiently with large amounts of data (Taulli, 2019).

As seen in the Figure 9, which explains the survival of the Titanic based on sex, age and number of spouses or children (sibsp), the starting point of a decision tree is at the top. From this root node, decision paths will emerge, called splits. In the splits, an algorithm will be used to make the next path choice based on computational probabilities of variables. The tree would come to an end when there are no more splits and therefore, the leaf or outcome is reached (Taulli, 2019).

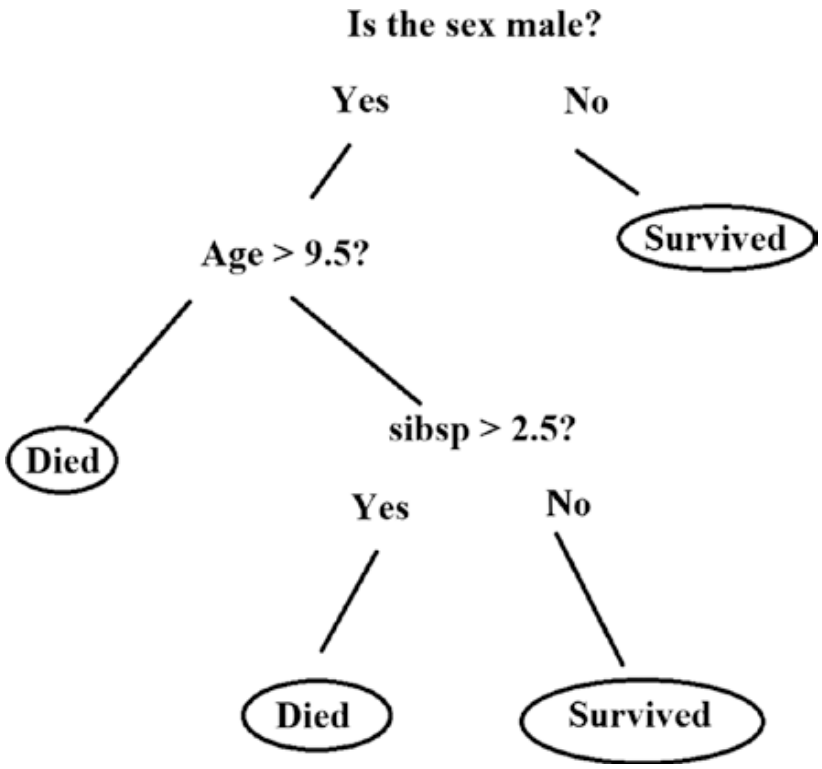


Figure 9. Decision Tree for predicting Titanic passengers' survival (Taulli, 2019).

Drawbacks of this model are the possibility of an error propagation if one of the splits' outcome is wrong and a poor performance if the number of algorithms required to be used is too large due to the complexity of the tree (Taulli, 2019).

2.4.2.3.3. Random Forests

A Random Forest (RF) is just a compilation of Decision Trees which provides an improved prediction over a single Decision Tree. Like them, they are intuitive and transparent supervised ML algorithms that solve classification and regression problems. Contrary to other ML models like Artificial Neural Networks, they provide a clear view of what variables are important for the classification or regression (Rebala, Ravi, & Churiwala, 2019).

2.4.2.3.4. Support Vector Machines

Support Vector Machines (SVM) are supervised ML algorithms used for classification and regression issues since the 1990s. They are large-margin estimation methods used for probabilistic models. It is halfway between parametric and nonparametric approaches. While the former uses parameters to model like linear regression, the later do not rely on them and includes training data directly like Decision Trees do. The main advantage is its ability to work with small data quantities and hence, smaller training datasets whereas its main disadvantage is its large *black box* component (See section 2.5.1). As binary linear classifiers, they divide the data in the space into two classes according to a hyperplane boundary. The main objective is to get the optimal hyperplane that correctly divides the data points, maximizing the margin. Looking at Figure 10, data points in orange represent one class, blue points another one and the red line *b* is the optimum hyperplane. Support vectors are those data points along the dotted lines displaying the maximum margin (Rebala, Ravi, & Churiwala, 2019).

In this case, the example shows a simple two-dimensional space. Hence, the hyperplane takes the shape of a line. However, SVM can work with very large or infinite feature space and the complexity of the classification problem will increase (Koller & Friedman, 2009).

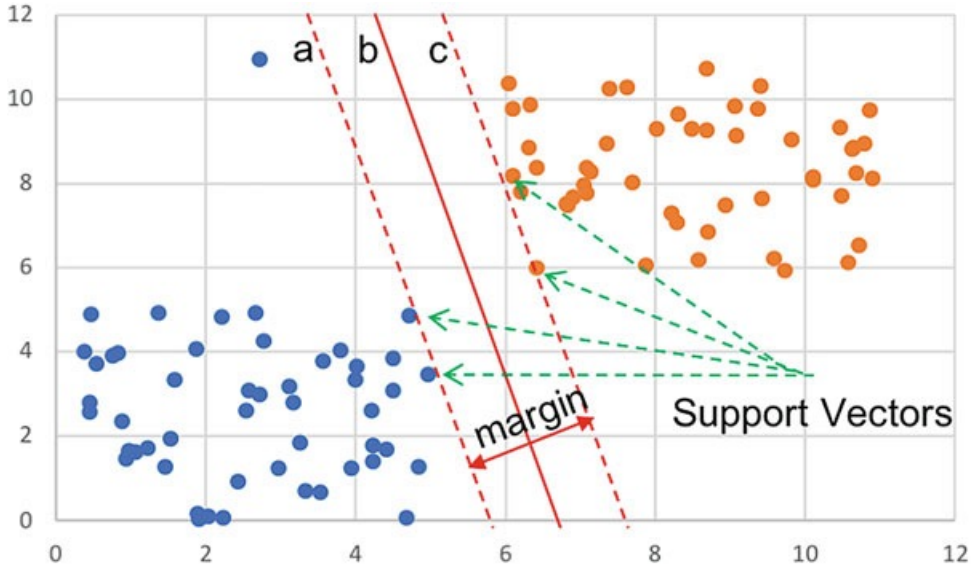


Figure 10. Representation of an SVM (Rebala, Ravi, & Churiwala, 2019)

2.4.2.3.5. K-Means

Not that typical are k-Means clustering algorithms. They are used in Unsupervised Learning to divide unlabeled data into different groups or clusters according to their similar characteristics. The letter *k* refers to the number of clusters and the centroids are the midpoints of the clusters. The k-Means algorithm will calculate the average distance of the centroids and change their location to position them in the center of each cluster (Taulli, 2019). To determine the right amount of clusters or *k*s to be used, bootstrapping cluster analysis are typically used in combination with k-Means (Hofmans, Ceulemans, Steinley, & Van Mechelen, 2015).

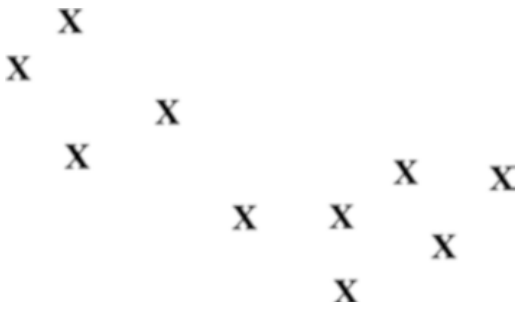


Figure 11. Initial dataset (Taulli, 2019).

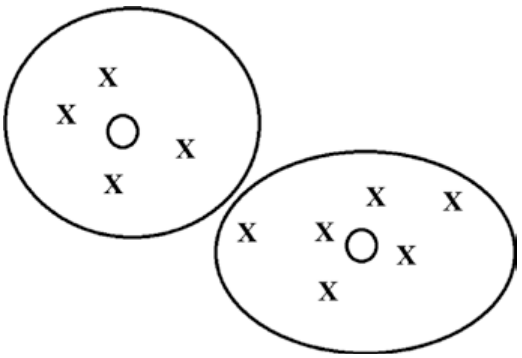


Figure 12. Two clusters with its respective centroids (Taulli, 2019).

2.5. Deep Learning

Deep Learning (DL) is subcategory of ML. To clarify the concept, Figure 13 positions DL in relation with AI basic components.

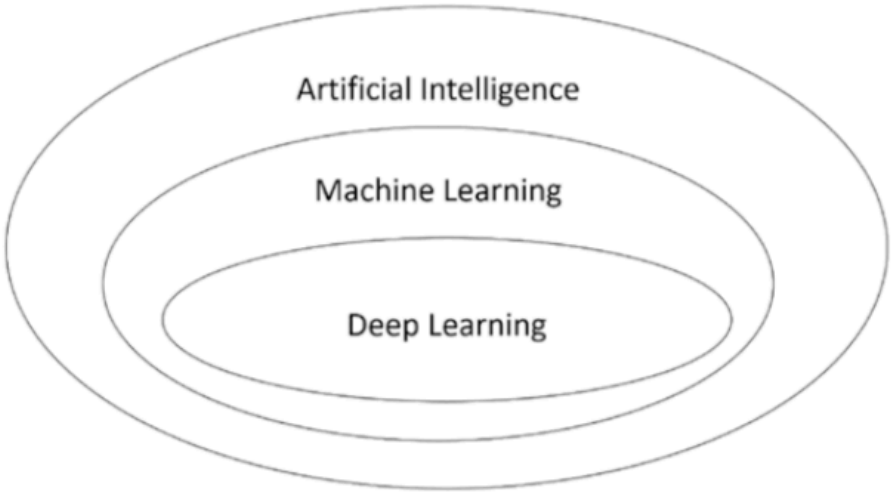


Figure 13. High-level look at the main components of the AI world (Taulli, 2019).

As we can see, AI comprises Machine Learning (ML) and at the same time, DL is a subset of ML, which is also a subgroup of AI. This hierarchy comes from the fact that AI emerged first, ML after and last but not least, Deep Learning (DL) as part of ML.

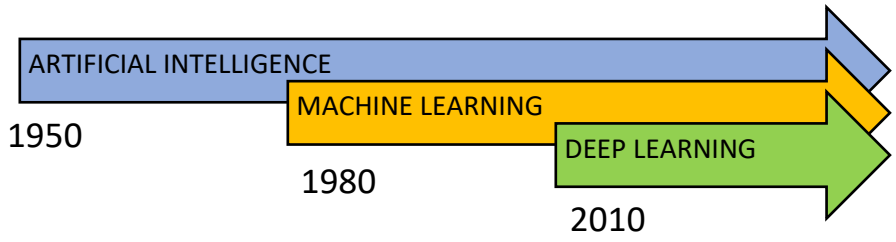


Figure 14. Timeline of the development of AI.

DL consists of an evolution of ML, not so much conceptually, but in its processing capacity, based on the goal of automation, whose intensified development started this decade (LeCun, Bengio, & Hinton, 2015)

DL algorithms are approximations of the processing system of a human brain. In the same way that our brains detect patterns to be used for classification, DL algorithms are thought similarly to do the same, to emulate human neural networks. First, the information is received (input layer) and then, it is compared to a known item (hidden layer), if there is one, to make sense of it. The output would come afterwards (output layer).

2.5.1. Artificial Neural Networks (ANN)

A popular method used in DL are Artificial Neural Networks (ANN). As shown in Figure 15, all the nodes in the same vertical line are input units that correspond to the same layer. And in all layers, except for the input layer, the nodes represent neurons -or artificial neurons-, which have an activation function in them. The rightmost blue layer is constructed by the output units and comprises the output layer, named after the result of its activations. The white nodes in the middle are the hidden units which conform the hidden layers (Rebala, Ravi, & Churiwala, 2019).

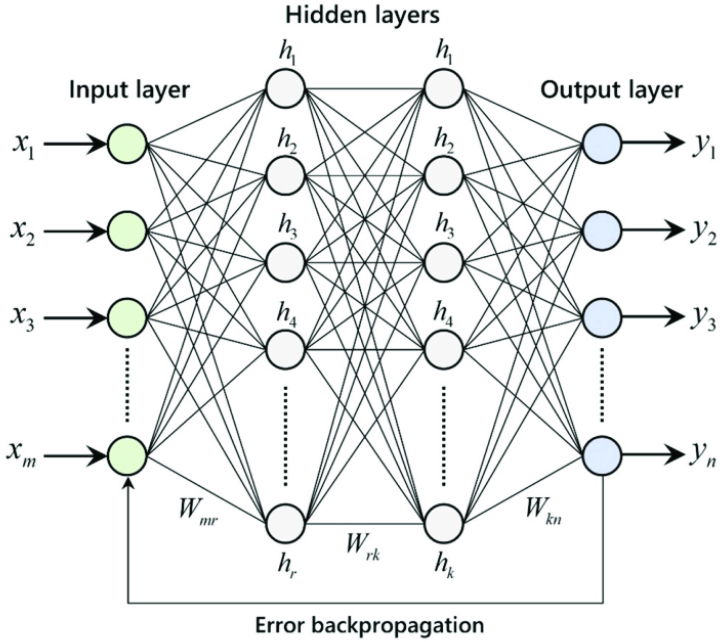


Figure 15. Functioning of an ANN (Fernández-Cabán, Masters, & Phillips, 2018).

They are called hidden because despite being aware of what the neural network receives (input units) and what it comes out of it (output units), what happens in-between (the middle process, decisions, behaviors) is unknown, taking the name of *black box*. Consequently, being able to find the origin of an error or which are the predominant factors influencing the result is very complicated.

Figure 16. Representation of the *black box*.

DL, along with its subfields such as Artificial Neural Networks (ANN) (Bhatt, Bhatt, & Prajapati, 2017), seek the adaptation and replication of the biological structure of a human brain and its capabilities (Martín del Brío & Sanz, 2006).

It is considered a differentiated field in ML due to its greater self-sufficiency to learn and sort out tasks with little human supervision. For example, DL can make a classification of the data taking note of its features (transform unstructured into structured data) whereas ML requires of a human being to find out those features (structure them) and introduce them to the machine previously. This is the reason why DL algorithms are used in Unsupervised Learning.

In order to carry out this complex human free task, another disparity is that Deep Learning needs way more quantity of training data to produce accurate results than another ML techniques and the machines required to support this technology need to have high-end performance. This is why it is not as often as with ML that interactions with DL models occur in our daily lives. However, cancer treatment recommendations and early detection of heart disease are among its remarkable uses (Patrick, 2020) Still, this field has a long way to go. As a matter of fact, Google did not start using DL for improving its search engine until 2015 (Taulli, 2019).

2.6.AI Today

Even though Turing forecasted that a computer would have passed his test by the beginning of this century, many systems have tried but none has succeeded yet, not even Google's Assistant in 2018 after being able to make an appointment with the hairdresser through the phone. And why? Because the chat was over a specific topic and not open-ended. (Taulli, 2019).

With its ups and downs throughout history, AI finds itself in an advanced stage of weak AI after the real explosion of interest in AI started around 2010. The hype cycle occurred due to the impressive growth of computer power, the massive amount of data available leading to Big Data sources and the improvement of some AI approaches such as the just explained Machine Learning (Mochon, 2019).

This acceleration of AI has released new applications, prompted industries and activities and has put a lot of value at stake. In the hype cycle in which AI is encountered,

the speed at which these technologies are evolving is exponential and therefore, resourceful companies like Google or DeepMind are investing and developing research on AGI (Taulli, 2019).

According to a model proposed by Harbers, Peeters and Neerincx (2017), at the moment, we can distinguish three types of AI models depending on the degree of human-machine interaction involved on them:

(1) *Man-in-the-loop*, when AI needs constant human contributions in order to carry out its assignments;
(2) *Man-on-the-loop*, if the machine is capable of acting on its own based on a previous programming, despite the possibility of a human intervention to interrupt or modify its actions at any time;
(3) Man-out-of-the-loop, in which the machine performs independently over some periods of time and during these, the human being has no influence on it.

But just because machines are able to have real-time human interaction does not mean that they have full autonomy. In this regard, certain philosophers of the mind have agreed on considering autonomy not as a quality of dichotomous nature, but as a continuous dimension otherwise. So, it will not be so simple to determine if a behavior performed by a machine holds autonomy or lacks it (Steinfeld, et al., 2006).

For example, automating a response to a given setting taking into account some variables in such a way that a machine, once in operation, can no longer be stopped, would not make AI autonomous. The autonomy of an AI comes from the real capacity to adapt its decisions to different contexts upon the one that was programmed for. For instance, if the Google's Assistant conversation would have progressed until it reaches a point where it is no more about what time is more suitable but what kinds of hair products are preferred, and is still able to follow the storyline, a high level of autonomy would have been achieved. But again, determining if a machine has or has not autonomy is not that simple.

An autonomous machine with such characteristics would belong to a fourth class, *No-man-on-the-loop*, in which the learning process would not stem from a human action but from the machine itself on its own (Miró Llinares, 2018).

All in all, as of yet, machines still need from even the slightest intervention of a human being to operate. It is the man that decides for the machine which stimuli to respond to

and how, which ones to ignore, and feeds it with the data from which it will start its learning process. In essence, at the present time machines are not able to understand complex contexts or introduce new variables on their own. They just act according to basic logical premises that have been previously introduced in their system by humans and therefore, AGI is still far from being achieved (Taulli, 2019).

There is a strong probability that computer systems capable of learning autonomously from the environment and adapting to it, similarly to a state of consciousness (McDermott, 2007), will be designed in the future, but at the moment that is no more than a utopian reality. Nevertheless, the scope of NAI subfields should not be underrated although individuals only interact with a small number of them.

3. THE FATE OF AI

Each of the terms of the Fairness, Accountability, Transparency and Ethics (FATE), refer to the fundamental features that should characterize any system involving AI. After a brief introduction to the origin of this movement, each of the characteristic will be described in the subsequent sections.

3.1. Origin of FATE

The ubiquity of AI presently is unquestionable. AI has the potential to solve problems in many areas such as Policing or Education but it also threatens the human-centered functioning way of the current world (Porayska-Pomsta & Rajendran, 2019). In summary, there is a dilemma between AI as a human replacing machine (AI) or as an intelligent assistant (IA) (Korinek & Stiglitz, 2017).

Artificial Intelligence and corresponding technologies have rapidly become part of many real-world applications and started to be used to make high stake decisions in several areas such as education or the Criminal Justice System. The problem is that those domains can easily affect fundamental rights and liberties in significant manners (Porayska-Pomsta & Rajendran, 2019).

As a matter of fact, there is a growing concern that AI systems have a tendency to reinforce social inequalities and injustice instead of diminishing them. AI acts as a mirror reflecting our current understandings of human intelligence which therefore include intellectual and empirical limitations that may result in biases. The two mains reasons for supporting this belief are (Porayska-Pomsta & Rajendran, 2019):

1) The data required for the functioning of AI models have implicit socio-cultural biases, mainly due to prejudices based on race, gender or ethnicity and a lack of representative data of the society as a whole, which results in biased models as well.

2) The scarce inspections or discussions made about most of the AI models, especially in ML. The most dangerous contributor is the *black box*[4] component of AI systems, that prevents human from noticing the existing biases or their origin that may come from the data, the algorithms used or both. This is also known as the *interpretability problem* (Brinkrolf & Hammer, 2018).

Nevertheless, it is important to remember that AI technology is applied in already imperfect social contexts (Education, Healthcare, Social Justice, etc.). Those systems are the reflection of the current knowledge and social structures that determine who has the power to influence on them. Hence, they are inherently biased representations of the world that are far from representing absolute truths (Porayska-Pomsta & Rajendran, 2019).

3.2.Fairness

The term Fairness in the FATE realm (algorithmic Fairness) refers to the impartial, just and non-discriminatory way of treating people. But Fairness, as well as discrimination, are difficult concepts to delimit. Fairness has many different interpretations and each of them is designed to content a specific social group, which will determine its meaning. Nevertheless, the definition will ultimately be portrayed by the social context.

On the other hand, discrimination, according to philosopher and legal scholar Deborah Hellman (2008), is actually wrongful discrimination: *we make distinctions all the time, but only cultural context can determine when the basis for discrimination is morally wrong*.

Unfairness is present in "any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable" (Holstein, Wortman Vaughan, Daume III, Dudik, & Wallach, 2019). Therefore, for an AI/ML model to be

[4] A computer based system is usually defined as a *black box* composed by its inputs, outputs and the relationship between the two of them (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). See ANN in Chapter 2 for further details.

fair, a person's experience with it should not vary depending on personal features such as their belonging to historically discriminated groups based on race, gender, sexual orientation, ethnicity, religion or age (Pedreschi, Ruggieri, & Turini, 2008). The problem is that sometimes, in order for Fairness to be attained, the speed of an algorithm has to be sacrificed and, many of those times, the importance of the rapidity will outreach the importance of Fairness.

Imagine an AI model designed to aid in hiring processes. It will try to achieve the best predictive accuracy possible on how good a candidate will perform in a job position but will not pay attention to any gender or racial bias (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). Similarly, a risk assessment tool designed to help taking decisions at the different stages of the Criminal Justice System, will predict things like the likelihood of failure to appear in court or to commit a crime in the future (VanNostrand & Lowenkamp, 2013) but will not guarantee the Fairness of the decision.

In AI, the attention is focused on achieving the objective that has been set for the model regardless of Fairness issues (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). This, added to the potential of AI to intensify social inequalities can became a serious problem. To mitigate it, statistical definitions of Fairness and algorithmic methods to palliate biases against those definitions have been developed (Narayanan, 2018).

3.2.1. Fairness in ML

Due to this, it has been recently developed a field that looks after a fairer and more just machine learning models, denominated the Fairness-aware Machine Learning (fair-ML). Even though Fairness and Justice are features of social and legal contexts and not of technological devices, the purpose of this movement is to introduce Fairness into the equation, or more precisely, into the algorithms, making it part of the *black box* system (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019).

For a ML system to work, it requires from abstraction of the social context or the specific aspect of a problem. Failure to consider the interactions between the social and the technological can lead the system to fall into an abstraction or category error, which could result in five different traps (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019).

- The first and most common one is called the *Framing Trap,* defined as the failure to model the entire system over a social criterion, such as Fairness, will be enforced. In simple words, the efficacy of the algorithm is evaluated as output (outcomes) related to the input (data) and the goal is to create a model that best obtains the relationship between them, with no concern about fairness.

  From the fair-ML perspective, it is important to understand how the chosen data and outcomes affect the resulting model and seeks to include Fairness as a goal as well. Fairness cannot exist just as a technical standard, with no human involvement taking part of the scheme (Latour, 2005).

