MACHINE LEARNING IN CRIMINAL JUSTICE: A PHILOSOPHICAL ENQUIRY

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in State of Wisconsin v. Eric L. Loomis

Chiara Andreoli
Master thesis
MSc Philosophy of Science, Technology and Society
Faculty of Behavioral, Management and Social Sciences
University of Twente
Enschede, The Netherlands

Supervisor – Dr. Nolen Gertz
Second reader – Dr. Koray Karaca

September 2019

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I wish to express my deepest appreciation to my supervisor Dr. Nolen Gertz and to my second reader Dr. Koray Karaca. You made the last few months a period of growth for me, and for that I am very grateful. Many thanks to my fellow PSTS students, in particular to Hidde and Nils for serving as readers for my final draft. Special thanks to Nils, Laura and the kids for their friendship and for welcoming me in their home. A special thank you also goes to Danique and her family for their generosity. I am extremely grateful to my family for the love, trust and support, and to my good friends Francesco and Marco for always bringing so much joy into my life.

SUMMARY

Forecasting in criminal justice can be dated back to at least the 1920s, while the machine-learning version of it is a fairly recent development. In sentencing, machine learning was introduced to substitute the judge in assessing the recidivism risk of convicts eligible for parole. These assessments became interesting for criminal justice because of the promises they bare in terms of predictive accuracy in forecasting criminal behavior and in terms of the possibility of eliminating forms of bias in sentencing. Scholars and practitioners have identified both opportunities and challenges that can be associated to the introduction of this technology in criminal justice. My analysis focuses on a specific case: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in *State of Wisconsin v. Eric L. Loomis*. Through the analysis of the legal documents of the case and the literature available about COMPAS, I identified three main assumptions which indicate that the current debate fails to properly address: first, the meaning of terms such as fairness, predictive accuracy, and risk when it comes to machine learning; second, whether risk assessments provide evidence; third, what it means for the assessments to contribute to the reduction of costs in criminal justice, and whether the assessments are fit to address the problem of mass incarceration. Each of those shortcomings can be addressed by a branch of philosophy: the first refers to epistemological questions, and as such can be addressed by philosophy of science; the second refers to a question about the nature of technology, and therefore it can be addressed by philosophy of technology; the third refers to the nexus between criminal justice reform, economics, and mass incarceration, which can be addressed by philosophy of criminal justice. The main research question is: *how can philosophy of science, philosophy of technology and philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?* This thesis suggests that philosophy of science, through the discussion of value-ladenness, induction, the difference between explanation and prediction, and inductive and epistemic risk, can contribute to the understanding of the value of the knowledge associated to machine-learning risk assessments, and the implications of the purposes that knowledge is expected to serve. Philosophy of technology in turn can provide insight into the cost-benefit analysis that risk assessment entails, by taking a perspective that sees beyond an instrumental view of technology, and addresses panopticism and normalization in machine learning through the development of a Foucauldian argument. Last, philosophy of criminal justice, through an analysis that draws from Angela Davis' insights into detention in the United States, can provide the analytical tools needed to address the political and social implications of machine learning in criminal justice, and therefore contribute to the understanding of the role machine learning plays in relation to the goals of criminal justice reform.

CHAPTER 1 – Machine learning in criminal justice: introducing a philosophical enquiry

1.1 Introduction

Forecasting in criminal justice can be dated back to at least the 1920s (Berk, 2012; Borden, 1928; Burgess, 1928), while the machine-learning version of it is a fairly recent development (Berk, 2012). In sentencing, machine learning (hereinafter "ML") was introduced to substitute the judge in assessing the recidivism risk of convicts eligible for parole. These assessments became interesting for criminal justice because of the promises they bare in terms of predictive accuracy in forecasting criminal behavior and in terms of the possibility of eliminating forms of bias in sentencing. Scholars and practitioners[1] have identified both opportunities and challenges that can be associated to the introduction of this technology in the criminal justice. Among the challenges, for example, there is the potential threat to the right to individualized sentencing, and other aspects such as the possibility that ML risk assessments might actually be biased towards race, gender, or other factors. This fact not only raises doubts about the value that ML risk assessments are supposed to bring to criminal justice, but also would raise questions of whether it is unconstitutional to have this type of technology in sentencing. To limit the scope of the analysis to be performed in this thesis and try to better understand what ML in criminal justice actually entails, I decided to focus on a specific case: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in *State of Wisconsin v. Eric L. Loomis* (hereinafter "State v. Loomis"). In this chapter, after describing the legal case (1.2), I introduce the current debate around COMPAS based on the analysis of the available literature (1.3). I then identify three main assumptions that I will challenge with my analysis in this thesis (1.4) and, as part of explaining my methodology (1.5), I indicate the research questions for this project and explain the reasons for choosing to address these assumptions by drawing on three disciplines: philosophy of science, philosophy of technology and philosophy of criminal justice. I will now begin by introducing COMPAS in *State v. Loomis*.

1.2 COMPAS in *State of Wisconsin v. Eric L. Loomis*

Eric Loomis was accused of participating in a drive-by shooting that took place in La Crosse County, Wisconsin, in 2013. After being charged by the circuit court considering his case, Loomis waived his right to a trial and pleaded guilty. The court sentenced him to six years of confinement followed by five years of extended supervision. When Loomis was to be considered for parole, a ML tool called COMPAS (acronym that stands for Correctional Offender Management Profiling for Alternative

[1] For a study that provides insight from the perspective of judges on the introduction of ML risk assessment in sentencing, s*ee* Hyatt, J., & Chanenson, S. L. (2016). The use of risk assessment at sentencing: Implications for research and policy. *Villanova Law/Public Policy Research Paper*, (2017-1040). For an example of a ML risk assessment account that provides the perspective of legal practitioners in support of the application to criminal justice, s*ee also* Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, *27*(4), 222-228.

Sanctions) was used to assess Loomis for recidivism risk.[2] COMPAS assigned Loomis a score that indicated high risk of recidivism. The score was considered along with other types of evidence in the decision-making process aimed at choosing whether or not to release Loomis on parole. The Court decided not to release him. Loomis filed a motion arguing that the court's use of COMPAS during sentencing violated his rights to due process for three reasons: first, because he was confronted with the impossibility of challenging the assessment's scientific validity, due to the fact that the company producing COMPAS (Equivant, previously known as Northpointe) has refused to provide the details of the functioning of COMPAS because "it is all proprietary information and protected trade secrets" (State v. Loomis, 2016, p. 21); second, because it violated his right to an individualized sentence, as COMPAS doesn't in fact provide an individual assessment; and third, because it exposed him to a gender-biased sentence, as COMPAS takes gender into account when processing information. The State considered that Loomis had failed to prove that the court relied upon inaccurate information and on gender at sentencing, and therefore denied the motion. Loomis then decided to address the issue to the Court of Appeals, asking for a chance at a new sentence. The court of appeal said that even though COMPAS seemed to represent a positive improvement in sentencing, there could be a problem of due process in case of impossibility for defendants to test the validity of those tools (State v. Loomis, 2016), which would raise a question of whether it could be unconstitutional to use COMPAS in sentencing. The Supreme Court of Wisconsin (hereinafter "the Court") which was asked to provide on opinion on the case, concluded that "if used properly […] a circuit court's consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process" (State v. Loomis, 2016, p. 36) and therefore affirmed the order of the circuit court.

Both Loomis and the Court agree on what COMPAS is and what it delivers: a tool for recidivism risk assessment. COMPAS as a ML tool is software designed to provide an output based on input data such as past criminal activity, general attitude towards crime, and other personal characteristics. This data used for input is gathered through a questionnaire.[3] The output of the tool, which is supposed to be the prediction of the risk of recidivism of the person whose being evaluated, takes the form of a bar chart, with three bars that represent pretrial recidivism, general recidivism risk,

[2] While studying COMPAS, *State v. Loomis*, and recidivism in criminal justice in the United States, I have come across several mentions of recidivism, but rare definitions. The Supreme Court does not provide definitions of recidivism or recidivism risk. Equivant (2019) does not provide definitions either. In the current debate, as I will discuss further in this thesis, recidivism is discussed either in terms of the former convict falling back into criminal behavior, or a re-arrest for an offense committed after release (Brennan et al., 2009). Recidivism risk is described as "the likelihood that a prisoner would reoffend after being paroled" (Farabee et al., 2010). The National Institute of Justice (NIJ) defines recidivism as "a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime" (National Institute of Justice [NIJ], n.d.). The NIJ adds "recidivism is measured by criminal acts that resulted in rearrest, re-conviction or return to prison with or without a new sentence during a three-year period following the prisoner's release" (National Institute of Justice [NIJ], n.d.). Both definitions are available at: https://nij.ojp.gov/topics/corrections/recidivism. Last visited on September 3rd, 2019.

[3] Copy of COMPAS sample risk assessment questionnaire is available online at: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html Last visited June 25th, 2019.

and violent recidivism risk. Each bar indicates a defendant's level of risk "on a scale of one to ten" (State v. Loomis, 2016, p. 7).