- Secondly, the *Portability Trap*. Portability, a primary objective in system design, aims to create a solution applicable to different social settings (either predicting risk of recidivism or the capability of a new employee). Nonetheless, if applied to social purposes, it prompts The *Portability Trap,* which refers to the unsuccessful attempt of using an algorithmic system designed for a certain social context to another one, not even within the same domain. Just a change in location can make the local Fairness concerns to be different, for example, from one court jurisdiction to another.
- In the third place, the *Formalism Trap* is the failure to account for the full meaning of social concepts such as Fairness, which can be procedural, contextual and contestable, and cannot be resolved through mathematical formalism. Anyway, the community of fair-ML has been trying to define Fairness in mathematical terms as a way to include it into ML.
- In the fourth place, the *Ripple Effect Trap* accounts for the lack of understanding of how an introduced technology interacts with a pre-existing social setting. To avoid it, it is essential to understand the response of the actors to these technical interventions in each context
- Ultimately, the *Solutionism Trap* means that technology is not always the solution to our problems. No technological intervention is preferred in two situations: when Fairness definitions are changeable or dependent upon political forces because the system might struggle to keep up; and when the model is computationally unmanageable due to its sophistication. In order to answer the question of whether technology is the best solution to a problem, a comprehension

of the context is needed (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019).

In order not to get caught by these traps, fair-ML researchers propose the following solutions to overcome each of them (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019):

1. For the Framing, include data and social factors to construct a heterogeneous frame that takes into consideration Fairness.
2. For the Portability, introduce enough social and technical provisions of the desired context.
3. For the Formalism, remind that social conditions such as Fairness are at times procedural, contextual and political. Therefore, ensure that the model can handle them.
4. For the Ripple Effect, anticipate the way in which technology will affect the social context to assure that the tool will still solve the problem that it was meant to fix in the first place.
5. For the Solutionism, design a system, if necessary, that befits the specific social context.

In the light of the foregoing, technical solutions could be formed as long as all these traps are borne in mind and establishing its limitations. In any context, the first question to be asked is if it is an appropriate solution. A research of the legal, social and political context should precede any implementation, and also, a definition of what the concept of Fairness embraces in there. Once it is established that is either the best way or that there is no other way because, for example, its adoption is mandatory by law, then a look to the previous traps will be highly recommendable.

Some recommendations given for ML industry are developing a Fairness-aware data collection, so that instead of trying to fix the models with de-biasing algorithms afterwards, high quality data is gathered the first place; and watching the blind spots while doing so, ensuring that enough data is collected from potentially problematic subpopulations and anticipating other forms of unfairness that may arise in the specific ML application. Predicting bias and unfairness before implementation is critical and still most are only noticed after customer complaints or bad media coverage (Holstein, Wortman Vaughan, Daume III, Dudik, & Wallach, 2019).

For that aim, proactive auditing processes are important. Efforts to find biases are not usually rewarded neither are part of the workflow, but auditing is key to achieve Fairness. Those audits have to be domain-specific or they will fell into the *Portability Trap* as metrics used for one context will not be valid for another. Also, having access to individual-level demographics (e.g. gender, race) is essential to find bias problems. Even though encryption mechanisms have been proposed to ensure anonymity or confidentiality, this is still a controversial issue. In the third place, auditing allows to find systemic problems induced from specific ones (Holstein, Wortman Vaughan, Daume III, Dudik, & Wallach, 2019).

Ultimately, note that any interventions in datasets or models to achieve Fairness can have unwanted side effects, turning it into a bigger problem rather than a solution. However, that should not be enough reason to adopt "band-aid" fixes instead of addressing the root of the problems (Holstein, Wortman Vaughan, Daume III, Dudik, & Wallach, 2019).

3.3.Accountability

Accountability means taking responsibility for the action taken and the situation that they led to. This dimension takes great relevance in terms of decision-making issues, both human-made or those involving AI. The difference relies on the fact that there has not been time to properly understand the nature of the implications that AI can have in the way society is organized or the chance to determine necessary accountability measures to control and take protection against any harm. Although the definition is not clear, Accountability is about giving autonomy to people in order to take action through knowledge, a feature that should characterize any democratic society (Porayska-Pomsta & Rajendran, 2019).

It is undeniable the human tendency to avoid having to decide between too many choices and delegating the responsibility of doing so to others. Human decision-making is prone to be based solely on prior knowledge due to their inherent resistance to change, and to choose simpler rather than complex strategies relying on first impressions with no logical thinking involvement. Contrary to this, AI presents an alternative for finding the optimal strategy that releases human beings from the efforts of making decisions. However, there has to be a way to hold these decisions accountable (Rebala, Ravi, & Churiwala, 2019).

The two prevailing perspectives adopted in law and policing are the *post-factum* and the *pre-factum* accountability. The former involves blaming an agent who manipulated another agent for their own benefit and hold them accountable for the consequences afterwards. The latter requires a prior blamable agent or agents that represent institutional preferences regarding aspirational values such as justice, democracy or racial discrimination. The goal of this type of accountability resides on achieving a change in society or mass realization before blameworthy actions are taken (Porayska-Pomsta & Rajendran, 2019).

Nevertheless, apart from providing a moral and legal framing, these two perspectives do not clarify how to effectively operationalize accountability so that it can be actioned in different contexts and in a constantly changing world. As a matter of fact, AI representations of the world showed that current systems rely on historical and socially biased data and that the principal accountability measures and law are having difficulties catching up with predominant social values and shifting norms (Porayska-Pomsta & Rajendran, 2019).

The necessity of having a system that holds accountable AI solutions and decisions is out of question taking into consideration the potential that it has to intensify the socio-cultural biases inherent to every society (Porayska-Pomsta & Rajendran, 2019). Also, finding a solution to the *interpretability problem* mentioned before is seen as a key for accountability and trustworthiness of decisions taken with the support of AI models (Lipton & Steinhardt, 2019). In brief, AI systems have to be contestable and open to modifications by the users (Brinkrolf & Hammer, 2018).

Therefore, a new more flexible ethics-based approach has been developed in which Accountability is composed by adjustable rules and moral codes depending on needs and changes occurring in society. It becomes a compromise between conflicting interests of decision makers in which there is no unique way of making others accountable, and where the blameworthiness of the actions will depend on the interests of the stakeholders affected. In that sense, inclusion of every group of stakeholders is also fundamentally important (Porayska-Pomsta & Rajendran, 2019).

3.4.Transparency

It claims for a clear exposition of the processes behind any AI system and the emptying of the *black box* as much as possible. Transparency adds value to a model as it allows for further details in explanations beyond "these are the inputs, and these are the outputs". There has to be an explanation for every outcome and therefore, simple algorithms like RF or logistic regressions are usually chosen over more complex ones such as SVM or ANN. The last two have large *black boxes* which make it harder to figure out what the model is doing and where results came from (Veale, Van Kleek, & Binns, 2018).

Many companies and organizations are actually publishing the weights used by their algorithms, their scoring guidelines or the limitations and ethical challenges they had to deal with in order to improve the transparency of their models. On the other hand, opacity is seen as the only way to sustain the utility of a model and the only protective measure against the *gaming* or manipulation of the data-driven systems (Veale, Van Kleek, & Binns, 2018).

In any case, Transparency has to cope with the respect to privacy. Closely related to Ethics and Fairness, privacy matters are very valuable to avoid discrimination and promote autonomy. Autonomy allows for the liberty of decision, free of manipulation and coercion, and hence, threats to privacy result in a limitation of that freedom and a degradation of welfare (Doyle, 2019).

Large amounts of data are gathered uninterruptedly every day without people's awareness of what is actually being collected, for what purpose or who will have access to it. All that information is used for identification, classification, assessment and distribution of people in an effective discriminatory way for the profit of businesses (Doyle, 2019).

However, privacy does not equate anonymity. In order to maintain privacy, anonymity is not enough but it helps in a significant way to protect it. Even when the specific information of a person is unavailable, acquired facts on similar others can be sufficient. Advanced data analysis techniques have enabled this. Then is when obfuscation acquires relevance. Its chore is to make data ambiguous and thus, it helps with anonymity and consequently, privacy and autonomy. Nonetheless, the disadvantages

are considerable as it makes data harder to use and reduces its value by ultimately, undermining Transparency (Doyle, 2019).

3.5.Ethics

The role of Ethics comes into play when deciding whether certain characteristics should be used as predictive variables. Gender, age, ethnicity or sexual preference are protected characteristics disallowed by the U.S. Constitution to be used by a model in some contexts. Sometimes, even when they are not forbidden or there is actually pressure to use them (e.g. race as a predictor for re-offending) they should be avoided on ethical grounds (Veale, Van Kleek, & Binns, 2018). The use of proxies in the attempt to avoid using protected characteristics is also risky, as the proxy outcomes could be related to other variables different to the variable of real interest (Veale, Van Kleek, & Binns, 2018).

Fighting for the avoidance of the incorporation of these kind of personal information is very difficult because individuals are usually unaware what information is being collected in the first place and in the second place, they are not in the position to discuss its use either (Doyle, 2019). But this problem, as stated before, is tied up with the privacy and anonymity issues discussed in the Transparency section.

Another ethical issue concerns a fundamental distinction between AI and human intelligence. Even though AI tries to imitate human behaviors, it is not limited by human cognition or functioning. Indeed, it can bypass human capacity in many areas. The delegation of tasks to AI machines is already transforming the interactions and the environment in which humans used to live (Pagallo & Durante, 2016).

For that delegation to be successful, it has to be supported by trust. Trust is the center of any relationship involving giving over duties and its best asset is that is provides security. More trust means more security and that is something that the human being always looks for. Security guarantees that individuals are in a protected condition from any danger where they can enjoy a carefree life (Pagallo & Durante, 2016).

This can result in a sense of disempowerment for humans, but it is also an inspiring tool for to improving their abilities and reflect on who they are and who they want to be. To bring an example, professional players of the board game Go are learning strategies

from AlphaGo[5] since the computer program beat the world champion with one unknown to him. Indeed, AI has a great potential to motivate human learning and creativity However, it is limited in its capacity for new inventions (e.g. a new board game), imagination, critical thinking -apart from weighting gains and losses- and lacks moral judgement (Porayska-Pomsta & Rajendran, 2019). Again, the main dilemma relies between seeing AI as an autonomous mechanism that attempts to substitute human beings or it is used for enhancing human capabilities and as a supportive technology (Korinek & Stiglitz, 2017).

In summary, encoding social practices to be followed when AI is involved is very difficult. But despite the ethic challenges raised by the AI models, many public agencies and private companies are working in their own in-house ethical codes for any activities involving AI. Some additional proposals include the creation of informal and dynamic knowledgebases and virtual communities to share and discuss ethical questions related to algorithmic practices. Nevertheless, Ethics are impossible to arrange because each individual will have its own view of what it concerns (Veale, Van Kleek, & Binns, 2018).

[5] AlphaGo, or AlphaGO-DM because of its creation by Google Deep Mind, is a computer program designed to play the ancient Go boardgame, considered one the most challenging games ever invented because the solutions require more than knowing rules (Silver, y otros, 2017).

4. RISK ASSESSMENTS IN THE U.S. CRIMINAL JUSTICE SYSTEM

4.1.History

Since the 1970s, the U.S. criminal justice system has had a backward-looking retributive punishment perspective where an offender's who breaks the law should suffer from blameworthiness with the corresponding conviction (Monahan & Skeem, 2016). Due to this, the United States has a prison population higher than any other country in the world. Even when the number of lock ups has declined progressively over recent years, still 1.489.400 inmates were living in federal and state correctional facilities by the end of 2017 (Bronson & Carson, 2019). The U.S. is the leader in incarceration by far, with imprisonment rates exceeding those of any other country in the world (Wagner & Sawyer, 2018). Furthermore, the racial proportions of the U.S. population given by the 2014 Census accounts for 62,1% White, 13,2% Black or African American and 17.4% Hispanic yet the prison population is categorized disproportionately as 37% Black, 32% White and 22% Hispanic (Flores, Bechtel, & Lowenkamp, 2016).

Besides the negative effects that come with incarceration, there is a tremendous economic cost that exceeds official correctional budgets. According to a report published in 2012 by the Vera Institute of Justice (Henrichson & Delaney, 2012), incarceration financial costs hit the sum of 39.5 billion dollars per year. Lately, an increased concern about these massive economic and human costs involved in mass incarceration has led to the emergence of a reform movement in sentencing and corrections (Lawrence, 2013). This movement has changed the limelight from punishment and retribution to ways of reducing over-incarceration and recidivism (Subramanian, Moreno, & Broomhead, 2014). The sentencing reform has put the emphasis in forward-looking approaches of an offender's potential recidivism, more geared towards a utilitarian rather than retributive theory of punishment. As a result, the existing model of criminal punishment is a combination of both theories, known as "limiting retributivism" (Morris, 1974; Frase, 2004).

A proposed solution to decompress mass imprisonment without jeopardizing public safety is to use risk assessment instruments to lighten and speed up the workload behind sentencing and corrections. The finality of risk assessment is to forecast the likelihood of an offender to reoffend or recidivate as per determined risk factors such as age or criminal history record (Monahan & Skeem, 2016). Statutes and regulations over the U.S. are

increasingly implementing risk assessments to provide with the necessary information required to decide about the locking up high-risk offenders, the supervised release of low-risk ones and the prioritization of treatment facilities to minimize potential offender's risks (Lawrence, 2013). As a matter of fact, the state of Virginia used risk assessments at sentencing and released 25% of non-violent low-risk offenders from prison without increasing the crime rate (Kleiman, Ostrom, & Cheesman, 2007).

Indeed, the Model Penal Code establishes in its section 1.02 (2) that sentencing decisions are taken "within a range of severity proportionate to the gravity of offenses, the harms done to crime victims, and the blameworthiness of offenders (...) to achieve offender's rehabilitation, general deterrence (and) incapacitation of dangerous offenders (American Law Institute, 2017). Therefore, it is a mixture between the retributive aspect that sets a timeframe for the sentence and the utilitarian view through risk assessments that determine the specific sentence. For example, the punishment for robbery in the state of California ranges from up to 5 or 9 years in state prison depending on if it was a first degree or second-degree conviction (Cal. Pen Code § 211). If an offender scores low risk in the risk assessment, the sentence he or she will get will likely be in the lower end of the interval, while a high-risk offender will be closer to the maximum penalty. It must be recalled that in any case should risk assessments be used to determine a longer sentence than the one they would have received otherwise according to what they morally deserve (Skeem & Lowenkamp, 2016).

Although it has been roughly a century since the beginning of risk assessments usage, nowadays the central role of them is undeniable. The first tools were built to predict future re-offending by usually selecting factors inferred to be risk predictors, each of them with a given weigh depending on the level of predictability and integrating all of them into a risk score. Based on the outcomes, supervision resources could be administrated more efficiently, intensifying it for those at most risk and lessen it when the risk is low. Instruments that came afterwards, however, put the focus on reducing risk by including risk factors that are variable or dynamic as "needs" to be tackled in supervision and treatment. These "needs" would support the principles of an Evidence-Based correctional system where the risk level indicates who ("risk" principle) and what ("need" principle) should be primarily treated (Skeem & Lowenkamp, 2016). For instance, the Sentencing Reform and Corrections Act of 2015 imposed the use of risk assessments to allocate federal inmates in recidivism reduction programs suitable for them (e.g. drug

rehabilitation, work and education) and those who accomplish them could benefit from an early release up to 25% of their sentencing time left.

4.2. Evolution of Correctional Assessments

Correctional practice, in terms of activities related to treatment, punishment and supervision of people convicted of crimes, has been evolving from first-generation (1G) to fourth-generation incipient methods during the last three decades (Andrews, Bonta, & Wormith, 2006). Failures and weaknesses of previous stages have been overcome to reach the 4G approach today.

The first-generation (1G) approach was based solely on clinical and professional judgment due to a lack of objective scoring systems. Hence, its main problem was subjectivity, inconsistency, bias and potential stereotyping, legal vulnerability and lower predictive validity introduced by humans (Grove & Meehl, 1996). Anyway, it was the main approach used in corrections and it is still preferred by some correctional decision-makers.

Subsequently, second-generation (2G) assessments grounded on additive point scales where each selected factor had a weight (Austin, 1983). However, these simplistic scales were, according to Dawes (1979), no more than "improper" linear models with few standardized factors and its respective weights decided by either common sense or professional consensus instead of statistically. The emphasis was put on risk prediction, concision and efficiency and the weaknesses were based on an absence of theoretical background, limited selection of risk and need factors with disregard to the dynamic ones, no treatment implications, little explicative use and arguable applicability for female offenders (Jones, 1996). Then again, these linear models were highly effective when it came to validity, exceeding professional opinions in most of the cases (Grove & Meehl, 1996).

On the late 1970s and 1980, the third generation (3G) emerged. These approaches were more objective, with theoretical and empirical foundations, and a wider range of risk factors, including dynamic ones. A benchmark of this type of assessments is the Level of Service Inventory-Revised (LSI-R), designed by Andrews and Bonta (1995). However, its main criticisms as well as for the rest of 3G methods, included an insufficient theoretical framework (basically Social Learning Theory), neglect of gender sensitivity,

a rampant attention on risk and a lack of concern for offender's strengths or protective factors that are of primary attention as claimed by the Good Lives Model (GLM)[6] (Andrews, Bonta, & Wormith, 2006; Ward & Brown, 2004; Ward & Stewart, 2003).

Finally, among 4G instruments, the Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) can be found along with others such as the Correctional Assessment and Intervention System (CAIS) and Level of Service/Case Management Inventory (LS/CMI) (Andrews, Bonta, & Wormith, 2004; 2006). The characteristics that distinguish 4G approaches to the rest are (Brennan, Dieterich, & Ehret, 2009): a more extended theoretical background; additional risk and need factors that provide content validity; introduction of the strengths perspective of the GLM; more sophisticated statistics; a perfect integration of the need or risk domain with the management information system (MIS)[7], criminal justice databases and web-based implementation of assessment technology.

The consolidation of all these features enables the tracking of offenders during its passage through the criminal justice system, from the very beginning to the very end, and provides support for subsequent case management monitoring, information feedback and decision making. In these dissertation, two instruments widely used will be described in detail for further analysis: LS -specifically, LSI-R and LS/CMI- and COMPAS.

4.3.The Level of Service (LS) Assessments: LSI-R AND LS/CMI

The creation of the initial version of the Level of Service (LS) assessment started by Andrews (1982) in the late 1970s, in collaboration with the Ontario Ministry of Correctional Services (Casey, et al., 2014). The aim was to create a comprehensive

---

[6] The GLM is a strengths-based theory of rehabilitation premised on the idea that crime is just the wrong way of satisfying primary needs that every human being has. Therefore, given offenders the tools to develop and implement meaningful life plans future offending would not take place. It is a positive approach conversely to the prior ones concerned mainly with risk factors and individual deficits (Ward & Brown, 2004) .

[7] A management information system (MIS) is a computer system (composed by hardware and software) that is the main core of an organization's operations. An MIS collects data from multiple online systems, analyzes the information, and reports data to support in management decision-making. (McLeod, 1995)

instrument that registered the characteristics of offenders and would help establishing the level of supervision required for each of them. Subsequently, improved versions of the tool were designed until the release in 1995 of a 3G tool, The Level of Service Inventory-Revised (LSI-R) and its 4G updated version, The Level of Service/Case Management Inventory (LS/CMI) in 2004, that are still currently used in the U.S., Canada and other countries worldwide (Andrews & Bonta, 1995; Andrews, Bonta, & Wormith, 2004).

All of the versions have fundamentally the same features, differentiated only by some innovations and adaptations to the context. In essence, LS instruments are quantitative assessments based on static and dynamic risk and need factors, scored dichotomously, given a 1 if the item is present and 0 otherwise. The design is intended to be applied across populations of different ages, gender, race and ethnicities (Wormith & Bonta, 2018).

The purpose of the instrument is to predict recidivism and other criminogenic conducts on the short (less than 6 months) and the long (more than 2 years) term. Originally, the LS was oriented to sentenced offenders either imprisoned or on probation or parole. With the years, an increased introduction on pre-sentence stages took place, serving as a tool to inform about the offender's needs and indicate its suitability for community supervision, bail and other pre-conviction decisions (Wormith & Bonta, 2018).

4.3.1. LS Versions

Due to the development of theory and practice research in the field, the original LS instrument turned into successive improved versions. At the very beginning, more than 25 years ago, an extensive document containing more than a hundred predictive risk factors was produced by Andrews (1982). In order to make the instrument more user-friendly for probation officers, several meetings were arranged to reduce the number of items. The subsequent meetings led to the fourth version of the Level of Supervision Inventory (LSI-IV) (Wormith & Bonta, 2018).

Not satisfied enough, Andrews kept on investigating to create the following-up LSI-V and LSI-VI versions. The modifications were minimal such as adding or dropping items. For example, LSI-VI had 58 items which did not include age or gender. Later on, though, research proved that not considering them did not affect the results (Wormith & Bonta, 2018).

The LSI-R came after LSI-VI. Andrews and Bonta (1995) started working on it in 1993. The name was changed by substituting the word "supervision" to "service" to better show its treatment purpose. This version was almost identical to its precursor except for the inclusion of two age related items (under 18 years old or under 20) and the further reduction to just 54 items by deleting the probation and parole conditions subcomponents which lacked predictive validity. The final number of subcomponents were 10, with Education/Employment as the main predictors and Leisure/Recreation as the least. The publication of the LSI-R was accompanied by a user guide with instructions on how to conduct the interview and other scoring guidelines as well as the theoretical foundations of the tool up to the date.

The LSI-R, a 3G instrument, became the main classification instrument for institutions in the course of the following years (Wormith & Bonta, 2018) but as its use grew, it was noticed that jurisdictions with high work volumes were in the need of a tool to triage their cases. Thus, a screening instrument called LSI-R Screening Version (LSI-R:SV) was devised with just 8 items (Andrews & Bonta, 1998).

Later on, an increased interest on the Evidence-Based Correctional Practices started to emerge. In 1994, a review of the LSIR-R was instituted by the Ontario Ministry of Community Safety and Correctional Services to address some concerns such as the validity of the tool when applied to specific groups of offenders or the absence of strengths and non- criminogenic needs. As a result, the Level of Service Inventory-Ontario Revision (LSI-OR) (Andrews, Bonta, & Wormith, 1995) was created. The main changes on this version were three:

1. Rearrangement of the subcomponents in accordance to the Central Eight risk/need factors. The Antisocial Pattern subcomponent was created, and Financial Problems, Accommodations and Emotional/Personal subcomponents were either included there or under a new section called Other Client Issues.
2. New sections for the assessment of concrete risk/need components such as sexual or violent risk factors, prison experience and responsive factors. However, those sections had no scores and were not used for classification.
3. The assessment of strengths or protective factors was introduced. For each component, the interviewer could note a potential strength in an organized way.

In 2004, the LSI-R and LSI-OR were updated to give birth to the Level of Service/Case Management Inventory (LS/CMI). In the LSI-OR, case management planning sections were missing so this version included them extra sections for additional support in the development of an offender's case plan and tracking its progress once it is implemented (Casey, et al., 2014). The structure of the sections in LSI-OR was mostly maintained with only few item modifications. The combination of case management with assessment turned the LS/CMI into a 4G tool (Andrews, Bonta, & Wormith, 2010).

In any case, for those correctional agencies which already had a case management system, the Level of Service/Risk, Needs and Responsivity (LS/RNR) was published in 2008 (Andrews, Bonta, & Wormith, 2008). Additionally, youth versions of the tool were also created, being the Youth Level of Service/ Case Management inventory (YLS/CSMI) the current and most popular one (Hoge & Andrews, 2002).

4.3.2. LS Design

Items on the scale have been selected on theoretical and empirical grounds, mainly the General Personality and Cognitive Social Learning theory (GPCSL) (Bonta & Andrews, 2017). This theory relies in the belief of multiple causes triggering the antisocial behavior that can be grouped in eight major areas of influence, the Central Eight risk/need factors. A comprehensive analysis of the individual is key in GPCSL as crime is not considered simply a consequence of substance abuse, poor self-control, etc.

The GPCSL is a general theory of criminal behavior which postulates that any kind of behavior is learned and repeated over time according to the fundamentals of the Social Learning Theory: behavior is learned through the observation of others and conducts are maintained if reinforced or eliminated if punished. The sources of the rewards and punishments can stem from different domains and that is why supported by the GPCSL, several subcomponents are addressed. The influence of GPCSL can also be appreciated in the items. If each subcomponent and its items are examined, information about the level of rewards or punishment of prosocial or criminal behavior can be acquired. For instance, in the Education/Employment subcomponent, if a person is unemployed, there would not be any observation of prosocial behaviors or activities. However, if the individual has a job, the co-workers can disapprove antisocial conducts and encourage prosocial ones. The GPCSL also posits the application of the Central Eight across gender, age and race. These features as well as poverty are considered in a separate

section (Special Responsivity Considerations) for its responsivity potential but they are not central factors in the model (Wormith & Bonta, 2018).

Table 3. LSI-R and LS/CMI Subcomponents (Casey, et al., 2014).

Subcomponents	**Items**	
	LSI-R	**LS/CMI**
Criminal History	10	8
Education/Employment	10	9
Financial	2	
Family/Marital	4	4
Accommodation	3	
Leisure/Recreation	2	2
Companions	5	4
Alcohol/Drug Problems	9	8
Emotional/Personal	5	
(Procriminal) Attitudes/Orientation	4	4
Antisocial Pattern		4

The rearrangement of subcomponents to reflect the Central Eight risk/need factors did not happened until LS/CMI and LS/RNR because the theory was not fully developed when LSI-R was created. The Emotional/Personal subcomponent was substituted by Antisocial Pattern because the former was excessively focused on emotional distress and mental illnesses and did not pay proper attention to antisocial personality traits. Also, static factors were deleted to emphasize the characteristics of the offender that could be changed through case planning and treatment interventions except for Criminal History.

This lone static factor is preserved due to its relationship with a life of many rewards and scarce penalties for criminal conducts if the criminal history is long, as the Social Learning of the GPCSL explains. Most importantly, GPCSL brought to light the importance of strengths to reduce offender's risk of recidivism. Each subcomponent could be seen not only as a risk predictor if present, but also identify strengths that could be help in the success of a treatment plan. For example, in the Criminal History subcomponent, a strength will be the occasional or inexistent criminal behavior (Wormith & Bonta, 2018).

Section 1 in the LSI-R includes 54 items divided in 10 subcomponents while LS/CMI has only 43 items across 10 subcomponents, as shown in Table 3.

The additional sections that receive no score in the LS/CMI are important though to collect information on influential factors that may trigger the criminal behavior. These are (Andrews, Bonta, & Wormith, 2004):

- Section 2: Specific Risk/Need Factors
- Section 3: Prison Experience-Institutional Factors
- Section 4: Other Client Issues
- Section 5: Special Responsivity Considerations
- Section 6: Risk/Need Summary and Override
- Section 7: Risk/Need Profile
- Section 8: Program/Placement Decision
- Section 9: Case Management Plan
- Section 10: Progress Record
- Section 11: Discharge Summary

An example of an LSI-R and a LSI/CMI Report can be found in Appendix A and B, respectively.

4.3.3. Data Collection Method

The data collection protocol consists essentially of an interview, supplemented by other sources of information such as file documents (e.g. criminal records, pre-sentence reports) or interviews with people close to the individual (e.g. family, co-workers) to assign a score to each item.

In order to guarantee a proper implementation of the tool, two directives must be followed: its application by trained staff and a constant monitoring of the implementation to ensure it is applied as intended. The training is provided by high-qualified personnel that met the standards set by the LS authors to provide the best training for new users, but the ongoing monitoring has to be made by the user agency. It is the responsibility of the institution to control carefully the quality of the implementation by the staff and that maintain the levels of competency. Override decisions have to be justified in all cases as they diminish the accuracy of the instrument so monitoring can ensure that it is not used excessively.

4.3.4. Scoring

All versions of the LS contain a User's Manual with instructions and scoring indications for every item. Both LSI-R and LS/CMI tools calculate a unique risk and needs score by summing the individual scores of each item in Section 1. The items receive a dichotomous scoring, that is, "1" if the item is present or "0" if the item is absent.

If an item cannot be scored because information is missing, users are encouraged to leave it blank instead of prorating. This decision is made on the basis that for an item to be scored with a "1", evidence of its existence must be present. Otherwise, the client gets the benefit of the doubt. Depending on each version, the number of items allowed to be unscored differ but usually is up to 4 even though that rarely happens. Although Sections 2 to 11 generate no score, the information contained in them takes relevance for supervision and treatment decisions.

Additionally to the 0 or 1 scores, after the increased interest on dynamic factors, the LS versions that came after LSI-VI include a number of dynamic items that are first given a score on a four-point scale, from very unsatisfactory with a clear need for improvement (0) to satisfactory with no need for improvement (3). Then ratings are grouped into problematic (1) and not problematic (0) for their addition to the final score. This modification was made with the purpose of pointing further differences between offenders and being able to notice small negative or positive changes over time that were not enough to change the overall score of an item. Moreover, most of the dynamic items were continuous variables subject to change in short periods of time such as performance at school or work. Hence, a four-point scale gives the chance to report changes in either direction to reward an offender or note a decrement in its performance, without sticking

to a 0 or 1 answer. This is of great help for both the case management staff and the offender that is able to acknowledge its situation. In the LS/CMI, the number of items that receive this supplementary rating are 13 out of 43.

The raw final score is then converted to a percentile according to suitable the normative group. The LSI-R divides offenders into three levels of risk (minimum, medium and maximum) whereas LS/CMI does it into five (very low, low, medium, high, very high) (Casey, et al., 2014). Both instruments provide normative data for 4 main groups: male inmates, male community offenders, female inmates and female community offenders but cut-off points are the same both either gender.

Another way of interpreting the assessment is to look at the individual scores of each subcomponent. The subcomponents with high marks indicate problematic areas that require intervention in the case plan while low marks highlight strengths. In the LS/CMI, this is done in Section 7 by transferring the subcomponent's scores to a table and rating them individually as very low, low, medium, high or very high.

4.4. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)

The Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) was initially created in 1998 by the Northpointe Institute for Public Management (rebranded as Equivant since 2017). The founders of Northpointe, Tim Brennan and Dave Wells, aimed to make a product that was better than the leading LS and the result was COMPAS. Its name points out the essence of values beneath this risk assessment tool: alternative options to mass incarceration, accurate risk assessment, public safety, institutional safety, fairness and racial equity in Criminal Justice decision support and community-based rehabilitative alternatives for risk offenders (Brennan & Dieterich, 2018).

Its design consists of an automated web-based software package that included risk and needs assessments to help Criminal Justice decision-makers at the different stages of the Criminal Justice System. Risk scales measure the likelihood of re-offending whereas needs scales capture information about offender's individual factors that are related to recidivism but can be changed for treatment interventions (e.g. education, substance abuse) (Casey, et al., 2014). The components of this instrument are a combination of static

—historical, such as the age at first arrest— and dynamic—criminogenic, such as employment status— risk factors and verified risk and need factors to serve as a guidance for correctional interventions and reduce the probability of recidivism (Brennan & Dieterich, 2018). Recidivism is understood in COMPAS as "a finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code" (Brennan, Dieterich, & Ehret, 2009) and scores are meant to forecast it for a period of two years after the COMPAS administration.

The areas to which COMPAS gives support cover the whole Criminal Justice System: from decisions involving risk, offender management, treatment, early release decisions, parole and reentry planning, and post-release supervision. Therefore, the institutions interested in its use range from pretrial release units, jails, prisons to probation and parole agencies or treatment providers. Lately its use has also been introduced into courts, not for determining sanctions but as an instrument to provide background information to aid in the preparation of presentence investigation reports (Brennan & Dieterich, 2018).

4.4.1. COMPAS Design

The design of COMPAS, apart from importing and integrating different MIS databases (criminal histories, current offense data, offender risk/needs, treatment goals, sentencing decisions, treatments/programs and outcomes monitoring), is characterized by the 4G characteristic features intended to move forward an Evidence-Based Practice (EBP) in Criminal Justice. Evidence-Based Practice traces its roots back to the early 1990s, when the emergence of evidence-based medicine occurred. Over the last decades, many other disciplines outside medicine have started to adopt this approach as well. The practice relies on scientific and mathematical evidence to find strong arguments to substantiate decision-making (Trinder & Reynolds, 2000).

Figure 17: Elements of transdisciplinary EBP model (Satterfield, et al., 2009)

The list of features that define 4G systems and outline COMPAS design are displayed as it follows (Satterfield, et al., 2009):

1. *Extended theoretical background*. COMPAS is a risk/need assessment tool designed relying on a comprehensive theory-based assessment approach. It is composed of key scales incorporated from the main theoretical frameworks in the Criminology field such as the ones explained below.

a. General Theory of Crime

The main assumption of this theory is that crime occurs in absence of self-control. In fact, people with a high level of self-control are usually concerned about long-term consequences whereas those with poor self-control are unaware (Gottfredson & Hirschi, 1990).

Therefore, low self-control is a relevant predictor of crime and a combination of low self-control and a criminal opportunity, will be the major cause of crime (Gottfredson & Hirschi, 1990).

b. **Criminal Opportunity Theories**

In this view, individuals make decisions rationally. Prior to an act, an underlying weighting of costs and benefits is made. If profits outweigh the risk, the misbehavior takes place. Thus, the most desirable reward equals a choice that requires little effort and risks (Cornish & Ronald, 1986). As typical example is that of the bike theft chain noticed by Van Dijik (1994): if a victim of a stolen bike sees the opportunity to steal another bike in order to replace it and steals it, then the owner of that bike would become a new victim and would steal from someone else, and so on.

Crime rates will then vary according to three variables, converging in time and space: 1) motivated offenders, 2) suitable targets and 3) the absence of guardianship. If the three elements take place in the same place and at the same time, crime occurs. Therefore, opportunity is a necessary but not sufficient condition for crime to happen; all crimes require opportunity but not every opportunity is followed by crime (Cohen & Felson, 1979; Felson, 1998; Cohen, Felson, & Land, 1980).

c. **Routine Activities Theory**

This theory is a variation of the prior one, known as the Routine Activities Theory or Opportunity Theory, emerged in the late 1970s, that studies how variations in routine activities can increase or decrease opportunities for crime (Cohen & Felson, 1979).

What daily activities create is a constant repetition of the three elements necessary for a crime in the same place and at the same time. If many of people's daily activities happen far from home and surrounded by strangers, the guardianship decreases. Also, the suitability of targets would depend on the value, size and attractiveness of the object or person of interest. Finally, a motivated offender is presumed to be a constant variable, present everywhere at any time. This third element is based on the assumption that every human being, given the right opportunity, that is, prone with suitable and unguarded targets, will commit an offense. The likeliness of a crime to occur will reflect how the three factors relate to each other in a social context (societies, cities, communities and local areas) (Cohen & Felson, 1979).

d. **Social Learning Theory**

Originally, Social Learning theory was meant to explain behaviors of individuals. However, some theorists broadened its scope and applied it too to explain crime rate

variations between social organizations (Akers, 1998; Sutherland, 1924; Wilson & Hermstein, 1985).

For that end, Sutherland and his individual-level theory of involvement in crime, viewed "differential social organization" as the explanation for those variations of crime rates in groups of different association. According to him, in a heterogeneous cultural social context, due to culture conflicts, crime-favorable messages will exceed unfavorable ones. Nevertheless, the theory requires further development as it lacks from an explanation for how structural arrangements and differential criminogenic learning produce higher crime rates in some social entities than others.

On the other hand, Wilson and Hermstein (1985) claim that learning of criminal behavior will be boosted or contained depending on the emphasis put on institutions commissioned to education, character formation and control. On his part, Akers (1998) sustains that changes in social structure and culture exert an influence on the reinforcements for misbehaviors and hence, on crime rates. Social structure variations can refer to alterations on the demographic composition, regional and geographical attributes and other changes that affect how social bodies and subcultures such as neighborhoods or families, organize internally (Tittle, Charles R., 2000).

e. Subculture Theory

According to Cohen (1955), crime is the result of the youth gathered in subcultures where deviant behaviors and dissident values and morals are predominant. The grounds of this theory rely on the conviction that young criminals are part of something bigger than them, a criminogenic subculture, defined as a subsystem of society that live under its own rules and beliefs in opposition to the majority class.

Subcultures are the result of nothing but the adjustment response and status conflicts of their members provoked by the inequality steaming from the existing class society (Cohen A. K., 1955).

f. Social Control Theory

Another theory to explain criminal behavior suggests that social and psychological integration in a group where crime is perceived as reputable and fear to exclusion from its negative responses can inhibit misconducts. In other words, social bonds (a job, a family, friends) can constrain individual criminal conducts.

The theory was first articulated by Durkheim (1952; 1933), but many others have developed his theme further on. For instance, Hirschi (1969) and his stipulation that individuals strongly bonded to conventional groups or institutions will have less probability of violating the law than those who are weakly engaged in society. The reason behind that is that the freedom they possess is more reduced (Horwitz, 1990). In accordance to his specification, freedom emanates from four different sources:

2. No preoccupation about other people, their thoughts or their responses to criminal behavior (Reiss, 1951)
3. Discrepancy between oneself moral beliefs and those of others (Hirschi, 1969)
4. Scant time and energy invested in the achievement of conventional goals that would be put at risk by a deviant behavior (Toby, 1957).
5. Disengagement from any traditional activity that requires time and energy (Hirschi, 1969).

In relation and contrary to this, the two main deterrence strategies involved in this theory are shaming and gossip (Tittle, Charles R., 2000).

g. Strain Theory

The main exemplar or this theory, Merton (1938), contend that strain is a factor that increases dramatically the likelihood of an individual to commit a crime. The relation between the set goals and the means available to achieve them will establish the level of strain that individuals who conform a society will suffer.

An improvement of this theory, the General Strain Theory (GST) was carried out by Robert Agnew on a macro-level perspective, who studied the effect of strain on individuals (1992) and on crime rates (1999). His line of thought implies that some social entities are more susceptible to crime than others because of their social, economic and cultural characteristics. These features result on strained people prone to commit crimes. The most consequential feature is the nearly inexistent informal social control they employ.

2. *Additional risk and need factors that provide content validity*. Extensive coverage of criminogenic factors to achieve a broader comprehensive assessment by adding theoretically relevant factors to the Central Eight criminogenic predictive factors (history of antisocial behavior; antisocial personality pattern; antisocial cognition;

antisocial associates; family and/or marital; school and/or work; leisure and/or recreation; and substance abuse) (Andrews, Bonta, & Wormith, 2006).

3. *Introduction of the strength-perspective of the Good Lives Model (GLM).* This (Ward & Brown, 2004) is an enhancement of the broad comprehensive assessment pursued because of the benefits it brings in for correctional assessments (Andrews, Bonta, & Wormith, 2006). The integration of this perspective is made by including strength and protective factors such as job and educational skills, history of successful employment, family bonds, social and emotional support, adequate finances, safe housing and so on which have demonstrated to be valuable to reduce risk and protect offenders from the impact of criminogenic needs (Brennan, Dieterich, & Ehret, 2009).
4. *More sophisticated statistics*. COMPAS uses advanced statistical models for its purposes of prediction and classification. This includes the introduction of Artificial Intelligence technology. By this way, COMPAS uses logistic regression, survival analysis and bootstrap classification methods (Brennan, Breitenbach, & Dieterich, 2008).
5. *Perfect integration with Criminal Justice databases*. In order to facilitate EBP for correctional agencies, COMPAS integrates the need or risk domains with the management information system (MIS) (Andrews, Bonta, & Wormith, 2006).

Additionally, COMPAS holds another two remarkable design features:

6. *Treatment-explanatory classification to address specific responsivity*. Specific responsivity is a persistent challenge when it comes to matching each offender with a treatment regime appropriate to them (Brennan T. , 2008a) and it is the least studied of the elements of the risk-need-responsivity mode (RNR)l[8] (Andrews, Bonta, & Wormith, 2006). Therefore, COMPAS addresses this problem by offering a person-centered assessment chart of decile scores for each risk and need scales and providing a treatment typology that integrates risk and

[8] The risk-need-responsivity (RNR) model is used in criminology to give recommendations of how inmates should be treated based on their risk factors, their needs and what it's the best environment for them to reduce their likelihood of re-offending (Andrews, Bonta, & Wormith, 2011)

need. Each typology presents several pathways to guide the treatment of the diverse type of offenders (Brennan, Dieterich, & Ehret, 2009).

7. *Gender-sensitive approach*. Unlike 2G and 3G methods that are built around male samples and applied identically with females (Brennan T. , 2008b), COMPAS uses separate female and male samples for gender-specific calibrations of all risk and need factors (Brennan, Dieterich, & Ehret, 2009)

4.4.2. COMPAS Model

COMPAS or Core COMPAS is, originally, a 137-question survey that covers different domains of information such as defendant's criminal history, environment or personality. A sample of the survey, from Wisconsin state, is provided in Appendix C. This state uses the tool at every stage of the Criminal Justice System once the individual has been convicted (Brennan & Dieterich, 2018).

From the traditional Burgess (1928) additive scaling where items are simply weighted and summed for a final score, COMPAS has evolved to use contemporary data analytics and ML methods for risk prediction models as well as explanatory/treatment offender classification methods.

For risk prediction models, LASSO[9] regression, logistic regression — to predict the probability of re-offending to happen depending on risk factors— and survival analysis — to predict the time until the next re-offending case—, were used to select and assign weight to the variables (Brennan, Breitenbach, & Dieterich, 2008).

For the explanatory/treatment offender classification, COMPAS differentiates between risk and needs scales to better understand explanatory factors, needs and intervention targets for case planning. Additionally, to develop an "Internal Prison Classification" to improve management and treatment of offenders, bootstrapped K-means were used (Brennan, Breitenbach, & Dieterich, 2008; Hofmans, Ceulemans, Steinley, & Van Mechelen, 2015).

---

[9] LASSO stands for Least Absolute Shrinkage and Selection Operator. This method performs variable selection and regularization to improve the prediction accuracy and interpretability of its statistical model (Tibshirani, 1996).

From the evolution of COMPAS, two different risk-predictive regression models emerged:

a) *COMPAS General Recidivism Risk scale (GRRS)*. This scale is a linear equation which originally comes from a LASSO regression model (Tibshirani, 1996) in a sample of pre-sentence investigation and probation intake cases in the state of New York, in 2002 (Brennan & Dieterich, 2018). The scale is composed by 28 questions concerning offender's criminal history, criminal environment, drug abuse and other early indicators of juvenile criminal tendency, which are validated predictors of recidivism (Desmarais & Singh, 2013; Gendreau, Goggin, & Little, 1996; Northpointe Inc., 2011). The items were selected through diagnostic modeling strategies such as LASSO and logistic regression, but updates are made constantly to improve the accuracy of the predictions as the final purpose of this scale is to, indeed, predict any new offense arrest in the next two years starting the day of the intake (Brennan & Dieterich, 2018).
b) *COMPAS Violent Recidivism Risk scale (VRRS)*. This regression model was created in 2006 to distinguish any misdemeanor or felonies from more violent crimes such as murder, manslaughter, rape, robbery and aggravated assault. The equation is also inferred from a sample of pre-sentence investigation and probation intake cases using survival modeling. Its intent is to forecast violent offenses, either misdemeanors or felonies, likely to happen in the next two years. Thus, the 28 items cover violent history, vocational or educational problems, history of non-compliance, the age at which the first arrest took place and the age at intake (Brennan & Dieterich, 2018). All of these are previously established as risk predictors of future violence (Gendreau, Goggin, & Little, 1996).

As COMPAS is designed to help in many stages of the Criminal Justice process, these two risk scales are usually used for pre-screening and triaging a case first and then, individuals who score high are assessed in more depth using additional scales or other COMPAS versions. As an example, a pre-trial case might include just the scales needed to support a release decision while at post sentence, additional scales would be added to help with supervision and treatment decisions (Northpointe Inc., 2012).

The scales that can be added to the main two in Core COMPAS, as well as the number of items to address each of them, are shown Table 4 (Northpointe Inc., 2011).

Table 4. COMPAS Core Needs and Risk Scales

Scale	Items
Criminal Involvement	4
Noncompliance History	5
Violence History	9
Current Violence	7
Criminal Associates	7
Substance Abuse	10
Financial Problems	5
Vocational/Education Problems	11
Family Crime	6
Social Environment	6
Leisure	5
Residential Instability	10
Social Adjustment	15
Socialization Failure	13
Criminal Opportunity	14
Social Isolation	8
Criminal Thinking	10
Criminal Personality	13

All of those scales are composed of items selected by instrument developers on the grounds of their connection to factors theoretically associated with criminal conducts and their statistically proven bond with those constructs (Northpointe Inc., 2011).

Furthermore, COMPAS has recently incorporated ML classifiers to attain higher reliability in offender risk prediction and classification, mainly in internal prison in order to assign each offender to their appropriate offender category in the Internal Prison Classification (Breitenbach, Dieterih, & Brennan, 2009; Brennan, Breitenbach, & Dieterich, 2008). Among others, the ML classifiers included are SVM and RF explained in Chapter 2.