Loomis and the Court, however, disagree on the quality of the knowledge that COMPAS provides, with consequences for the recognition of the role COMPAS is expected to play both in sentencing and in criminal justice more broadly. When it comes to discussing the role of COMPAS in sentencing, COMPAS is described as allowing the judge to better decide on whom should be released. However, there is disagreement between Loomis and the Court on how much of an improvement COMPAS actually brings to sentencing. Loomis claims that he's being discriminated on the basis of gender, and that he's unable to verify the validity of the assessment. The Court, on the other hand, believes that the combination of risk assessment tools and professional judgment. The Court confirms COMPAS as providing information that aligns with the current accepted standards for accuracy in risk assessment, and on whether that information is individualized or refers to a group rather than the specific convict the disagreement is settled because while it is true that COMPAS is based on group data, the data for the assessment of each convict is still collected individually with a questionnaire.[4] When it comes to the matter of being able to challenge the assessment, the problem of transparency is also dismissed because in any case the information on which the assessment was based was equally available to the judge and the defendant. The Court refers to risk assessments in relation to the goals of criminal justice reform (American Bar Association, n.d.). In particular, when connecting COMPAS to criminal justice reform goals, reference is made to the positive impact expected from the implementation of the so-called evidence-based practices in criminal justice,[5] as a way to reduce criminal justice costs and contribute to solving the problem of mass incarceration.

The Court substantiates its position by referring to academic literature explicitly discussing COMPAS (Brennan et al., 2009; Brennan & Oliver, 2013; Farabee et al., 2010; Fass et al., 2008; Klingele, 2016; Tallarico et al., 2012). Since the evaluation of the Court, the current debate has flourished, enriching the discussion through contributions by more authors (Angwin et al., 2016; Baird, 2009; Barabas et al., 2018; Bavitz et al., 2018; Beriain, 2018; Freeman, 2016; Hamilton, 2015; Kehl & Kessler, 2017; Starr, 2014; van Eijk, 2017). I will now present the most important points of discussion in the current debate. It is important to keep in mind that not all not all the sources were available at the time of the evaluation of the Court. All academic sources, those that were considered at the time of sentencing and those that were not, will be presented jointly.

1.3 The current debate around COMPAS

The current debate around COMPAS follows the three levels of analysis touched upon by *State v. Loomis* (tool level, sentencing level, and criminal justice reform level). COMPAS is described as

[4] Ibid.

[5] The National Institute of Corrections, which is an institute of the Department of Justice, defines evidence-based decision making (EBDM) as "a strategic and deliberate method of applying empirical knowledge and research-supported principles to justice system decisions made at the case, agency, and system level" (National Institute of Corrections, n.d.). Definition available online at: https://nicic.gov/evidence-based-decision-making. Last visited on September 3rd, 2019.

designed to “predict new offenses in a probation sample” (Brennan et al., 2009, p. 25), which means that the tool is trained to recognize a target, in this case the potential recidivist, whose characteristics are defined by the so-called “COMPAS norm group” (Brennan et al., 2009, p. 25). COMPAS is trained to perform the risk assessment and recognize the target based on a regression model (Brennan et al., 2009). Equivant states that the software is designed according to the best scientific theories on recidivism and crime prediction, but when it comes to the question of what COMPAS does, authors Skeem & Louden (2007) provide a compelling analysis indicating that “there is little evidence that the COMPAS predicts recidivism” (Equivant, 2019, p. 29), which should raise greater concerns even on the accepted understanding of the nature of COMPAS as a recidivism risk assessment tool. The strongest disagreements between the additional contributions presented in the debate and in *State v. Loomis* are about the quality of the knowledge provided by COMPAS. These disagreements generate diverging opinions on whether or not COMPAS should be used in sentencing. I will recall these contributions in relation to: first, the qualities of the knowledge provide by COMPAS; and second, the characterization of the operational function of COMPAS.