4.4.3. COMPAS Versions

From that Core COMPAS, different specific versions have been developed to better adjust to the different target populations according to the age (Youth COMPAS), sex (Women's COMPAS) or stage in the Criminal Justice System (Reentry COMPAS). These are discussed next:

4.4.3.1.Youth COMPAS

The target population here are youngers between 12 and 17 years old who have already had a first contact with the justice system. Theoretical foundations were used for a deeper understanding of the youth and select the best explanatory and predictive scales for these range of age.

The number of scales is reduced from the 137 shown in Appendix C to 33, addressing family, friends, school, community, leisure activities and personality. It also includes both risk assessments and an explanatory treatment-oriented youth classification, which uses a person-centered ML pattern-seeking approach for boys and girls separately, replicated in several different states (Brennan, Breitenbach, & Dieterich, 2008; Brennan T. , 2008a; Brennan & Breitenbach, 2009).

4.4.3.2.Reentry COMPAS

This version is directed to detainees that have been imprisoned for a long-term, usually two or more years. Both risk assessments and explanatory treatment classifications for internal management (housing placement, rehabilitation programs and reentry planning) are included. A specification for reentry assessments was needed because criminological needs such as peer relations or family support, after a while

incarcerated, they can change. Moreover, the system has more information than when the compliance of the sentence started. That information comes from the behavior while imprisoned such as disciplinary history, prison adjustment, program attendance and performance, etc. (Brennan & Dieterich, 2018).

4.4.3.3.Women's COMPAS

This version is specifically designed for incarcerated women. It uses Reentry COMPAS gender-neutral scales plus a gender sensitive approach through gender-responsive test instruments (Van Voorhis, Bauman, Wright, & Salisbury, 2009) that include critical factors for women (trauma, self-efficacy, mental health, relationship problems, safety, parenting, etc.). It gives support for the person-centered ML approach for internal classification of women, needs assessment, case interpretation and treatment planning (Brennan & Dieterich, 2018).

4.4.4. Data Collection Method

Before the instrument's implementation begins, users are previously required to complete a two-day training course. Regarding the administration of this tool, it is flexible, and it can take from ten minutes to an hour depending on the COMPAS version being used but all of them have three components to gather information equally valid (Brennan, Breitenbach, & Dieterich, 2008). Approximately one-third of the information is collected from official records, one-third from self-report questions and one-third from an interview with the defendant/inmate (Blomberg, Bales, Mann, Meldrum, & Nedelec, 2010):

a) *Official Records*. Information about the individual is collected, preferably before the interview, through source documents such as state and national criminal histories, police reports, pre-sentence investigation reports, etc. This information can be introduced manually by the administrator of the test or automatically, if an appropriate transfer software is present.
b) *Interview*. A test administrator conducts the interview guided by a standardized script provided by the software. The questions to be asked will appear in the computer screen and answers will be entered in the program by the interviewer.

The type of interview conducted will depend on the skills of the interviewers and interviewee. For example, if the respondent is unable to read, que interviewee would read the questions out loud and enter the responses in the system.

In the training, the interviewer will acquire Active Listening skills such as paraphrasing, Motivational Interviewing skills such as role-playing exercises to provide guidance to the interview. Also, it ensures that the interviewer is familiarized with all the COMPAS items and is able to explain its meaning to any question that the interviewee does not understand. Moreover, training on how to deal with difficult situations and handle anxiety, suspicion or resistance is covered (Brennan & Dieterich, 2018).

c) *Self-Report*. This section is a pencil questionnaire to be completed by the individual on its own in a private protected environment. It focuses on personal items regarding life history, family, education, attitudes and so on. Any doubts concerning the interpretations of the questions may be asked to the assistance staff (Brennan & Dieterich, 2018).

Additionally, three hidden data validity tests are included to detect inconsistent or distorted answers aimed to give a good impression ("faked good") (Northpointe Inc., 2012).

- *The Defensiveness Test*. It is included to detect defensive or not self-revealing offenders by asking bizarre and unlikely items.
- *Random Responding Test*. It aims to find careless, inconsistent responding done on purpose.
- *The Inconsistency Test*. It compares an offender's predicted risk level with his general social history and profile of risk factor. If there is no reasonable coherence between them, it can mean there is false reporting, data missing or classification errors.

At the end of the interview, the software would alert the interviewer of any inconsistent or "faked good" answers for him to verify and the interview would not be closed until those responses are reviewed and corrected (Brennan & Dieterich, 2018).

4.4.5. Scoring

Both COMPAS risk scales (General and Violent) are integrated by items addressing offender's characteristics proven to predict recidivism by statistical models. For instance, the Violent Recidivism Risk Scale includes a History of Noncompliance Scale, a Vocational Education scale, a History of Violence Scale, the current age and age-at-first-arrest (Northpointe Inc., 2011).

Every item has an assigned weight (*w*) depending on the strength or potential to provoke a person's recidivism. Raw scores of each item is multiplied by its *w* to transformed them into weighted items (deciles) that would then be add together to calculate the final score (Northpointe Inc., 2011). The equation is the following:

Violent Recidivism Risk Score = (age∗−*w*) + (age-at-first-arrest∗−*w)* + (history of violence∗ *w*) + (vocation education ∗ *w*) + (history of noncompliance ∗ *w*)

The final scoring guidelines vary depending on the site where COMPAS is being applied. Each location establishes its own cut points for decile levels, computes probability scores or uses simple low, medium or high score designations, based on a large norming sample distribution (Brennan & Dieterich, 2018). Low, medium and high labels are typically used when the number of local cases is not large enough to infer stable normative scores. For referencing scale scores, COMPAS includes scale distributions of eight normative groups: (1) male prison/parole, (2) male jail, (3) male probation, (4) male composite, (5) female prison/parole, (6) female jail, (7) female probation and (8) female composite (Northpointe Inc., 2012).

The ranges are divided using decile scores intervals for which 1-4 is low, 5-7 is medium and 8-10 is high. However, deciding the cutting points requires discussion with local administrators and policymakers although COMPAS can provide recommendations (Brennan & Dieterich, 2018).

Anyway, when using, for instance, the Violent Recidivism Risk Scale, results can speak out against expectations. There are features such as a young age, absence of a job and an early age-at-first-arrest and a history of supervision failure, that could make a person get a medium or high score, even if that person never had a violent offense arrest.

On the contrary, a person who scores high (D10) on History of Violence (e.g. prior assault, domestic violence, property and weapon offenses); medium (D6) on vocational education issues; but is 50 years old and has a late age-at-first-arrest (e.g. 35) with no history of noncompliance (D1), can get a low Violent Recidivism Risk score (D3). Being age one of the best predictors of violent recidivism, it carries a lot of weight so, if the person's age was 25 and the first arrest took place when he was 16 years old instead, the score would change to D8 (High) (Northpointe Inc., 2011).

For the final assessment report of risk and needs of the offender, COMPAS uses an ad-hoc report generator that processes the results and forms statistical summaries and trend charts. The document does not provide a single final score but separate rates for violence, recidivism, failure to appear and community failure and a written description of which are the offender's risk and need scale results, a statement from the interviewer and a recommendation of treatments to be applied. Current charge and criminal history information are also presented. In any case, COMPAS designers developed the tool expecting disagreements with the result in 10% of the cases due to exceptional circumstances that could mitigate or aggravate the situation of the individual. In such circumstances, the professionals conducting the test are encouraged to give their personal opinion and override the scale results (Casey, et al., 2014).

Finally, must be noted that all the report results and the COMPAS database is compatible with PDF, Word, Excel and other statistical programs for further customized studies (Casey, et al., 2014). An example of how the final report showing the results would look can be found in Appendix D.

5. CRITICAL ANALYSIS OF COMPAS AS A CRIMINAL JUSTICE DECISION MAKING TOOL

As an exemplification of a real-world application of AI in the Criminal Justice System, the COMPAS risk assessment tool has been described in Chapter 5. Furthermore, as the dissertation has shown in Chapter 3, the debate concerning the legal, ethical and statistical use in judicial decision-making of these computerized risk assessments algorithms is now commonplace. Therefore, a deep analysis of the instrument regarding the FATE components of it is made in this Chapter.

5.1.Fairness of COMPAS

The tool has been under various internal and external validation studies in different locations, Criminal Justice agencies and populations such as the Michigan Department of Corrections (MDOC), New York State Office of Probation and Correctional Alternatives (OCPA), California Department of Corrections and Rehabilitation (CDCR) and Broward County Sheriff's Department (Dieterich, Brennan, & Oliver, 2011; Dieterich, Oliver, & Brennan, 2011; Brennan, Dieterich, & Ehret, 2009; Lansing, 2012; Farabee, Zhang, Roberts, & Yang, 2010; Flores, Bechtel, & Lowenkamp, 2016). Apart from the validity, the non-discriminatory use and results of the instrument have been examined. Being Fairness ultimately defined by the social context, which is, in this case, the Criminal Justice setting, the tool should seek for an accurate prediction of an offender's future criminal behavior by performing an equal application of the tool regardless of the age, gender, sexual orientation, race or ethnicity of the offender.

First of all, for any instrument to be considered fair, it has to actually do what it is supposed to do, and do it so, accurately and consistently, evaluated by the validity and reliability of the tool. Secondly, it has to be free of any type of bias. And thirdly and additionally for being COMPAS an ML system, it has to overcome the Abstraction Traps.

5.1.1. Validity

To begin with, Validity addresses the question of whether an instrument is accurate or not in its results. The accuracy of COMPAS has been analyzed in those studies by its Criterion, Content and Construct validity. Nonetheless, it is worth noting that risk assessments are not meant to measure a psychological construct or an internal disposition

but to estimate the probability of a criminal behavior. Consequently, statistical criteria applicable for psychological test such as the factorial validity of the items, the internal consistency, the test-retest and the construct validity, has a lower relevance when assessing risk in Criminal Justice (Dutton & Kropp, 2000).

5.1.1.1. Criterion Validity

The Criterion or Concrete Validity examines if there is a relationship between the outcome and the scales of COMPAS. It is commonly divided in Predictive and Concurrent Validity. Predictive Validity is the likeliness of the instrument to predict well by comparing the scores of COMPAS with the real future behavior of the offenders while Concurrent Validity compares them with other measurements already established as valid, like the Area Under the Curve (AUC). The AUC is commonly used to measure discrimination ability, fundamental in risk assessment instruments as they depend largely on its capacity to discriminate successfully between recidivists and non-recidivists. For binary outcomes like such, the AUC can be understood as the probability that if two participants are selected randomly, a recidivist will achieve a higher risk score than a non-recidivist.

The predictive validity of the COMPAS General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) have been validated in diverse geographical areas, diverse Criminal Justice agencies, and diverse gender and race categories. Table 5 presents the AUC results from studies that focused on the functioning of the tool by the MDOC, New York OCPA, CDCR and Broward Jail (Brennan, Dieterich, & Ehret, 2009; Dieterich, Brennan, & Oliver, 2011; Dieterich, Oliver, & Brennan, 2011; Farabee, Zhang, Roberts, & Yang, 2010; Flores, Bechtel, & Lowenkamp, 2016; Lansing, 2012).

The benchmarks most commonly used for predictive accuracy establish that AUC scores above .56, .64, and .71, correspond to “small,” “medium,” and “large” effect sizes, respectively (Rice & Harris, 1995). Most of the predictive scores in Table 5 indicate a large effect size in the mentioned studies.

Table 5. Average AUCs of Six Studies made on the Predictive Validity of COMPAS GRRS and VRRS (Brennan & Dieterich, 2018).

Study	N	Year	AUC
NY Probation[10]	(*n*=2,328)	2009	.700
NY Probation[11]	(*n*=13,993)	2012	.710
MDOC Reentry[12]	(*n*=25,347)	2011	.710
MDOC Probation[13]	(*n*=21,101)	2011	.710
CDCR Reentry[14]	(*n*=25,009)	2010	.680
Broward Jail[15]	(*n*=6,172)	2016	.710

5.1.1.2. Content Validity

The Content Validity manifests if the of a risk/need assessment tool is considering all the relevant factors to accomplish its purpose successfully. A low Content Validity value will indicate that the instrument is simplifying the coverage of factors that comprise the model. For example, 2G risk assessments include a very small number of variables (5-15), typically static, and only focus on risk, resulting in poor Content Validity and making them unable to orientate treatment, internal management, reentry planning or apply RNR principles.

[10] (Brennan, Dieterich, & Ehret, 2009).

[11] (Lansing, 2012).

[12] (Dieterich, Brennan, & Oliver, 2011).

[13] (Dieterich, Oliver, & Brennan, 2011).

[14] (Farabee, Zhang, Roberts, & Yang, 2010).

[15] (Flores, Bechtel, & Lowenkamp, 2016).

The characteristics that made COMPAS a 4G Correctional Practice tool were the risk and needs scales added to the Central Eight criminogenic predictive factors previously displayed in Table 4 and the theory-based selection of them using the major criminological theories, explained in Chapter 4. These two features were included to provide Content Validity and make COMPAS an instrument with efficient risk models and comprehensive need factors competent to provide a detailed profile of each participant, help with the internal classifications of prisoners and support the implementation of RNR principles (Brennan & Dieterich, 2018).

5.1.1.3. Construct Validity

Construct Validity deals with whether the assessment is measuring what is intended to measure and if the outcomes, such as recidivism, are correlated with the scales measured. But for that, it has to be established the factorial validity first, that is, if each scale is measuring a well-defined construct of a single dimension. All COMPAS subscales were subject to this factorial analysis and results proved that they were all unidimensional (Brennan & Oliver, 2002).

Another way of demonstrating Construct Validity is through Convergent Validity by examining correlations between two scales that are meant to measure the same or a similar construct. For example, drug abuse is typically measured by the Substance Abuse Subtle Screening Inventory (SASSI), so the COMPAS scale meant to measure it was compared to SASSI and a positive correlation of .44 was found (Brennan & Dieterich, 2018). To demonstrate the Convergent Validity of overall COMPAS, its scales were correlated with the LSI-R subscales due to its similarity and significant correlations were found between pairs of scales. The results found a correlation of .64 for Criminal Involvement; .48 for Vocation/Education; .39 for Financial Problems; and .57 for Residential Stability. Other pairs had lower correlation values, but it was clearly determined that they were assessing distinct aspects of the same construct (Farabee, Zhang, Roberts, & Yang, 2010).

5.1.2. Reliability

The reliability of the tool measures the trustworthiness of the instrument and its consistent performance over time. It can be measured by different means, as the ones that follows: Internal Consistency, Test-Retest and Inter-Rater Reliability.

5.1.2.1. Internal Consistency Reliability

The internal consistency of COMPAS scales and subscales is examined by conventional item analysis and factor analysis for scale improvement as well as determination of the factorial validity (Brennan & Dieterich, 2018). To evaluate the internal consistency of all scales and subscales, the Cronbach's alpha coefficient is used. It is established that alphas of .70 and above denote satisfactory internal consistency (Nunnally & Bernstein, 1994). As shown in Table 6 below, most of the Core COMPAS subscales provide alphas above .70 (Brennan & Dieterich, 2018).

5.1.2.2. Test-Retest Reliability

Another reliability measure is verified by replicating the test over time. The scales that constitute COMPAS have been proven to be in a good or even excellent range of test-retest reliability (Farabee, Zhang, Roberts, & Yang, 2010). In this research, Farabee et al. (2010) compared the COMPAS Core scales to the LSI-R subscales described in Chapter 4 and, in fact, it turned out the COMPAS scales had a higher test-retest reliability than LSI-R, with correlations between .70 to 1.00 and an average correlation above .80, versus .64 of the LSI-R scales.

5.1.2.3. Inter-rater reliability

Inter-rater reliability refers to the level of agreement reached among raters. In the risk assessment field, reaching a good degree of inter-reliability is not common (Desmarais & Singh, 2013) and in particular when the instruments are being used in a context of an overwhelmed Criminal Justice System. The challenges to be faced in this environment are diverse, ranging from workers lack of enthusiasm due to high pressures, employee replacements, unqualified personnel participating in the project, and logistical difficulties that arise when the same case needs from multiple interviewers for its assessment (Brennan & Dieterich, 2018).

In order to improve COMPAS reliability, the instrument is enhanced with the mentioned ML technology, support hyperlinks to solve interviewers' questions immediately and a periodic revision of the problematic questions. (Brennan & Dieterich, 2018).

Table 6. Cronbach Alpha of COMPAS Core Scales.

Scale	Alpha
Criminal Involvement	.75
Noncompliance History	.65
Violence History	.52
Current Violence	.53
Criminal Associates	.71
Substance Abuse	.76
Financial Problems	.70
Vocational/Education Problems	.71
Family Crime	.62
Social Environment	.81
Leisure	.86
Residential Instability	.71
Social Adjustment	.54
Socialization Failure	.69
Criminal Opportunity	.66
Social Isolation	.83
Criminal Thinking	.80
Criminal Personality	.70

5.1.3. Gender and Racial Biases

Addressing the issue of a non-discriminatory application of the model, Northpointe has tried to palliate the inequalities that an offender might experience in the COMPAS assessment process by, first, designing specific versions of the tool for those subpopulations with different needs (Youth COMPAS, Reentry COMPAS and Women's COMPAS); and secondly, avoiding the inclusion of any items concerning racial, gender, religious or national origin although their exclusion does not imply the elimination of disparities (Casey, et al., 2014; Dressel & Farid, 2018). Despite this, COMPAS has been highly criticized and accused of suffering from gender, and mostly, racial biases. The allegations involved finding black defendants more likely to be categorized at a higher risk in opposition to White defendants, which were more often categorized as low risk defendants (Angwin, Larson, Mattu, & Kirchner, 2016).

The results of various studies have proven the Predictive Validity of COMPAS among gender and racial groups. The outcome of the study performed by Brennan et al. (2009) confirmed the equity among Black and White males in the COMPAS GRRS. The AUCs for any arrest, scores were .67 for Blacks and .89 for Whites, and for felony arrests, .73 and .71, respectively.

Notwithstanding, the much talked-about polemics on COMPAS risk score involved an alleged racial bias against African Americans in a study made in Broward County, Florida. ProPublica's article stated that the equality of error rates across groups was violated not only by COMPAS, but by all risk assessments used in Criminal Justice, making all of them biased against blacks (Angwin, Larson, Mattu, & Kirchner, 2016). However, posterior reports developed using the same Broward County data proved otherwise, confirming that the test accuracy was leveled between Blacks and Whites and found some inconsistencies in their procedures (Dieterich, Mendoza, & Brennan, 2016; Flores, Bechtel, & Lowenkamp, 2016). Both studies demonstrated equal predictive results for Black and White offenders in COMPASS GRRS and VRRS. The study of Flores et al. reached an AUC of .70 for Blacks and .69 for Whites and .71 globally for the GRRS forecasting any arrest within two years and .70, .68 and .71 for the VRRS predicting violent arrest within two years. On the other hand, the study of Dieterich et al. obtained analogous findings. The results suggest an adequate predictive validity of the

GRRS and very similar ones for men and women, as well as for White, Black, and Hispanic defendants.

Another independent study done by Farabee et al. (2010) measured Predictive Validity across men and women, and Whites, Blacks and Hispanics in the CDCR with a sample of more than 20,000 prisoners. The inmates were released on parole and their conduct was studied for the next two years. AUCs for the outcomes of "any arrest" and confirms that AUCs, most of them above .70, did not differ for men, women, White, Black, and Hispanics.

AUCs for a parallel study in the MDOC on COMPAS Reentry made on a large sample resulted in large effect sizes. A three years follow-up was done since the prisoners were released. As provided in the other studies, the results were presented for the overall sample and for the men, women, and White, Black, and Hispanic groups. The only difference was a .71 for Blacks versus .78 for Hispanics (Dieterich, Brennan, & Oliver, 2011).

In summary, the allegations made by ProPublica about COMPAS or any other risk assessment, have been proven to have many flaws and faulty statistics. The study of Flores et al. (2016) concluded that the reason why Blacks obtained higher scores on COMPAS was not a racial bias but an actual higher recidivism rate. Not only that but many other researches have well-founded the equal predictive validity of COMPAS, determining it is racially unbiased and considering it a gender-responsive instrument.

5.1.4. Abstraction Traps

Due to the ML techniques used in the operating of COMPAS, a look at the Fairness-aware ML field and its Abstraction Traps appears necessary due to their capacity to hamper Fairness.

The first question to be asked is whether introducing COMPAS in the system brings true benefits to it. Considering the overwhelmed U.S. Criminal Justice System and the concerning mass incarceration problem from which the country suffers, it can be stated that definitely a change was needed. With the application of the COMPAS software, judges are relieved of great part of their workload that were not able to perform properly. Judges are expected to review carefully the files in every case to make a determination,

but the reality is that when they have that many cases, an in-depth examination gets out of their hands. In such situations, an intelligent system which gives them a risk prediction of a defendant based on that same information they were supposed to study seems like an ideal solution. Not only could they make the prediction faster, but also more objectively as biases introduced by, for example, prejudices, fatigue, personal beliefs, would not be part of the final outcome because the human intervention required is minimum. Based on this, COMPAS actually helps addressing those problems, but other risk assessments such as the LS instrument do it so as well. Hence, if the *Solutionism Trap* can be ruled out is not clear.

Secondly, in social systems where stakes are high, as in the Criminal Justice, the first step should consist on examining what the consequences of its implementation will be before actually doing so to avoid the *Ripple Effect Trap* (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). COMPAS designers anticipated that the tool could affect the response of judges, prosecutors, officers and more (behavior, perception of his job, etc.) and clearly established the auxiliary function of the instrument to them. With regard to arrestees, the creators also foresaw that many of them could presume that some of the questions addressed by the questionnaire aim to indicate the degree of criminal risk of themselves —for example, how many of their friends have been arrested—so they would probably answer deceptively. The assumption of honest responses disregarding these reactive behaviors could alter the algorithm. To fix this problem, they included the Defensiveness Test, the Random Responding Test and the Inconsistency Test. In any case, the introduction of a novelty in a pre-existing setting can always detonate unexpected responses such as fading away other social premises such as rehabilitation when focusing the attention in the risk assessment.

Moving forward the *Formalism Trap,* it requires consideration of how to contemplate Fairness and Justice in the system. Bias introduced by individuals that participate in the Criminal Justice System has to be taken into account, as well as how social factors affect criminal trends to understand the concept of Fairness in assessing someone's risk. A widely accepted mathematical definition of Fairness has not been found yet, neither has Northpointe-Equivant included one in the COMPAS system. Nevertheless, the procedural Fairness of the instrument has been validated by the previously described studies (Brennan, Dieterich, & Ehret, 2009; Dieterich, Brennan, & Oliver, 2011; Dieterich,

Oliver, & Brennan, 2011; Farabee, Zhang, Roberts, & Yang, 2010; Flores, Bechtel, & Lowenkamp, 2016; Lansing, 2012).

The *Formalism Trap* also entails the desire for contestability. Fairness can be procedural, contextual and contestable. Currently, a score given by a risk assessment cannot be fought back if the defendant is in disagreement. This desire could be satisfied if interpretable models are established and a formal procedure to allow the defense attorney to replicate and discuss the results with a judge is created, in case of the appreciation of data errors or an exceptional situation of the individual (Selbst & Powles, 2017; Wexler, 2018). This, however, falls out of Northpointe-Equivant hands. Equally important is to make public the value judgements used to build the assessment in case that if changes in the social context take place, the instrument can be readjusted at its best. Over and above, this is indeed their choice and despite the proprietary nature of the company, most of the design and components of their model are available to the public.

Consecutively, The *Portability Trap* brings out that context is essential and so, an algorithm designed to predict good employees would not be appropriate for a risk assessment. Not only has the algorithm to be different if the domain changes but also when applying it within the same one if there a minimum variation. If the information about the specific social setting is insufficient or not concrete enough, an approximated model can make things even worse (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). For this reason, COMPAS has developed the Youth COMPAS, Reentry COMPAS and Women's COMPAS versions and has referencing data on eight different normative groups (male and female for prison/parole, jail, probation or composite) for establishing the cutting points and the scoring guidelines differ according to the jurisdiction and location where is utilized (Brennan & Dieterich, 2018).

Finally, the *Framing Trap* reveals that a sociotechnical frame has to be the basis of the system. Populations commonly underrepresented should be adequately included in the model as well, instead of making suppositions about their needs (Eubanks, 2018). In the context where COMPAS works, this applies to the poor, the minority races such as Hispanics or African Americans, women and the youth. It has to be understood how politics work and which are the powerful social groups that keep the needs of those subpopulations buried. For that, partnerships with advocacy organizations, social scientists or the troubling populations itself can be created. Effort should be made by the community to avoid closure and maintain interpretive flexibility until the technologies

address a wider array of concern from various social groups (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). Northpointe itself cannot be heavily criticized in this matter as it has gathered sufficient data collection to be able to design the specific versions of COMPAS to modification and adapt in the best way they could to the needs of a heterogeneous population of offenders.

5.2.Accountability of COMPAS

COMPAS is used to help in high-stake decision-making where fundamental rights such as freedom are put at risk. Decisions with such significant effect require a higher burden of accountability. More and more risk assessments are being introduced into the mainstream of the Criminal Justice System and there is an increased need for policy makers to regulate this practice. The question of who should hold accountable for the false positives and false negatives of COMPAS predictions and the consequent mistaken decisions taken based on them is unresolved.

COMPAS and other risk assessments depend upon Politics and the prevalent public policies and thus, the tools can be designated to lighten the prison population in a jurisdiction while locking up offenders in another one. The mission assigned to the software can even change over time. A main criticism related to policies is the operationalization of recidivism with rearrests, as it does not directly indicate reoffending and underlies structural problems. For example, racial inequality in the enforcement of the law added to policies that target the Black community, lead to the over-criminalization of their members and create a baseline figure used in crime statistics. Therefore, racial disparities are found in rearrests, being Blacks, for example, about 4 times more likely to be arrested than Whites. The same happens with employment, used as a predictive risk factor, when the unemployment rate for Black people have been higher for decades in the United States (Dressel & Farid, 2018).

On the other hand, being Equivant a for-profit company working for a public institution makes the accountability of instrument even more ambiguous. The Criminal Justice System buys the COMPAS software to help making more objective and accurate resolutions. The personnel working for the institution follows the determination of using COMPAS to enhance and facilitate their jobs. Thus, there is a reliance on the COMPAS predictions to be accurate but obviously, an expertise on AI systems works is not among the abilities of every single worker. The responsibility of the accurate functioning of

COMPAS goes down partly to its designers, who are the first interested parties in not committing any mistakes as they can have a negative impact on their profits. However, fairness and other relevant aspects for the Criminal Justice System are not a main concern for them and that commercial interest should prevent the Criminal Justice System to blindly trust on the tool. In that sense, both the Criminal Justice Institution and Equivant would be partially accountable: the former according to the pre-factum accountability perspective and the later, as the post-factum blamable agent.

Moreover, clearly, one of the biggest advantages of this AI systems is the relief of responsibilities for humans who use it, but the delegation of responsibility is not total. COMPAS is an intelligent assistant used to make predictions that can help in decision-making. Nevertheless, the scores it provides are not a binding and the final decision is taken by a human. In fact, judges are not supposed to be given longer sentences based on a higher result. COMPAS was primarily designed to reduce crime so that the score would determine a defendant's eligibility for probation or treatment programs, not their sentence. Nevertheless, the use of scores has been wrongfully interpreted and used in several sentencing decisions (Angwin, Larson, Mattu, & Kirchner, 2016). Also, even though the COMPAS software is totally automated, it requires from a human being to transfer the raw data collected by the stuff who conducted the test, leading to the possibility of errors in the introduction of the data in the computer. For that reasons, humans should also take an accountable part.

When relying on its risk assessment algorithms, the risk of leaving important decisions unaccountable is taken. Despite the mistakes and biases possibly made by humans, at least a certain degree of rationalization and accountability can be demanded to them, whereas who responds for COMPAS errors is not established yet. It is essential to find a way to hold the decisions accountable considering the impact that the use of the COMPAS software can have in the lives and well-being of criminal defendants but still, further research is needed on what people want and require for it.

5.3.Transparency of COMPAS

One way of achieving better accountability is prompting regulations to make the system more transparent (Binns, et al., 2018). In Europe, for instance, due to the increasing use of these predictive systems by both public and private bodies, The European Union's General Data Protection Regulation (GDPR) was created to compel

organizations to give an explanation of the logic behind their predictions. The purpose of these regulations is to find a solution to the *interpretability problem* by clearing the *black box*. If the algorithmic decision-making systems are explained, the affected individuals and stakeholders are able to assess the fairness of their processes and outcomes (Binns, et al., 2018).

Due to the private nature of the Equivant company and their for-profit interests, many details of how their algorithms work are not revealed to the public. Despite their reservations, a lot of information is published, such as the scoring guidelines, instructions and limitations of the instrument and COMPAS has provided insight information under request for the development of independent studies. Even though certain degree of opacity is necessary for the model to be useful, profitable and protective against manipulative strategies, is that confidential information what puts transparency of COMPAS on the line. How do their algorithms really work? How is data processed by the system? What are the weights of each items? How does the software provide the outputs? COMPAS still mainly uses regression models but has incorporated new ML techniques such as SVM, which as explained before in Chapter 2, do not provide a clear explanation of the decision-making process and cause transparency issues. In this respect, more simple risk assessments like LS or less complex algorithms are often preferred. Nevertheless, new explanation ways for ML models are being investigated to overcome the challenges they present and thus, be able to provide "meaningful information" about how the predictions are made (Binns, et al., 2018).

Further, predictive analytics like the examined COMPAS and LS risk assessments use information to forecast. In the case of COMPAS, it is a complex model based on AI systems with ML and for it to be effective, it is essential to constantly introduce fresh data to bring the model up to date. But being data such a powerful part in the success of an AI system, has also its risks. We can think the more the merrier, but quantity cannot put aside quality (Taulli, 2019). On the basis that AI systems are built and learn from large amounts of data gathered from the past, it can be a reflection of the social, historical and political conditions in which it was created. According to the AI Now Institute at New York University, a research center committed to study the social implications of Artificial Intelligence, not only for this reason but together with many others, if the grounds from which the system learns is skewed, the algorithm can produce inaccurate and unfair results leading to inequality and bias (Kak & Richardson, 2020). COMPAS

currently deals with data provided from a Criminal Justice System which suffers from underlying inequalities in the application of the law. Taking this into consideration, COMPAS or any other ML system working with historically biased data inputted, will continue to produce biased data, as the algorithms will learn from those past references for its application in future cases.

Additionally to this problem is the fact that the model works with personal data. In this regard, privacy issues can be entailed. Consent of the individuals subject to study are necessary to deploy the tool but keeping safe and anonymous the information gathered for purposes other than the original is critical and responsibility of Equivant. Limited access to the results to just the people working with it should be guaranteed to prevent a wrongful use of it.

In summary, the Transparency problems that COMPAS might encounter are both internal and external. On the one hand, the opacity with regard to the inner methods of the tool and the *black box* component of the instrument related, among others, to the highly criticized use of SVM, makes it difficult for the public to get an explanation behind the outcomes. On the other hand, COMPAS provides governments with ways and means for collecting, tracking and analyzing large amounts of data without people's realization. This, apart from dealing with privacy issues, raises awareness about the risk of creating a feedback loop that prolongs and strengthens institutional bias in policing.

5.4.Ethics of COMPAS

The use of COMPAS in such an important field as it is the Criminal Justice System raises different dilemmas related to basic human rights and Ethics. To begin with, COMPAS avoids the use of protected characteristics such as race or gender as predictive variables for the simple reason that they are outside the individual's control. The tool focuses on dynamic factors that can be subject to change. However, gender, race and age are part of the personal informative box found at the beginning of any questionnaire. So even though it is not used for prediction, gender, for instance, is important to decide whether the Women's COMPAS versions should be applied to better address the different needs of males and females. Similarly occurs with age, which is also considered for legal concerns, as an offender under 18 years old is separated from adult prisons.

As stated before, COMPAS was originally designed as a support tool in pre-trial and bail decisions and not as a decisive tool for sentencing. With the years, COMPAS has been adapting to the necessities of the Criminal Justice System but always as an auxiliary instrument. In general, risk assessments are meant to be used as recommendations that judges can then take into account or not, leaving the ultimate response to the judge. If the model does not include how a judge needs to use the risk assessment score for the decision, fairness will also be in endangered but, in this case, COMPAS clearly states its supportive functionality.

Before the risk assessments existed, judges listened to the opinions of the probation officers and other professionals and looked at the evidence in order to make their own determinations. With risk assessments like COMPAS, they attend to a mathematically calculated score to make their decisions. The response of the judge is unknown though. Several studies suggest that some COMPAS users over-rely on the instrument scores whereas others have a complete distrust to this kind of intelligent systems despite their proven accuracy. An increased transparency on algorithmic decisions could improve trust (Binns, et al., 2018). However, too much reliance on the given score can result in the appearance of the *Automation Bias* (Christin, 2017) and a sense of disempowerment for their own role as humans.

The automation of the system makes it also hard to fight back the outcomes. The chance of negotiation available when interacting with a human disappears with the use of COMPAS, perceiving it as dehumanizing and impersonal for the subject of the decision (Binns, et al., 2018). If more transparent explanations on how the tool functions are given to the individuals in order to understand how the algorithm works, it jeopardizes privacy but it can help in the sense that if they know where their scores come from, they can improve their behavior to change it.

Finally, another ethical concern regarding COMPAS is its lack of moral judgement. There is no really an opportunity to discuss the morality of the instrument if the system is just doing what it was intended to do. At the end of the day, it is primarily designed to seek for efficiency, not other goals like Fairness (Binns, et al., 2018). Certain margin of error is though considered with the override policy that experts are encouraged to use in both LS and COMPAS when aggravating or mitigating circumstances appear. But, in comparison to LS/CMI, COMPAS does not have any section specifically destined to

address contingencies and individual circumstances that can deviate the standard outcome based on generalization and statistical inference. In any case, people are reluctant to impute morality to automated systems so the model can still be seen as statistically fair (Binns, et al., 2018).

6. CONCLUSION

The use of AI is more than settled these days, I would venture to say, in almost every aspect of our lives. The Criminal Justice System is no exception. Any modern society should aspire not just to detect crimes and cast offenders into prison but also, to prevent offenses from happening in the first place. Crime prevention has become a priority over the years and AI has brought many different possibilities to help in this matter by providing interesting tools for the prevention, investigation and judicial determination of crimes. But as a world that straddles between several disciplines, how to make the best use of these intelligent machines is complicated. AI is also a prevailing topic in Criminal Justice discussions regarding good practices, equity and potential distortions because of, mainly, gender and racial biases underlying the Criminal and Police systems. As a consequence, this AI reality has to be faced as a whole, understanding how it works, the benefits as well as the challenges and limitations that its application presents.

In the Criminal Justice setting, to deal with the dilemma of whether AI should or should not be applied, or to what extent, two risk assessment tools have been described in Chapter 4. The U.S. is a pioneer country implementing AI in the field so an instrument with an AI component and another without it were selected for examination: COMPAS, and the traditional LS tool, considered the gold standard of correctional risk and need assessment. After the analysis of COMPAS and its comparison with LS, the following conclusions were drawn, which might be of interest to public and private companies developing risk assessment tools, people working for the Criminal Justice System and, last but not least, Governments.

At first sight, these automated assessment technologies provide nothing but benefits to the courts because they claim to be able to determine the risk of future criminal behaviors without requiring a judge examining the circumstances of the case and saving the costs of a full forensic evaluation. Therefore, the current slow and overburdened U.S. Criminal Justice System will be relieved as the frequency of the decision-making expediency will increase. Moreover, they reduce the possibility of biased determinations based on prejudices and personal perceptions, consequently achieving higher levels of Fairness in the application of the law.

However, the truth is that drawbacks also exist. As for now, AI still requires from a minimum human involvement to operate. This means that imperfections will be present

in the models as every society and person in this world is skewed in its own way. In other words, it cannot be expected from a system grounded in human interventions to provide perfect solutions and decisions when they are non-existent in the first place. A judge's ruling will never be completely unprejudiced and impartial neither will an artificial model designed by humans. Thus, the ultimate goal should not be to design machines better than the ideal man, but instead, better than a real man.

As for now, the resemblance to a human brain capacity of reasoning is far from being the same. Besides its capability to surpass humans in many fields due to their larger mathematical processing, in comparison to individuals, these machines lack the personal flexibility, context-relevant judgements, empathy, as well as the ability to develop complex moral judgements. They are programmed to perform concrete functions and nothing else so, how can we question their morality if they just do what they were intended to do? This demonstrates the euphemism surrounding the AI field because, even though some believe that artificial systems capable of assimilating all human mental processes, or even generate new ones (Holland, 2004), will be developed in the near future, several others think strong AI might never be achieved. Researchers need to keep investigating AI models in general, and specifically, computer risk algorithms, to find evidence in order to, more than support or reject these tools, put an end to the controversy of whether AI models are intended to operate on its own or as a support tools.

In this regard, one possible solution to the problems that have been presented, and concretely, for COMPAS, could be the incorporation of professional judgment combined with the model, using the later more as an auxiliary tool more than as an independent and sufficient instrument itself. Along with the improvement of the Transparency of the procedures behind the outcomes, this will address the issue of people in complete opposition with the strict statistical model of judicial decision-making and overcome the perception of risk assessment algorithms as "reducing a human being to a percentage" (Binns, et al., 2018).The human factor is many times overlooked by trying to transform qualitatively features into quantitative items, which is not very realistic. Additionally, more emphasis should be put on promoting proactive auditing of the systems to seek problems that go beyond the efficiency and efficacy of the instruments, which, in combination with a better access to individual-level demographics would facilitate finding the origin of one-off but also systematic biases.

Nevertheless, regarding our area of interest, the use of AI in detecting or predicting crimes or an individual's risk of recidivism is a promising field that requires far more investigation, as well as education. There is a real need for educating judicial decision-makers about the strengths and weaknesses of these tools because there is still a huge lack of information aggravated by a massive sensationalism provoked by fake news that can result in very unfavorable consequences for many people. Politicians, judges, lawyers, officers and any person related to the Criminal Justice System should be fully informed of how AI works in order to prevent misinformation and guarantee Fairness. In that sense, it is also important the role that private companies play in this process. It cannot be ignored the fact that their main goal is to gain benefit and thus, they are interested in getting a portion of the total spend from governments in this segment. Therefore, companies that provide innovative and promising ML tools will take advantage of the potentially profitable market area of crime prevention the best way they can with a general disregard to FATE practices. On top of that, Transparency will be sacrificed many times due to proprietary rights.

There is a lot of work ahead to achieve the creation of Fair, Accountable, Transparent and Ethic AI instruments but it is extremely important in a setting such as the Criminal Justice System, where the stakes are as high as involving the strip of one of the most important values for the human being: the freedom. But for that purpose, the reality is that a structural change is necessary beforehand. Even when a tool such as COMPAS is demonstrated to have no gender or racial bias, how can it be equally fair when data is systematically skewed? Systematic bias within the Criminal Justice System can alter the recidivism rates (Olver, Stockdale, & Wormith, 2014). In the United States, the fact that the prison population is disproportionately black makes racial bias a recurrent concern in risk assessments. Still, due to this disproportion, a proper use of them to unwind mass incarceration will benefit minorities the most (Flores, Bechtel, & Lowenkamp, 2016). All in all, AI models ought to have a solid ground-base in order to comply with the FATE standards.

In conclusion, should an AI device make a decision about human justice? Despite the several limitations, there are good enhancing arguments for accepting the use of actuarial risk assessments like COMPAS. Even knowing that they are not perfect, they can be considered the least bad choice available at the moment. They actually bring the possibility of a sentencing reform and can help in the unwinding of the mass incarceration

problem present in the United States. Further on, by making the proper adjustments on them and implementing the FATE practices, they possess great potential for growth over and above the U.S. Although misinformation and misunderstanding of how the AI systems work can hold this change back (Flores, Bechtel, & Lowenkamp, 2016), the truth is that, at the moment, risk assessment algorithms are doing nothing more than what judges have been doing for decades using their own criteria, but faster and with lower costs. In the words of Steve Jobs (2007) *"Let's go invent tomorrow rather than worrying about what happened yesterday".*

7. BIBLIOGRAPHY

Agnew, R. (1992). Foundation for a General Strain Theory of Crime and Delinquency. *Criminology, 30*(1), 47-88.

Agnew, R. (1999). A General Strain Theory of Community Differences in Crime Rates. *Journal of Research in Crime and Delinquency, 36*(2), 123-155.

Akers, R. L. (1998). *Social learning and social structure: A general theory of crime and deviance.* Boston, MA: Northeastern University Press.

American Law Institute. (2017). Model Penal Code: Sentencing (Black-letter of proposed final draft). Philadelphia: American Law Institute.

Andrews, D. A. (1982). *The Level of Supervision Inventory (LSI): The first follow-up.* Toronto, ON : Ontario Ministry of Correctional Services.

Andrews, D. A., & Bonta, J. (1995). *The Level of Service Inventory-Revised.* Toronto, CA: Multi-Health Systems.

Andrews, D. A., & Bonta, J. (1998). *The Level of Service Inventory-Revised: Screening Version.* Toronto, ON: Multi-Health Systems.

Andrews, D. A., Bonta, J. L., & Wormith, J. S. (2004). *The Level of Service/Case Management Inventory (LS/CMI): An Offender Assessment System: User's Manual.* Toronto, CA: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. (1995). *The Level of Service Inventory-Ontario Revision (LSI-OR).* Toronto, ON: Ontario Ministry of the Solicitor General and Correctional Services.

Andrews, D. A., Bonta, J., & Wormith, J. (2008). *The Level of Service/Risk, Need, Responsivity (LS/RNR). User's manual.* Toronto, ON: Multi-Health Systems.

Andrews, D. A., Bonta, J., & Wormith, J. (2010). The Level of Service (LS) assessment of adults and older adolescents. In R. K. Otto, & K. Douglas, *Handbook of violence risk assessment tools* (pp. 199-225). New York, NY: Routledge.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Needs Assesment. *Crime & Delinquency, 52*(1), 7-27.

Andrews, D. A., Bonta, J., & Wormith, S. J. (2011). The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention? *Criminal Justice and Behavior, 38*(7), 735-755.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 1-27. Retrieved from ProPublica.

Austin, J. (1983). Assessing the new generation of prison classification models. *Crime and Delinquency, 29*(4), 561-576.

Bellman, R. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco, CA: Boyd & Fraser.

Bhatt, N., Bhatt, N., & Prajapati, P. (2017). Deep Learning: A New Perspective. *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), 6*(6), 136-140.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's Reducing a Human Being to a Percentage; Perceptions of Justice in Algorithmic Decisions. *CHI'18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Montreal, CA: ACM.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's Reducing a Human Being to a Percntage; Perceptions of Justice in Algorithmic Decisions. *CHI'18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Montreal, CA: ACM.

Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *Validation of the COMPAS Risk Assessment Classification Instrument.* Tallahassee, FL: Florida State University.

Bonta, J., & Andrews, D. A. (2017). *The Psychology of Criminal Conduct (6th ed.).* New York, NY: Routledge.

Breitenbach, M., Dieterih, W., & Brennan, T. (2009). Creating risk-scores in very imbalanced datasets. In Y. S. Koh, & N. Rountree, *Rare association rule mining and knowledge discovery: Technologies for infrequent and critical event detection* (pp. 231–254). Hershey, PA: IGI Global.

Brennan, T. (2008a). Explanatory diversity among female delinquents: Examining taxonomic heterogeneity. In R. Zaplin, *Female crime and delinquency: Critical perspectives and effective interventions* (pp. 197-232). Boston: Jones and Bartlett.

Brennan, T. (2008b). Institutional assessment and classification of female offenders: From robust beauty to person-centered assessment. In R. Zaplin, *Female crime and delinquency: Critical perspectives and effective interventions* (pp. 283-322). Boston: Jones and Bartlett.

Brennan, T., & Breitenbach, M. (2009). The taxonomic challenge to general theories of delinquency: Linking taxonomy development to delinquency theory. In O. Sahin, & J. Maier, *Delinquency: Causes, reduction and prevention* (pp. 1-38). Hauppauge, NY: Nova Science.

Brennan, T., & Dieterich, W. (2018). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). In J. P. Singh, G. D. Kroner, J. S. Wormith, S. L. Desmarais, & Z. Hamilton, *Handbook of Recidivism Risk/Needs Assesment Tools* (pp. 49-75). Chichester, UK: Wiley-Blackwell.

Brennan, T., & Oliver, W. L. (2002). *Evaluation of reliability and predictive validity of the COMPAS scales.* Traverse City, MI: Northpoibte Inc.

Brennan, T., Breitenbach, M., & Dieterich, W. (2008). Towards an explanatory taxonomy of adolescent delinquents: Identifiying several social-pschological profiles. *Journal of Quantitative Criminology, 24*(2), 179-203.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justie and Behavior, 36*(1), 21-40.

Brinkrolf, J., & Hammer, B. (2018). Interpretable Machine Learning with Reject Option. *Automatisierungstechnik , 66*(4), 283-290.

Bronson, J., & Carson, E. A. (2019). *Prisoners in 2017.* Washington, DC: National Criminal Justice Reference Service.

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine, 26*(4), 53-60.

Burgess, E. W. (1928). Residential Segregation in American Cities. *The ANNALS of the American Academy of Political and Social Science, 140*(1), 105-115.

Cardoso, J. (2006). Developing Dynamic Packaging Applications using Semantic Web based Integration. In A. F. Salam, & J. R. Stevens, *Semantic Web Technologies and eBusiness: Toward the Integrated Virtual Organization and Business Process Automation* (pp. 1-39). Hershey, PA: Idea Group.

Casey, P. M., Elek, J. K., Warren, R. K., Cheesman, F., Kleiman, M., & Ostrom, B. (2014). *Offender Risk & Needs Assesment Instruments: A Primer for Courts.* Williamsburg, VA: National Center for State Courts.

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). *It Takes Two to Tango: Towards Theory of AI's Mind.* Atlanta, GA: ArXiv.

Charniak, E., & Mcdermott, D. (1985). *Introduction to Artificial Intelligence.* Boston, MA: Addison-Wesley.

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology, 37*(1), 51-89.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big data & Society, 4*(2), 1-14.

Cohen, A. K. (1955). *Delinquent Boys: The culture of the gang.* New York: Free Press.

Cohen, L., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activities Approach. *American Sociological Review, 44*(4), 588-608.

Cohen, L., Felson, M., & Land, K. (1980). Property Crime Rates in the United States: A Macrodynamic Analysis. 1947–1977, with Ex Ante Forecasts for the Mid- 1980s. *American Journal of Sociology, 86*(1), 90-118.

Cornish, D. B., & Ronald, V. C. (1986). *The Reasoning Criminal.* New York: Springer-Verlag.

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Models. *American Psychologist, 34*(7), 571-582.

Desmarais, S. L., & Singh, J. P. (2013). *Instruments for assessing recidivism risk: A review of validation studies conducted in the U.S.* New York, NY: Council of State Governments Justice Center.

Dieterich, W., Brennan, T., & Oliver, W. L. (2011). *Predictive validity of the COMPAS Core risk scales: A probation outcomes study conducted for the Michigan Department of Corrections.* Traverse City, MI: Northpointe Inc.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity.* Traverse City, MI: Northpoint. Inc.

Dieterich, W., Oliver, W. L., & Brennan, T. (2011). *Predictive validity of the Reentry COMPAS risk scales: An outcomes study with extended follow-up conducted for the Michigan Department of Corrections.* Traverse City, MI: Northpointe Inc.

Doyle, T. (2019). Obfuscation and Strict Online Anonymity. In D. Berkich, & M. Vincenzo d'Alfonso, *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence* (pp. 359-370). Cham, CH: Springer.

Dressel, J., & Farid, H. (2018). The Accuracy, Fairness and Limits of Predicting Recidivism. *Science Advances, 4*(1), 1-5.

Durkheim, E. (1933). *The Division of Labor in Society.* New York: Macmillan.

Durkheim, E. (1952). *Suicide: A Study in Sociology.* Glencoe, IL: Free Press.

Dutton, D. G., & Kropp, P. R. (2000). A review of domestic violence risk instruments. *Trauma, Violence and Abuse, 1*(2), 171-181.

El Abid Amrani , N., Youssfi, M., & Abra, O. E. (2018). Semantic interoperability between heterogeneous multi-agent systems based on Deep Learning. *6th International Conference on Multimedia Computing and Systems (ICMCS)* (pp. 1-6). Rabat, MA: IEEE.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor.* New York: St. Martin's Press.

Farabee, D., Zhang, S., Roberts, R. E., & Yang, J. (2010). *COMPAS validation study: Final report.* Los Angeles, CA: UCLA ISAP.

Felson, M. (1998). *Crime and Everyday Life.* Thousand Oaks, CA: Pine Forge Press.

Fernández-Cabán, P. L., Masters, F. J., & Phillips, B. M. (2018). Predicting Roof Pressures on a Low-Rise Structure From Freestream Turbulence Using Artificial Neural Networks. *Frontiers in Built Environment, 4*(68), 1-16.

Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A New Approach to the Application ofTheorem Proving to Problem Solving. *Artificial Intelligence, 2*(3-4), 189-208.

Fishel, S., Flack, D., & DeMatteo, D. (2018, January). Computer Risk Algorithms and Judicial Decision-making. *Judicial Notebook, 49*(1), 35-36.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Federal Probation, 80*(2), 38-46.

Frase, R. S. (2004). Limiting Retributivism. In M. Tonry, *The Future of Imprisonment* (pp. 83-120). New York: Oxford University Press .

Gendreau, P., Goggin, C., & Little, T. (1996). A meta-analysis of the predictors of adult off ender recidivism: What works! *Criminology, 34*(4), 575-607.

Genesereth, M. R., & Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

Gottfredson, M. R., & Hirschi, T. (1990). *A General Theory of Crime.* Stanford: Stanford University Press.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The Clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Waltham, MA: Elsevier.

Harbers, M., Peeters, M. M., & Neerincx, M. (2017). Perceived Automony of Robots: Effects of Appearance and Context. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, & G. S. Virk, *A World with Robots* (pp. 19-33). Cham, CH: Springer.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

Hellman, D. (2008). *When is Discrimination Wrong?* Cambridge, MA, and London, ENG: Harvard University Press.

Henrichson, C., & Delaney, R. (2012). *The Price of Prisons: What Incarceration Costs the Taxpayer.* New York: Vera institute of Justice.

Hirschi, T. (1969). *Causes of Delinquency.* Berkeley: University of California Press.

Hofmans, J., Ceulemans, E., Steinley, D., & Van Mechelen, I. (2015). On the Added Value of Bootstrap Analysis for K-Means Clustering. *Journal of Classification, 32*(2), 268-284.

Hoge, R. D., & Andrews, D. A. (2002). *Youth Level of Service/Case Management Inventory: User's manual.* Toronto, ON: Multi-Health Systems.

Holland, O. (2004). The future of embodied artificial intelligence: Machine consciousness? In F. Lida, R. Pfeifer, L. Steels, & Y. Kuniyoshi, *Embodied artificial intelligence* (pp. 37-53). Berlin: Springer.

Holstein, K., Wortman Vaughan, J., Daume III, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *2019 CHI Conference on Human Factors in Computing Systems.* Glasgow, SCT: ACM.

Horwitz, A. V. (1990). *The Logic of Social Control.* New York: Plenum Press.

Jobs, S. (2007, May 30). Bill Gates and Steve Jobs at D5. (K. Swisher, & W. Mossberg, Interviewers)

Jones, P. R. (1996). Risk Prediction in Criminal Justice. In A. T. Harland, *Choosing Correctional Options that Work* (pp. 33-68). Thousand Oaks, CA: Sage.

Kak, A., & Richardson, R. (2020). *Consultation: Proposals for Ensuring Appropriate Regulation of Artificial Intelligence.* New York: The Office of the Privacy Commissioner of Canada.

Kleiman, M., Ostrom, B., & Cheesman, F. (2007). Using Risk Assessment to Inform Sentencing Decisions for Nonviolent Offenders in Virginia. *Crime and Delinquency, 53*, 106-132.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* Cambridge, MA : MIT Press.

Korinek, A., & Stiglitz, J. (2017). *Artificial Intelligence and its implications for income distribution and unemployment.* Cambridge, MA: National Bureau of Economic Research.

Kumar, C. (2018, August 31). Artificial Intelligence: Definition, Types, Examples, Technologies. *Medium*, pp. 1-5.

Kurzweil, R. (1990). *The Age of Intelligent Machines.* Cambridge, MA: MIT Press.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety.* Stamford, CT: META Group.

Lansing, S. (2012). *New York State COMPAS-Probation risk and needs assessment study: Evaluating predictive accuracy.* Albany, New York: New York State Division of Criminal Justice Services, Office of Justice Research and Performance.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory* . Oxford, NY: Oxford University Press.

Lawrence, A. (2013). *Trends in Sentencing and Corrections: State Legislation.* Denver: National Conference of State Legislatures.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436-444.

Lipton, Z. C., & Steinhardt, J. (2019). Troubling Trends in Machine Learning Scholarship. *ACM queue, 17*(1), 1-33.

Luger, G. F., & Stubblefield, W. A. (1993). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* Redwood City, CA: Benjamin/Cummings.

Martín del Brío, B., & Sanz, A. (2006). *Redes neuronales y sistemas borrosos.* Zaragoza: Ra-Ma.

McDermott, D. (2007). Artificial Intelligence and Consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson, *The Cambridge Handbook of Consciousness* (pp. 117-150). Cambridge, ENG: Cambrige University Press.

McLeod, R. (1995). *Management Information Systems: A Study of Computer-Based Information Systems* . Upper Saddle River, NJ: Prentice Hall.

Merton, R. (1938). Social Structure and Anomie. *American Sociological Review, 3*(5), 672-682.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). An Overview of Machine Learning. In R. S. Michalski, J. G. Carbonell, T. M. Mitchell, R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 3-16). Berlin: Springer-Verlag and Heidelberg GmbH.

Miró Llinares, F. (2018). Inteligencia Artificial y Justicia Penal: Más Allá de los Resultados Lesivos Causados por Robots. *Revista de Derecho Penal y Criminología, 3*(20), 87-130.

Mochon, F. (2019). Editor's Note. *International Journal of Interactive Multimedia and Artificial Intelligence, 5*(4), 4-7.

Monahan, J., & Skeem, J. (2016). Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology, 12*(1), 489-513.

Morris, N. (1974). *The Future of Imprisonment.* Chicago, IL: University of Chicago Press.

Narayanan, A. (2018). 21 Fairness Definitions and Their Politics. *FAT* Conference* (pp. 1-2). New York: ACM.

Northpointe Inc. (2011). *Practitioners Guide to COMPAS.* Traverse City, MI: Northpointe Inc.

Northpointe Inc. (2012). *Selected Questions Posed by Inquiring Agencies.* Traverse City, MI: Northpointe Inc.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Olver, M. E., Stockdale, K. C., & Wormith, S. J. (2014). Thirty Years of Research on the Level of Service Scales: A Meta-Analytic Examination of Predictive Accuracy and Sources of Variability. *Psychological Assessment, 26*(1), 156-176.

Pagallo, U., & Durante, M. (2016). The pros and cons of legal automation and its governance. *European Journal of Risk Regulation*, 323-334.

Patrick, B. (2020). What is Artificial Intelligence? *Journal of Accountancy, 229*(2), 69-73.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware Data Mining. *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 560-568). Las Vegas, NE: ACM.

Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in Human and Artificial Intelligence Decision-Making as the Basis for Diversity and Educational Inclusion. In J. Knox, Y. Wang, & M. Gallagher, *Artificial Intelligence and Inclusive Education. Perspectives on Rethinking and Reforming Education* (pp. 39-59). Singapore: Springer.

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning.* Cham, CH: Springer.

Reiss, A. J. (1951). Delinquency as a Failure of Personal and Social Controls. *American Sociological Review, 16*(2), 196-207.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*(5), 737-748.

Rich, E., & Knight, K. (1991). *Artificial Intelligence.* New York: McGraw-Hill Education.

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., & Whitlock, E. P. (2009). Toward a Transdisciplinary Model of Evidence-based Practice. *The Milbank Quaterly, 87*(2), 368-390.

Schalkoff, R. J. (1990). *Artificial Intelligence: An Engineering Approach* . New York: Mcgraw-Hill.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Selbst, A. D., & Powles, J. (2017). Meaningful Information and the Right to Explanation. *International Data Privacy Law, 7*(4), 233-242.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 1*, pp. 59-68. Atlanta, GA: Association for Computing Machinery, Inc.

Shapiro, L., & Stockman, G. (2001). *Computer Vision.* Upper Saddle River, NJ: Prentice Hall.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Graep. (2017). Mastering the Game of Go Without Human Knowledge. *Nature, 550*(7676), 354-359.

Skeem, J., & Lowenkamp, C. T. (2016). *Risk, Race & Recidivism: Predictive Bias and Disparate Impact.* Berkeley, CA: University of California.

Smith, J. (2009). *Well said! Great speeches in American history.* Washington, DC: E & K Publishing.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-ronot interaction. *HRI '06: Proceedings of the 1st ACM SIG- CHI/SIGART conference on Human-robot interaction* (pp. 33-40). Salt Lake City, UT, USA: ACM New York.

Subramanian, R., Moreno, R., & Broomhead, S. (2014). *Recalibrating Justice: A Review of 2013 State Sentencing and Correction Trends.* New York: Vera Institute of Justice.

Sutherland, E. H. (1924). *Principles of Criminology.* Chicago, IL: University of Chicago Press.

Taulli, T. (2019). *Artificial Intelligence Basics: A Non-Technical Introduction.* Monrovia, CA: Apress.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society,, 58*(1), 267-288.

Tittle, Charles R. (2000). Theoretical Develpments in Criminology. *The Nature of Crime: Continuity and Change, 1*(1), 51-101.

Toby, J. (1957). Social Disorganization and Stake in Conformity: Complementary Factors in the Predatory Behavior of Hoodlums. *Journal of Criminal Law, Criminology, and Police Science, 48*(1), 12-17.

Trinder, L., & Reynolds, S. (2000). *Evidence-Based Practice: A Critical Appraisal.* Oxford, ENG: Blackwell Science.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind 49*, 433-660.

Van Dijk, J. (1994). Understanding crime rates: On the interactions between the rational choices of victims and offenders. *British Journal of Criminology, 34*(2), 105-121.

Van Voorhis, P., Bauman, A., Wright, E. M., & Salisbury, E. J. (2009). Implementing the Women's Risk/Needs Assessments (WRNAs): Early lessons from the field. *Women, Girls, & Criminal Justice, 10*(6), 81-96.

VanNostrand, M., & Lowenkamp, C. T. (2013). *Assesing pretriak risk without a defendant interview.* Houston, TX: Laura and John Arnold Foundation.

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems -CHI'18* (pp. 1-14). Montreal, CA: ACM Press.

Wagner, P., & Sawyer, W. (2018). *States of incarceration: the global context 2018.* Northampton, MA: Prison Policy Initiative.

Wang, P., & Goertzel, B. (2012). *Theoretical Foundations of Artificial General Intelligence.* Osnabrück, DE, Germany: Atlantis Press.

Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime and Law, 10*(3), 243-257.

Ward, T., & Stewart, C. (2003). Criminogenic needs and human needs: A theoretical model. *Psychology, Crime and Law, 9*(2), 125-143.

Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Standford Law Review, 70*(5), 1343-1429.

Wilson, J., & Hermstein, R. (1985). *Crime and Human Nature.* New York: Simon & Schuster.

Winston, P. H. (1992). *Artificial Intelligence.* Boston, MA: Addison-Wesley .

Wormith, S. J., & Bonta, J. (2018). The Level of Service (LS) Instruments. In J. P. Singh, D. G. Kroner, S. Wormith, S. L. Desmarais, & Z. Hamilton, *Handbook of Recidivism Risk/ Needs Assessment Tools* (pp. 117-145). Chichester, UK: Wiley-Blackwell.

APPENDIX

Apprendix A. Example of a Report done with LSI-R.

LSI-R

Level of Service Inventory-Revised

By D.A. Andrews, Ph.D. & James L. Bonta, Ph.D.

Profile Report

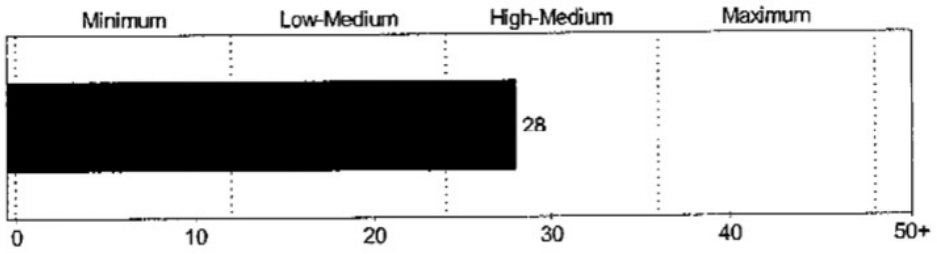
Name:	**Rex Darlington**
Assessment Age:	37
Gender:	Male
Social Security #:	
ID Number:	
Referral Source:	
Reason for Referral:	
Present Offenses:	
Disposition:	
Rater:	
Purpose of Report:	
Context	Community: Presentence Report/Predisposition Report
Other Client Issues:	None Specified
Assessment Date:	March 22, 2001

Introduction

The Level of Service Inventory-Revised is a risk and needs assessment tool. This report summarizes the results of the LSI-R administration, and provides information pertinent to the assessment of the individual.

Overall Assessment Based on LSI-R Total Score

The graph below shows the LSI-R Total Score and indicates the classification level associated with that score.



Assessment Based on LSI-R Score

Source/Purpose of Classification	LSI-R Score
Overall LSI-R Score	High-Medium
Risk Level (Community)	Maximum level of supervision/service is suggested, but consider medium supervision with management and/or treatment of dynamic risk factors.
Probation Guideline	Maximum Surveillance
Halfway House	Not appropriate unless intensive supervision and treatment are also provided.
Probability of Recidivism	44%

Comparison to Prison Inmates

The score is as high or higher than 59.7% of a normative group of prison inmates tested with the LSI-R.

Important Note

This LSI-R was re-scored once due to the availability of new information.

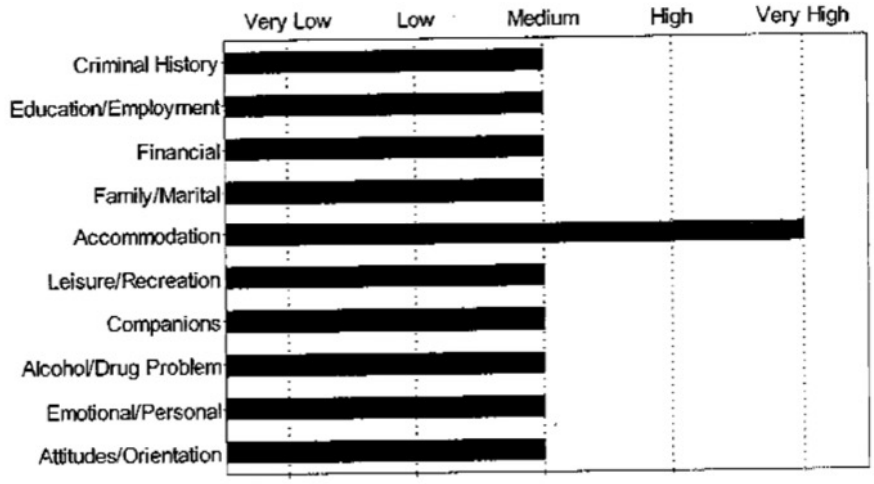
Professional Discretion/Override

The professional discretion/override was not used in this case.

MHS

Assessment of Risk/Needs Based on LSI-R Subcomponents

The graph below displays specific areas, and indicates whether they are low, medium, or high risk/needs areas.



Details Regarding Subcomponent Risks/Needs

Criminal History

1. Any prior adult convictions: Yes
4. Three or more present offenses: Yes
6. Ever incarcerated upon conviction: Yes
8. Ever punished for institutional misconduct: Yes

Education/Employment

11. Currently unemployed: Yes
13. Never employed for a full year: Yes
14. Ever fired: Yes
17. Suspended or expelled at least once: Yes

Financial

22. Reliance upon social assistance: Yes

Family/Marital

24. Non-rewarding, parental: A relatively unsatisfactory situation with a need for improvement
26. Criminal-Family/Spouse: Yes

MHS

Accommodation
27. Unsatisfactory: A relatively unsatisfactory situation with a need for improvement
28. 3 or more address changes last year: Yes
29. High crime neighborhood: Yes

Leisure/Recreation:
31. Could make better use of time: A relatively unsatisfactory situation with a need for improvement

Companions
32. A social isolate: Yes
33. Some criminal acquaintances: Yes
35. Absence of anti-criminal acquaintances: Yes

Alcohol/Drug Problem
37. Alcohol problem, ever: Yes
38. Drug problem, ever: Yes
41. Law violations: Yes
42. Marital/Family: Yes
44. Medical: Yes

Emotional/Personal
46. Moderate interference: Yes
47. Severe interference, active psychosis: Yes
49. Mental health treatment, present: Yes

Attitudes/Orientation
51. Supportive of crime: A relatively unsatisfactory situation with a need for improvement
54. Poor, toward supervision: Yes

Summary of LSI-R Item Responses

The rater entered the following response values for the items on the Level of Service Inventory-Revised Form.

Item	Response	Item	Response	Item	Response
1.	Y	19.	2	37.	Y
2.	N	20.	2	38.	Y
3.	N	21.	3	39.	2
4.	Y	22.	Y	40.	3
5.	N	23.	3	41.	Y
6.	Y	24.	1	42.	Y
7.	N	25.	2	43.	N
8.	Y	26.	Y	44.	Y
9.	N	27.	1	45.	N
10.	N	28.	Y	46.	Y
11.	Y	29.	Y	47.	Y
12.	N	30.	N	48.	N
13.	Y	31.	1	49.	Y
14.	Y	32.	Y	50.	N
15.	N	33.	Y	51.	1
16.	N	34.	N	52.	2
17.	Y	35.	Y	53.	N
18.	3	36.	N	54.	Y

Additional Item Information

1. Number of prior convictions: Not Specified
4. Number of present offenses: Not Specified
8. Number of times punished for institutional misconduct: Not Specified
40. Type of drug associated with current drug problem: None Specified
45. Other indicators of drug problem: None Specified
50. Area of psychological assessment indicated: None Specified

User-Defined Questions

1. Registered owner of a firearm?
   No

2. afasdf
   dsf

Date Printed: Friday, October 26, 2001

End of Report

MHS

Appendix B. Example of a Profile Report done with the LS/CSMI retrieved from Ms Lake's case in p. 81-94 of Andrews, Bonta and Wormith (2004).

Case Study

Probation Intake Report

Name: Louise Lake

Age: 37 years old

Date: April 1, 2017

Reason for Assessment

Ms. Lake is beginning a six-month period of probation. She pleaded guilty to one count of possession of narcotics for the purpose of trafficking. The conditions of probation are minimal and entail reporting as requested by the probation officer. According to the police report, Ms. Lake was found with 7 ounces of cannabis during a police raid at a local dance club. This is Ms. Lake's first conviction as an adult. During the interview, Ms. Lake presented as a cooperative and friendly woman. She answered all questions freely and appeared frank in her discussion of the present situation.

Criminal History

Official documents show no prior criminal history. Ms. Lake stated that she has never been arrested by police either as an adult or as a juvenile. Quite the contrary, she describes herself as a law-abiding citizen and ashamed of her present encounter with the law. Ms. Lake reported that the possession offence resulted from holding the cannabis for her husband who was described as a recreational user. She denied using cannabis herself and said that the police had no choice but to arrest her because "I had possession."

Education/Employment

Ms. Lake described herself as always liking school and never experiencing any behavioral or academic difficulties. She continued with school until graduating from the local university 10 years ago. She received a degree in business and presently Ms. Lake is employed by the Best-Thing department store where she is an accountant. Her employer is aware of the present offence but her job is not in jeopardy.

Best-Thing department store has been Ms. Lake's employer since she graduated from college. Her employer describes her as an excellent worker and valued employee who is well liked by the staff. Ms. Lake reported that she enjoys her work very much and that over the years she has been given increasing responsibility, which she finds both challenging and rewarding. Her supervisor is also a close friend of the family who has been quite supportive during the court proceedings. Ms. Lake works in an office with four other employees. They appear to have a very good collegial relationship, spending coffee breaks together and playing on a company bowling team.

Family/Marital

Mr. Lake's use of cannabis has been a long-standing concern for Ms. Lake. She had never liked his use of the substance, even though it was relatively infrequent (once a month). They have argued over his use in the past and these arguments have become more frequent as their daughter has become older. Ms. Lake feels that his drug use sets a bad example for their child (although her husband has never used cannabis in their daughter's presence). The present conviction has further added to the strain between the couple but Ms. Lake denies that the situation has become so intolerable that she wishes to seek a separation. Ms. Lake commented that power/control was not a relationship issue but she noted how her legal problems were linked directly to her relationship with her husband.

Ms. Lake's parents live in the city and they visit her regularly. Ms. Lake is particularly close to her mother. They have lunch at least once a week and her aunt and uncle, who are retired, look after the daughter while Ms. Lake and her husband are at work. Only one in Ms. Lake's family has been in conflict with the law. Mr. Lake was convicted of possession of a narcotic three years ago.

Leisure/Recreation

Ms. Lake is a member of her company's bowling team as well as the local Neighbourhood Watch and Block Parents organizations. The latter two activities involve monthly meetings and the preparation of a newsletter and periodic fundraising activity. In addition to these activities, Ms. Lake belongs to a neighbourhood book-reading club and during the summer, she enjoys gardening. In the winter, she takes weekend ski lessons.

Companions

To the best of her knowledge, none of Ms. Lake's friends has been involved with the criminal justice system. In fact, she finds it difficult to imagine herself associating with anyone who has been arrested by the police. Ms. Lake reported that her two closest friends (a colleague from work and an old childhood friend) know about the present offence and are shocked by it. However, they see this event as an unusual circumstance unlikely to be repeated. Actually, one of her friends drove her to the appointment for this interview.

Alcohol/Drug Problems

Ms. Lake denies ever having a drug or alcohol problem. She has never experimented with any drugs and expressed dismay that her husband still uses cannabis. The "harder" drugs are seen as substances that can destroy a person's life and she hopes that her daughter will never be exposed to its dangers. Ms. Lake drinks socially and in moderation. She will drink a glass of wine on special occasions with her last drink taken at the retirement party for a co-worker last month. Ms. Lake's description of alcohol and drug use is collaborated by her husband and mother who were interviewed by this examiner.

Attitude/Orientation

Ms. Lake admitted that she was in possession of cannabis and feels that the officer who made the arrest, did so appropriately: "Their job is to enforce the law; in the long run it is good for everybody." She thinks a re-occurrence is unlikely: She looks forward to a more normal life, working and continuing her involvement with the family and her community. Quite prepared to accept the penalty the court deemed appropriate, she feels that the judge made a fair decision and is pleased that probation was the final decision. I explained probation to her and the possibility that there may be some restrictive conditions accompanying the probation order. Ms. Lake understood and said, "Whatever is involved, I hope that the probation officer can help me put this part of my life behind me."

Antisocial Pattern

Ms. Lake presented without a single indication of a pattern of antisocial behaviour. There were no indicators of antisocial personality, no history of antisocial behaviour, no antisocial thinking, and no pattern of generalized trouble.

Other Client Issues

No other specific risk/need indicators were present. Similarly, an exploration of financial, accommodation, health, and emotional/personal issues revealed no problematic areas.

The combined family income for Ms. Lake and her husband is $93,000. Mr. Lake works as a landscape architect. They own their own home and a three-year-old car. Ms. Lake denies any difficulties in meeting mortgage or car payments. In fact, they have been able to save money for vacation trips each year and for the future education of their eight-year-old daughter. Neither Ms. Lake nor her husband has ever been on any form of social assistance.

The Lakes' home is in a quiet and well-established neighbourhood of the city. They have lived in the same residence for the past 8 years. Ms. Lake is a member of the Block Parents Association and the block captain for Neighbourhood Watch. Last year they upgraded their kitchen and bathroom and Ms. Lake hopes that this home will be their residence for many years to come.

According to Ms. Lake's mother, Ms. Lake was always a happy and sociable child. Ms. Lake did well in school and had no medical problems. She denied ever seeing a counsellor or mental health professional and described her life as very satisfactory. Her only wish is that her husband stop using cannabis.

Summary and Recommendations

Ms. Lake impresses as a sincere mature woman who appeared to have made one mistake that she wishes to forget. The results of the LS/CMI placed her in the Very Low risk-need range. Her score was two. Offenders with similar scores showed a very low likelihood of returning to crime (1). The only area that showed a potential for treatment targeting was her relationship with her husband, their disagreements over his cannabis use, and her apparent willingness to "carry" the substance on at least one occasion.

Community Case Management Plan

I discussed marital counselling with Ms. Lake and she will explore services available at a local family service agency. That agency is known to favour short term structured marital counselling with special attention to quality and equity in interpersonal relationships. There are no problematic special responsivity considerations beyond the possibility of power/control as a women's issue. Notably, marital counselling may well build on the many strengths noted in this case. They included Criminal History, Education/Employment, Leisure Recreation, Companions, Attitudes, and Pattern. Once counselling is underway and progress confirmed by the counsellor and participants, I anticipate a favorable early closure.

Jeff Atlas

Intake Probation Officer

Case Management Discharge Summary

Name: Louise Lake

Age: 37 years

Date: Sept 29, 2017

Background

Ms. Lake received a six-month period of probation for possession of cannabis. She was assessed at intake as a Very Low risk case with a multitude of strengths. Assigned to minimal supervision she was referred to a family service agency for marital counseling. The only identified criminogenic factor was marital dissatisfaction centering on her husband's occasional use of cannabis. A favorable early case closure was expected.

Case Management

Ms. Lake and her husband made early contact with the family agency and entered structured behavioral counseling with a focus on an equitable relationship. With only four counseling contacts over four weeks, the husband committed to cease drug use

and Ms. Lake committed to having no contact with the substance or with her husband during the occurrence of a lapse. With four additional contacts, the counselor and the Lakes reported to this probation officer that their counseling goals had been achieved. Family-counselor phone contacts were planned for once a month for the following three months.

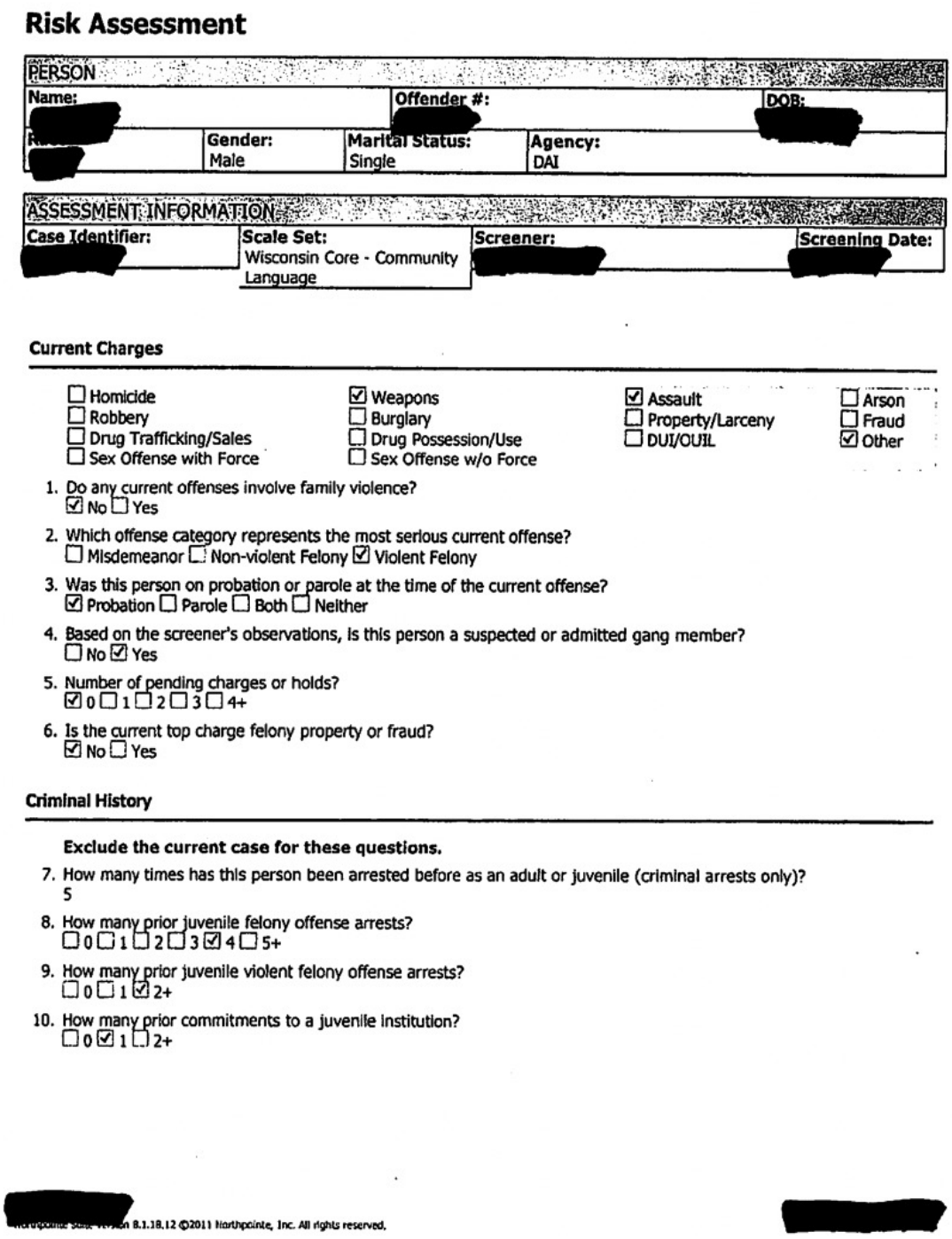
Case Closure

After the first eight weeks, the case was closed with the understanding that the probation officer (or Ms. Lake or the counselor) might initiate a contact at any time up to the end of the formal six- month probation period.

Sarah Repaz

Probation and Case Management Officer

Appendix C. COMPAS questionnaire version from Wisconsin, a state that applies COMPAS at every stage of the Criminal Justice System after an individual has been convicted (Angwin, Larson, Mattu, & Kirchner, 2016).



Risk Assessment

PERSON			
Name: [redacted]		Offender #: [redacted]	DOB: [redacted]
[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION			
Case Identifier: [redacted]	Scale Set: Wisconsin Core - Community Language	Screener: [redacted]	Screening Date: [redacted]

Current Charges

☐ Homicide	☑ Weapons	☑ Assault	☐ Arson
☐ Robbery	☐ Burglary	☐ Property/Larceny	☐ Fraud
☐ Drug Trafficking/Sales	☐ Drug Possession/Use	☐ DUI/OUIL	☑ Other
☐ Sex Offense with Force	☐ Sex Offense w/o Force		

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

[redacted] 8.1.18.12 ©2011 Northpointe, Inc. All rights reserved. [redacted]

Note to Screener: The following Criminal History Summary questions require you to add up the total number of specific types of offenses in the person's criminal history. Count an offense type if it was among the charges or counts within an arrest event. Exclude the current case for the following questions.

11. How many times has this person been arrested for a felony property offense that included an element of violence?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
12. How many prior murder/voluntary manslaughter offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
13. How many prior felony assault offense arrests (not murder, sex, or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
14. How many prior misdemeanor assault offense arrests (not sex or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
15. How many prior family violence offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
16. How many prior sex offense arrests (with force) as an adult?
☑0 ☐1 ☐2 ☐3+
17. How many prior weapons offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
18. How many prior drug trafficking/sales offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
19. How many prior drug possession/use offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
20. How many times has this person been sentenced to jail for 30 days or more?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
21. How many times has this person been sentenced (new commitment) to state or federal prison?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
22. How many times has this person been sentenced to probation as an adult?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+

Include the current case for the following question(s).

23. Has this person, while incarcerated in jail or prison, ever received serious or administrative disciplinary infractions for fighting/threatening other inmates or staff?
☑ No ☐ Yes
24. What was the age of this person when he or she was first arrested as an adult or juvenile (criminal arrests only)?
14

Non-Compliance

Include the current case for these questions.

25. How many times has this person violated his or her parole?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+
26. How many times has this person been returned to custody while on parole?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+
27. How many times has this person had a new charge/arrest while on probation?
☐0 ☐1 ☐2 ☐3 ☑4 ☐5+
28. How many times has this person's probation been violated or revoked?
☐0 ☐1 ☑2 ☐3 ☐4 ☐5+

[redacted]

[redacted]

29. How many times has this person failed to appear for a scheduled criminal court hearing?
☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

30. How many times has the person been arrested/charged w/new crime while on pretrial release (includes current)?
☐ 0 ☐ 1 ☐ 2 ☑ 3+

Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
☐ Both Natural Parents
☐ Natural Mother Only
☐ Natural Father Only
☐ Relative(s)
☐ Adoptive Parent(s)
☐ Foster Parent(s)
☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

35. Were your brothers or sisters ever arrested, that you know of?
☐ No ☑ Yes

36. Was your wife/husband/partner ever arrested, that you know of?
☑ No ☐ Yes

37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
☑ No ☐ Yes

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
☑ No ☐ Yes

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

40. How many of your friends/acquaintances served time in jail or prison?
☐ None ☐ Few ☑ Half ☐ Most

41. How many of your friends/acquaintances are gang members?
☐ None ☑ Few ☐ Half ☐ Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?
☑ None ☐ Few ☐ Half ☐ Most

43. Have you ever been a gang member?
☐ No ☑ Yes

44. Are you now a gang member?
☐ No ☑ Yes

Substance Abuse

What are your usual habits in using alcohol and drugs?

45. Do you think your current/past legal problems are partly because of alcohol or drugs?
☑ No ☐ Yes

46. Were you using alcohol or under the influence when arrested for your current offense?
☐ No ☑ Yes

47. Were you using drugs or under the influence when arrested for your current offense?
☑ No ☐ Yes

48. Are you currently in formal treatment for alcohol or drugs such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

49. Have you ever been in formal treatment for alcohol such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

50. Have you ever been in formal treatment for drugs such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

51. Do you think you would benefit from getting treatment for alcohol?
☑ No ☐ Yes

52. Do you think you would benefit from getting treatment for drugs?
☑ No ☐ Yes

53. Did you use heroin, cocaine, crack or methamphetamines as a juvenile?
☐ No ☑ Yes

Residence/Stability

54. How often do you have contact with your family (may be in person, phone, mail)?
☐ No family ☐ Never ☐ Less than once/month ☐ Once per week ☑ Daily

55. How often have you moved in the last twelve months?
☐ Never ☑ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

56. Do you have a regular living situation (an address where you usually stay and can be reached)?
☐ No ☑ Yes

57. How long have you been living at your current address?
☑ 0-5 mo. ☐ 6-11 mo. ☐ 1-3 yrs. ☐ 4-5 yrs. ☐ 6+ yrs.

58. Is there a telephone at this residence (a cell phone is an appropriate alternative)?
☐ No ☑ Yes

59. Can you provide a verifiable residential address?
☐ No ☑ Yes

60. How long have you been living in that community or neighborhood?
☐ 0-2 mo. ☐ 3-5 mo. ☐ 6-11 mo. ☑ 1+ yrs.

61. Do you live with family—natural parents, primary person who raised you, blood relative, spouse, children, or boy/girl friend if living together for more than 1 year?
☐ No ☑ Yes

62. Do you live with friends?
☑ No ☐ Yes

63. Do you live alone?
☑ No ☐ Yes

64. Do you have an alias (do you sometimes call yourself by another name)?
☑ No ☐ Yes

Social Environment

Think of the neighborhood where you lived during the past few (3-6) months.

65. Is there much crime in your neighborhood?
☑ No ☐ Yes

66. Do some of your friends or family feel they must carry a weapon to protect themselves in your neighborhood?
☑ No ☐ Yes

67. In your neighborhood, have some of your friends or family been crime victims?
☐ No ☑ Yes

68. Do some of the people in your neighborhood feel they need to carry a weapon for protection?
☐ No ☑ Yes

69. Is it easy to get drugs in your neighborhood?
☑ No ☐ Yes

70. Are there gangs in your neighborhood?
☐ No ☑ Yes

Education

Think of your school experiences when you were growing up.

71. Did you complete your high school diploma or GED?
☑ No ☐ Yes

72. What was your final grade completed in school?
9

73. What were your usual grades in high school?
☐ A ☐ B ☑ C ☐ D ☐ E/F ☐ Did Not Attend

74. Were you ever suspended or expelled from school?
☐ No ☑ Yes

75. Did you fail or repeat a grade level?
☑ No ☐ Yes

76. How often did you have conflicts with teachers at school?
☐ Never ☑ Sometimes ☐ Often

77. How many times did you skip classes while in school?
☐ Never ☑ Sometimes ☐ Often

78. How strongly do you agree or disagree with the following: I always behaved myself in school?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

79. How often did you get in fights while at school?
☐ Never ☑ Sometimes ☐ Often

Vocation (Work)

Please think of your past work experiences, job experiences, and financial situation.

80. Do you have a job?
☑ No ☐ Yes

81. Do you currently have a skill, trade or profession at which you usually find work?
☑ No ☐ Yes

82. Can you verify your employer or school (if attending)?
☑ No ☐ Yes

83. How much have you worked or been enrolled in school in the last 12 months?
☐ 12 Months Full-time ☐ 12 Months Part-time ☐ 6+ Months Full-time ☑ 0 to 6 Months PT/FT

84. Have you ever been fired from a job?
☑ No ☐ Yes

85. About how many times have you been fired from a job?
0

[redacted] 8.1.18.12 [redacted]

86. Right now, do you feel you need more training in a new job or career skill?
☑ No ☐ Yes

87. Right now, if you were to get (or have) a good job how would you rate your chance of being successful?
☑ Good ☐ Fair ☐ Poor

88. How often do you have conflicts with friends/family over money?
☐ Often ☐ Sometimes ☑ Never

89. How hard is it for you to find a job ABOVE minimum wage compared to others?
☐ Easier ☑ Same ☐ Harder ☐ Much Harder

90. How often do you have barely enough money to get by?
☐ Often ☑ Sometimes ☐ Never

91. Has anyone accused you of not paying child support?
☑ No ☐ Yes

92. How often do you have trouble paying bills?
☐ Often ☐ Sometimes ☑ Never

93. Do you frequently get jobs that don't pay more than minimum wage?
☐ Often ☐ Sometimes ☑ Never

94. How often do you worry about financial survival?
☐ Often ☐ Sometimes ☑ Never

Leisure/Recreation

Thinking of your leisure time in the past few (3-6) months, how often did you have the following feelings?

95. How often did you feel bored?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

96. How often did you feel you have nothing to do in your spare time?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

97. How much do you agree or disagree with the following - You feel unhappy at times?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

98. Do you feel discouraged at times?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

99. How much do you agree or disagree with the following -You are often restless and bored?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

100. Do you often become bored with your usual activities?
☑ No ☐ Yes ☐ Unsure

101. Do you feel that the things you do are boring or dull?
☑ No ☐ Yes ☐ Unsure

102. Is it difficult for you to keep your mind on one thing for a long time?
☑ No ☐ Yes ☐ Unsure

Social Isolation

Think of your social situation with friends, family, and other people in the past few (3-6) months. Did you have many friends or were you more of a loner? How much do you agree or disagree with these statements?

103. "I have friends who help me when I have troubles."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

104. "I feel lonely."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted] [redacted]

105. "I have friends who enjoy doing things with me."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

106. "No one really knows me very well."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

107. "I feel very close to some of my friends."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

108. "I often feel left out of things."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

109. "I can find companionship when I want."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

110. "I have a best friend I can talk with about everything."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

111. "I have never felt sad about things in my life."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Criminal Personality

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

113. "I always practice what I preach."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

114. "The trouble with getting close to people is that they start making demands on you."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

115. "I have the ability to "sweet talk" people to get what I want."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

116. "I have played sick to get out of something."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

117. "I'm really good at talking my way out of problems."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

118. "I have gotten involved in things I later wished I could have gotten out of."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

119. "I feel bad if I break a promise I have made to someone."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

120. "To get ahead in life you must always put yourself first."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Anger

121. "Some people see me as a violent person."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

122. "I get into trouble because I do things without thinking."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

123. "I almost never lose my temper."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

124. "If people make me angry or lose my temper, I can be dangerous."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted] [redacted]

125. "I have never intensely disliked anyone."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

126. "I have a short temper and can get angry quickly."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Criminal Attitudes

The next statements are about your feelings and beliefs about various things. Again, there are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

127. "A hungry person has a right to steal."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

128. "When people get into trouble with the law it's because they have no chance to get a decent job."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

129. "When people do minor offenses or use drugs they don't hurt anyone except themselves."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

130. "If someone insults my friends, family or group they are asking for trouble."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

131. "When things are stolen from rich people they won't miss the stuff because insurance will cover the loss."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

132. "I have felt very angry at someone or at something."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

133. "Some people must be treated roughly or beaten up just to send them a clear message."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

134. "I won't hesitate to hit or threaten people if they have done something to hurt my friends or family."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

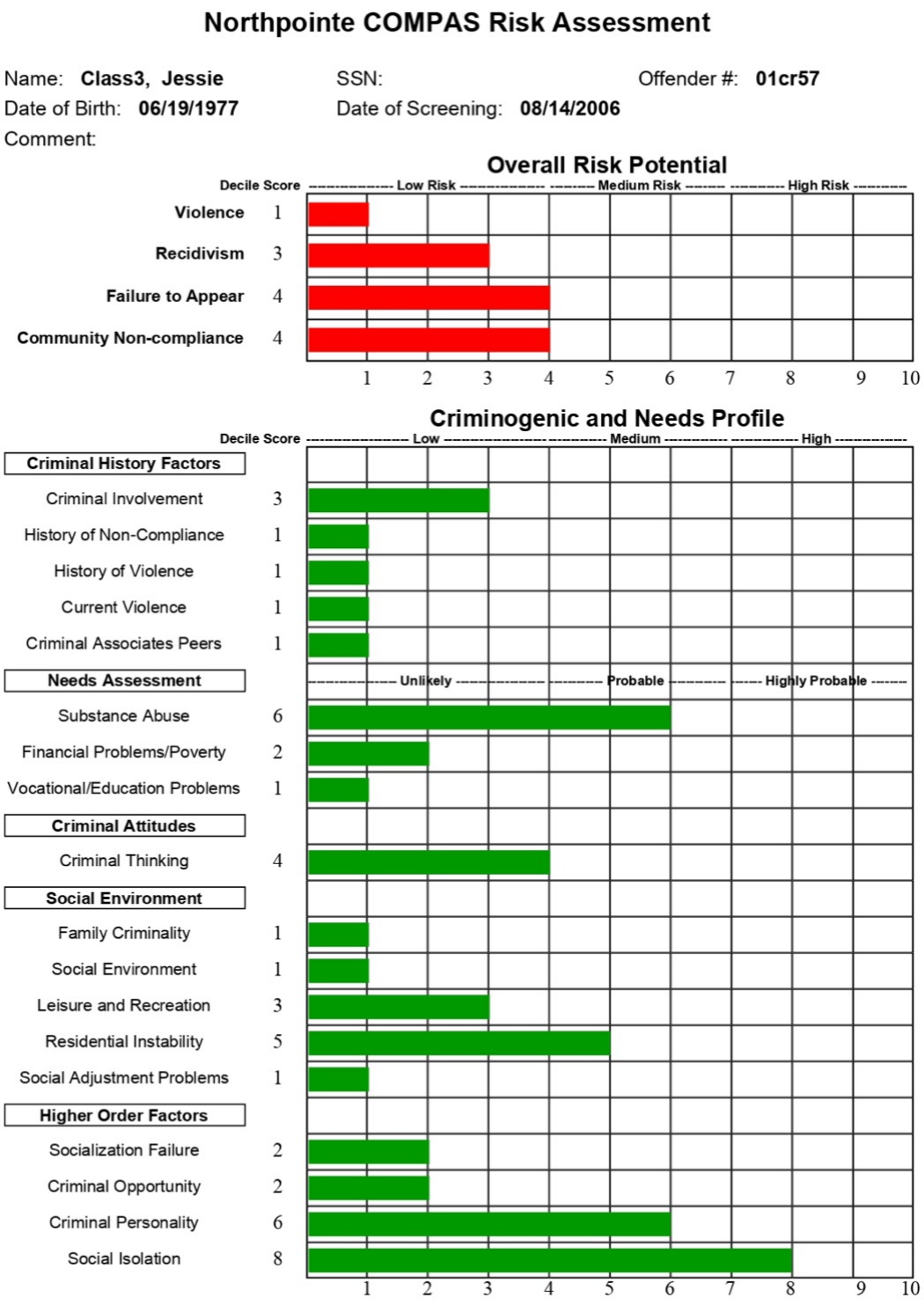
135. "The law doesn't help average people."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

136. "Many people get into trouble or use drugs because society has given them no education, jobs or future."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

137. "Some people just don't deserve any respect and should be treated like animals."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted] 8.1.18.12 [redacted]

Appendix D. Example of a COMPAS generated Final Report obtained from the Michigan Department of Corrections website (Angwin, Larson, Mattu, & Kirchner, 2016).



EXECUTIVE SUMMARY

In the past few years, an increasing use of Artificial Intelligence (AI) has been experienced in many fields all over the world, including the Criminal Justice System. Computerized assessment algorithms have been introduced with the aim of developing tools to prevent and reduce crimes, as well as perform tasks such as setting bail conditions or determining criminal sentences, among which risk assessments can be found. Over the years, that power of discretion that the judges hold has been seen as inappropriate and has been reduced (Angwin, Larson, Mattu, & Kirchner, 2016). This trend altogether with the substantial costs in which the Criminal Justice System of the United States is incurring to confront the mass incarceration has pointed out the forecasting of criminal behaviors as an optimal solution. The main goal of this dissertation is to address the controversial topic of whether AI methods can replace completely functions that human beings have been executing for ages, or if they are nothing else than a support instrument and, in any case, where should be the limits of its performance when decisions about human justice are on stake as they are on the Criminal Justice setting.

1. WHAT IS ARTIFICIAL INTELLIGENCE?

Artificial Intelligence (AI) is a broad-ranging branch of computer science aimed to design and build intelligent machines capable of performing tasks that would naturally require human intelligence. AI traces its roots back to the 1950s with the introduction of the term by Alan Turing, also known as the father of AI. He established the main goal and vision of AI by raising the question: *can machines think?* And that is, at its core, what AI attempts to answer in an affirmative way.

However, it is important to note that AI does not only perform tasks that humans are able to handle with their own brain but they can go beyond the capacity of a human brain (Mochon, 2019). The most common mistake is to narrow the term under computer science or mathematics, but AI is a puzzle conformed by pieces from many other domains such as economics, neuroscience, psychology, linguistics, electrical engineering and philosophy (Taulli, 2019) and it is used in a myriad of ways in our workaday without us even noticing (e.g. getting driving directions or looking for music recommendations).

1.1. Types of AI

- *Strong AI* or *Artificial General Intelligence* (AGI), is aimed to create machines with general intelligence at the human level or beyond (Wang & Goertzel, 2012). Unfortunately, AGI does not exist in reality yet.
- *Narrow AI* (NAI) or *weak AI*, is the attempt to create machines that perform tasks that would seem to require human intelligence. It is more simplistic than strong AI and specifies in systems designed to perform concrete instructions (e.g. Apple's Siri) (Taulli, 2019).

1.2. Machine Learning

ML is the most prevalent field and plays a major role within AI. Its goal is to create machines capable of learning how to perform tasks on their own from the data provided to them. For that aim, statistical algorithms emulate human cognitive tasks by working out their own procedures through the analysis of large training datasets. Some examples of it application are traffic predictions when using GPS navigation services or music and movie recommendations.

1.2.1. Application of ML techniques

To list a couple of the problems ML solve:

1) Classification: classify data into categories (e.g. divide emails into spam or not spam).
2) Predictions: forecast future values based on a model built upon historical data (e.g. predicting if an offender is likely to recidivate)
3) Clustering: take data and group items into clusters according to characteristics they have in common (e.g. customer segmentation).

1.2.2. ML Process

The first step to be taken is to select which data to feed the algorithm with. Secondly, what type of algorithm has to be decided by guesswork depending on the data available and the problem to be solved. The third step is the training phase, in which the algorithm will use the training data to find patterns and create a model, from which accurate predictions will be produced beyond the training data. Finally, the last step will consist on improving the algorithm by adjusting the values of their parameters (Taulli, 2019).

Data plays a key role in ML. While humans learn from past experiences, machines learn from data. Depending how is this information organized, data can be divided in three main groups: structured data, semi-structured data and unstructured data. Once it has been decided which type of data to work with, there are four ways in which machines can be thought how to do learn from it: supervised, unsupervised, semi-supervised and reinforcement learning (Taulli, 2019). For each teaching method, there is a huge list of algorithms to assist in the process. Among the most common ones, it can be found:

- *Linear Regressions*. Destined to expose the existent relationships between variables (y=ax+b).
- *Decision Trees*. From a starting point, decision paths will emerge, called splits. In the splits, an algorithm will be used to make the next path choice based on computational probabilities of variables until there are no more splits (Taulli, 2019).
- *Random Forests* are just compilations of Decision Trees.
- *Support Vector Machines* (SVM) are supervised ML algorithms used for classification and regression issues. As binary linear classifiers, they divide the data in the space into two classes according to a hyperplane boundary. The objective is to get the optimal hyperplane that correctly divides the data points, maximizing the margin. The main disadvantage is its black box component, that is, what the algorithm receives (input units) and what it comes out of it (output units) is known, but what happens in-between (the middle process, decisions, behaviors) is unknown. Consequently, being able to find the origin of an error or which are the predominant factors influencing the result is very complicated.
- *K-Means*. They are used in Unsupervised Learning to divide unlabeled data into different groups or clusters according to their similar characteristics. The letter k refers to the number of clusters and the centroids are the midpoints of the clusters. The k-Means algorithm will calculate the average distance of the centroids and change their location to position them in the center of each cluster (Taulli, 2019).

1.3. AI Today

With its ups and downs throughout history, AI finds itself in an advanced stage of weak AI after the real explosion of interest in AI started around 2010. The hype cycle occurred due to the impressive growth of computer power, the massive amount of data available leading to Big Data sources and the improvement of some AI approaches such as the just explained Machine Learning (Mochon, 2019).

2. THE FATE OF AI

Artificial Intelligence and corresponding technologies have started to be used to make high stake decisions in several areas such as education or the Criminal Justice System. The problem is that those domains can easily affect fundamental rights and liberties in significant manners. (Porayska-Pomsta & Rajendran, 2019). The movement Fairness, Accountability, Transparency and Ethics (FATE) of AI refers to the fundamental features that should characterize any system involving AI.

2.1. Fairness

The term Fairness in the FATE realm (algorithmic Fairness) refers to the impartial, just and non-discriminatory way of treating people. Therefore, for an AI/ML model to be fair, a person's experience with it should not vary depending on personal features such as their belonging to historically discriminated groups based on race, gender, sexual orientation, ethnicity, religion or age (Pedreschi, Ruggieri, & Turini, 2008).

2.1.1. Fairness in ML

It has been recently developed a field that looks after a fairer and more just Machine Learning models, denominated the Fairness-aware Machine Learning (fair-ML). the purpose of this movement is to introduce Fairness into the equation, or more precisely, into the algorithms, making it part of the *black box* system (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019).

Failure to consider the interactions between the social and the technological can lead the system to fall into an abstraction or category error, which could result in five different traps (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019): the Framing Trap, The Portability Trap, The Formalism Trap, the Ripple Effect Trap and the Solutionism Trap. In order not to get caught by these traps:

- For the Framing, include data and social factors to construct a heterogeneous frame that takes into consideration Fairness.
- For the Portability, introduce enough social and technical provisions of the desired context.
- For the Formalism, remind that social conditions such as Fairness are at times procedural, contextual and political. Therefore, ensure that the model can handle them.
- For the Ripple Effect, anticipate the way in which technology will affect the social context to assure that the tool will still solve the problem that it was meant to fix in the first place.
- For the Solutionism, design a system, if necessary, that befits the specific social context.

2.2. Accountability

AI presents an alternative for finding the optimal strategy that releases human beings from the efforts of making decisions. However, the necessity of having a system that holds accountable AI solutions and decisions is out of question taking into consideration the potential that it has to intensify the socio-cultural biases inherent to every society (Porayska-Pomsta & Rajendran, 2019)

2.3. Transparency

It claims for a clear exposition of the processes behind any AI system and the emptying of the *black box* as much as possible. There has to be an explanation for every outcome and therefore, simple algorithms like RF or logistic regressions are usually chosen over more complex ones such as SVM or ANN, which have large *black boxes* (Veale, Van Kleek, & Binns, 2018).

In any case, Transparency has to cope with the respect to privacy to avoid discrimination and promote autonomy. Autonomy allows for the liberty of decision, free of manipulation and coercion, and hence, threats to privacy result in a limitation of that freedom and a degradation of welfare (Doyle, 2019). In order to maintain privacy, anonymity is not enough but it helps in a significant way to protect it.

2.4. Ethics

The role of Ethics comes into play when deciding whether certain characteristics should be used as predictive variables (e.g. gender, age, ethnicity or sexual preference).

Also, the delegation of tasks to AI machines is already transforming the interactions and the environment in which humans used to live (Pagallo & Durante, 2016). This can result in a sense of disempowerment for humans, but it is also an inspiring tool for to improving their abilities and reflect on who they are and who they want to be.

3. RISK ASSESSMENTS IN THE U.S. CRIMINAL JUSTICE SYSTEM

The U.S. is the leader in incarceration by far, with imprisonment rates exceeding those of any other country in the world (Wagner & Sawyer, 2018). Furthermore, the racial proportions of the U.S. population given by the 2014 Census accounts for 62,1% White, 13,2% Black or African American and 17.4% Hispanic yet the prison population is categorized disproportionately as 37% Black, 32% White and 22% Hispanic (Flores, Bechtel, & Lowenkamp, 2016).

A proposed solution to decompress mass imprisonment without jeopardizing public safety is to use risk assessment instruments to lighten and speed up the workload behind sentencing and corrections. Correctional practice, in terms of activities related to treatment, punishment and supervision of people convicted of crimes, has been evolving from first-generation (1G) to fourth-generation incipient methods during the last three decades (Andrews, Bonta, & Wormith, 2006). Failures and weaknesses of previous stages have been overcome to reach the 4G approach today.

The characteristics that distinguish 4G approaches to the rest are (Brennan, Dieterich, & Ehret, 2009): a more extended theoretical background; additional risk and need factors that provide content validity; introduction of the strengths perspective of the GLM; more sophisticated statistics; a perfect integration of the need or risk domain with the management information system (MIS), criminal justice databases and web-based implementation of assessment technology. These features can be found in two of the main risk assessment algorithms used in the U.S.: the LS/CMI and COMPAS.

3.1. The Level of Service (LS) Assessments: LSI-R AND LS/CMI

In essence, LS instruments are quantitative assessments based on static and dynamic risk and need factors. The design is intended to be applied across populations of different ages, gender, race and ethnicities (Wormith & Bonta, 2018). The purpose of the instrument is to predict recidivism and other criminogenic conducts on the short (less than 6 months) and the long (more than 2 years) term.

All of the **versions** have fundamentally the same features, differentiated only by some innovations and adaptations to the context. The creation of the initial version of the Level of Service (LS) assessment started by Andrews (1982) in the late 1970s, in collaboration with the Ontario Ministry of Correctional Services (Casey, et al., 2014). The aim was to create a comprehensive instrument that registered the characteristics of offenders and would help establishing the level of supervision required for each of them. Subsequently, improved versions of the tool were designed until the release in 1995 of a 3G tool, The Level of Service Inventory-Revised (LSI-R) and its 4G updated version, The Level of Service/Case Management Inventory (LS/CMI) in 2004, that are still currently used in the U.S., Canada and other countries worldwide (Andrews & Bonta, 1995; Andrews, Bonta, & Wormith, 2004).

About the **design**, items on the scale have been selected on theoretical and empirical grounds, mainly the General Personality and Cognitive Social Learning theory (GPCSL) (Bonta & Andrews, 2017). This theory relies in the belief of multiple causes triggering the antisocial behavior that can be grouped in eight major areas of influence, the Central Eight risk/need factors. Section 1 in the LSI-R includes 54 items divided in 10 subcomponents while LS/CMI has only 43 items across 10 subcomponents. The additional sections 2-11 that receive no score in the LS/CMI are important though to collect information on influential factors that may trigger the criminal behavior. These are (Andrews, Bonta, & Wormith, 2004):

The **data collection** protocol consists essentially of an interview, supplemented by other sources of information such as file documents (e.g. criminal records, pre-sentence reports) or interviews with people close to the individual (e.g. family, co-workers) to assign a score to each item.

With regard to the **scoring**, both LSI-R and LS/CMI tools calculate a unique risk and needs score by summing the individual scores of each item in Section 1. The items receive a dichotomous scoring, that is, “1” if the item is present or “0” if the item is absent.

3.2. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)

The Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) was initially created in 1998 by the founders of Northpointe (rebranded as

Equivant), Tim Brennan and Dave Wells, who aimed to make a product that was better than the leading LS.

Core COMPAS is, originally, a 137-question survey that covers different domains of information such as defendant's criminal history, environment or personality. For risk prediction models, LASSO regression, logistic regression — to predict the probability of re-offending to happen depending on risk factors— and survival analysis — to predict the time until the next re-offending case—, were used to select and assign weight to the variables (Brennan, Breitenbach, & Dieterich, 2008).

From that Core COMPAS, different specific **versions** have been developed to better adjust to the different target populations according to the age (Youth COMPAS), sex (Women's COMPAS) or stage in the Criminal Justice System (Reentry COMPAS).

The **design** of COMPAS is characterized by the 4G characteristic features intended to move forward an Evidence-Based Practice (EBP) in Criminal Justice. It is composed of key scales incorporated from the main theoretical frameworks in the Criminology field such as the General Theory of Crime, Criminal Opportunity, Routine Activities, Social Learning, Subculture Theory, Social Control and Strain Theory. Among its more sophisticated statistics, COMPAS includes Artificial Intelligence technology such as SVM and RF.

Concerning the **data collection**, approximately one-third of the information is gathered from official records, one-third from self-report questions and one-third from an interview with the defendant/inmate (Blomberg, Bales, Mann, Meldrum, & Nedelec, 2010).

For the **scoring**, every item has an assigned weight (*w*) depending on the strength or potential to provoke a person's recidivism. Raw scores of each item is multiplied by its *w* to transformed them into weighted items (deciles) that would then be add together to calculate the final score (Northpointe Inc., 2011). The ranges are divided using decile scores intervals for which 1-4 is low, 5-7 is medium and 8-10 is high.

4. CRITICAL ANALYSIS OF COMPAS AS A CRIMINAL JUSTICE DECISION MAKING TOOL

4.1. Fairness of COMPAS

The predictive validity of the COMPAS General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) have been validated in diverse geographical

areas, diverse Criminal Justice agencies, and diverse gender and race categories. To evaluate the reliability, the Cronbach's alpha coefficient was used, where most of the Core COMPAS subscales provide alphas above .70 that denote satisfactory internal consistency. Also, it has been proven to be in a good or even excellent range of test-retest reliability. The critics on racial and gender biases have been proven to have many flaws and faulty statistics. The study of Flores et al. (2016) concluded that the reason why Blacks obtained higher scores on COMPAS was not a racial bias but an actual higher recidivism rate and many other researches have well-founded the equal predictive validity of COMPAS, determining it is racially unbiased and considering it a gender-responsive instrument.

4.1.1. Abstraction Traps

COMPAS can make a faster prediction than humans and more objectively as biases introduced by, for example, prejudices, fatigue, personal beliefs, would not be part of the final outcome. Based on this, COMPAS actually helps addressing those problems, but other risk assessments such as the LS instrument do it so as well. Hence, if the *Solutionism Trap* can be ruled out is not clear. To avoid the *Ripple Effect Trap* (Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019). COMPAS designers anticipated that the tool could affect the response of judges, prosecutors, officers and arrestees. Moving forward the *Formalism Trap,* it requires consideration of two things: 1) how to contemplate Fairness and Justice in the system. A widely accepted mathematical definition of Fairness has not been found yet, neither has Northpointe-Equivant included one in the COMPAS system but Fairness of the instrument has been validated.; 2) the desire for contestability, not satisfied yet but its solution falls outside Northpointe's hands. Consecutively, The *Portability Trap* is considered in COMPAS by its different versions and referencing data on eight different normative groups and different scoring guidelines according to the jurisdiction and location where is utilized (Brennan & Dieterich, 2018). In relation and to finish with, the *Framing Trap* reveals that a sociotechnical frame has to be the basis of the system and Northpointe itself cannot be heavily criticized in this matter as it has gathered sufficient data collection to be able to design the specific versions of COMPAS to modification and adapt in the best way they could to the needs of a heterogeneous population of offenders.

4.2. Accountability of COMPAS

Despite the mistakes and biases possibly made by humans, at least a certain degree of rationalization and accountability can be demanded to them, whereas who responds for COMPAS errors is not established yet. It is essential to find a way to hold the decisions accountable considering the impact that the use of the COMPAS software can have in the lives and well-being of criminal defendants.

4.3. Transparency of COMPAS

The Transparency problems that COMPAS might encounter are both internal and external. On the one hand, the opacity with regard to the inner methods of the tool and the *black box* component of the instrument related, among others, to the highly criticized use of SVM, makes it difficult for the public to get an explanation behind the outcomes. On the other hand, COMPAS provides governments with ways and means for collecting, tracking and analyzing large amounts of data without people's realization. This, apart from dealing with privacy issues, raises awareness about the risk of creating a feedback loop that prolongs and strengthens institutional bias in policing.

4.4. Ethics of COMPAS

The tool focuses on dynamic factors that can be subject to change. However, gender, race and age are part of the personal informative box. COMPAS clearly states its supportive functionality but the ultimate response of the judge is unknown; too much reliance on the given score can result in the appearance of the *Automation Bias* (Christin, 2017) and a sense of disempowerment for their own role as humans. The automation of the system also removes the chance of negotiation available when interacting with a human disappears with the use of COMPAS, perceiving it as dehumanizing, impersonal and lacking moral judgement (Binns, et al., 2018).

5. CONCLUSION

Through the analysis and comparison of COMPAS and LS, the following conclusions were drawn, which might be of interest to public and private companies developing risk assessment tools, people working for the Criminal Justice System and Governments.

- **Benefits:** they determine the risk of future criminal behaviors without requiring a judge examining the circumstances of the case and saving the costs of a full forensic evaluation, thus relieving the overburdened U.S. Criminal

Justice System as the frequency of the decision-making expediency increases; and they reduce the possibility of biased determinations based on prejudices and personal perceptions.

- **Drawbacks**: a judge's ruling will never be completely unprejudiced and impartial neither will an artificial model designed by humans. Also, they are programmed to perform concrete functions and nothing else so, how can we question their morality if they just do what they were intended to do?

Possible Solutions

- The ultimate goal should not be to design machines better than the ideal man, but instead, better than a real man.
- The incorporation of professional judgment combined with the model, using the later more as an auxiliary tool more than as an independent and sufficient instrument itself.
- The improvement of the Transparency of the procedures behind the outcomes.
- More emphasis on promoting proactive auditing of the systems to seek problems that go beyond the efficiency and efficacy of the instruments.
- Better access to individual-level demographics would facilitate finding the origin of one-off but also systematic biases.

Nevertheless, the use of AI in detecting or predicting crimes or an individual's risk of recidivism is a promising field that requires far more investigation, as well as education. There is a real need for educating judicial decision-makers about the strengths and weaknesses of these tools because there is still a huge lack of information aggravated by a massive sensationalism provoked by fake. Moreover, to achieve the creation of Fair, Accountable, Transparent and Ethic AI instruments, the reality is that a structural change is necessary beforehand. All in all, AI models ought to have a solid ground-base in order to comply with the FATE standards.

In conclusion, should an AI device make a decision about human justice? Even knowing that they are not perfect, they can be considered the least bad choice available at the moment. By making the proper adjustments on them and implementing the FATE practices, they possess great potential for growth over and above the U.S. At the moment, risk assessment algorithms are doing nothing more than what judges have been doing for decades using their own criteria, but faster and with lower costs.