The qualities of the knowledge provided by COMPAS

When it comes to discussing bias and discrimination in relation to risk-assessment tools, the transfer of decision-making to an automated system is indicated as an opportunity to reduce discrimination and bias (Bavitz et al., 2018) because of the understanding of bias and discrimination as (at least in part) human-generated issues. There are three factors that are discussed in the current debate: race (Angwin et al., 2016; Fass et al., 2008; Starr, 2014), gender (State v. Loomis, 2016) and socio-economic factors (Baird, 2009; Barabas et al., 2018; Beriain, 2018; Klingele, 2016; Starr, 2014; van Eijk, 2017). In addition to those three the most important features of COMPAS discussed in the debate are: accuracy, reliability and predictive validity. The most positive voices in the debate (Brennan et al., 2009; Brennan & Oliver, 2013; Farabee et al., 2010) judge COMPAS to be reliable. More cautious authors (Bavitz et al., 2018; Hamilton, 2015; Tallarico et al., 2012) nevertheless optimistic about what the tool can contribute to criminal justice, focus on the current potential issues that the technology can present in terms of bias, accuracy and predictive validity, but believe better tools could be available thanks to technological improvements in the future. Those authors believe that harm can be avoided thanks to policies that will make sure the proper cautions are taken when using the tool. Authors adopting a much more critical stance (Angwin et al., 2016; Baird, 2009; Beriain, 2018; Barabas et al., 2018; Fass et al., 2008; Freeman, 2016; Kehl & Kessler, 2017; Klingele, 2016; Starr, 2014; van Eijk, 2017) focus on the shortcomings of the tool and the potential undesirable implication for criminal justice by discussing the same technical issues discussed by less critical authors, but also trying to go beyond the mainstream debate over technical features and consider the implications for society at large.

Another issue that is raised when discussing ML and criminal justice is whether the tool provides individual or group risk assessments (Starr, 2014). Those adopting a critical stance (Angwin et al., 2016; Baird, 2009; Beriain, 2018; Barabas et al., 2018; Fass et al., 2008; Freeman, 2016; Kehl & Kessler, 2017; Klingele, 2016; Starr, 2014; van Eijk, 2017), judge that COMPAS does not provide individual assessments, which means that "although the models are frequently depicted as a means to predict which offenders will reoffend" (Baird, 2009, p. 3), the assessments are in fact indicating that an individual is considered risky because of the characteristics they share with others who recidivated in the past.

COMPAS is also discussed in relation to transparency and opacity (Barabas et al., 2018; Beriain, 2018; Freeman, 2016; Kehl & Kessler, 2017; Starr, 2014) because it does not allow the users access to the processes through which the scores are produced. One of the solutions to this problem that has been proposed in the debate is to make use of open source algorithms (Beriain, 2018; Freeman, 2016). Open-source as a solution to transparency is an idea based on an understanding of the problem of opacity and lack of transparency as due to property rights the company producing the tool has on the tool itself which do not allow access to the algorithms.

The operational function of COMPAS

COMPAS, as a risk assessment tool is characterized as contributing to implement in criminal justice the so-called evidence-based[6] practices in criminal justice (Kehl & Kessler, 2017; Klingele, 2016; Tallarico et al., 2012). Evidence-based practices are described as "based on criminological research about 'what works' to reduce convicted individuals' odds of committing future crimes" (Klingele, 2016, p.537). Because of the costs involved in criminal justice practices, policymakers are looking for "proof that these programs were worthy of investment" (Klingele, 2016, p. 552). The National Centre for State Courts and the American Bar Association discuss the need for reform in criminal justice in terms of the importance of guaranteeing that criminal justice is carried out in a fair way and that the costs of criminal justice are reduced. The increase in criminal justice costs has been attributed to the problem of mass incarceration (American Bar Association, n.d; Tallarico et al., 2012). In this framework, the solution reformers proposed to reduce people that stay in detention: both those that are detained and do not need to be there, and those released and do not need to come back. Specifically, the Supreme Court recalls, the American Bar Association expressed concerns about the incarceration of low-risk offenders together with medium and high-risk offenders, as the exposure of the low-risk offenders to the company of more serious ones could actually increase their risk of recidivating (American Bar Association, n.d.).

1.4 Assumptions in the characterization of COMPAS

Through the analysis of the legal documents of *State v. Loomis* and the literature available about COMPAS I have identified three main assumptions, one for each level of analysis:

- At tool level, the assumption relates to the characterization of knowledge when it comes to COMPAS as fair and accurate and as providing a prediction of recidivism risk. That characterization, however, does not take into account that those terms have precise meanings

[6] Evidence-based practices can be described as striving to "improve sentencing decisions by incorporating scientific and quantitative methods" (Kehl & Kessler, 2017, p. 7) used to predict future behavior.

in ML that do not seem to align with the meaning associated to the same terms in criminal justice.

- At sentencing level, the assumption is that COMPAS is a risk assessment tool for the provision of evidence. In the debate, when it comes to discussing the assessments in terms of technology, that technology is only discussed in instrumental terms, meaning that ML software is considered a tool used to perform risk assessments, but because the starting point is an assumption that a risk assessment provides evidence about the individual, whether or not that is actually the case is neither questioned nor properly justified.
- At criminal justice reform level, the assumption is that COMPAS will contribute to the achievement of the goals of criminal justice reform: solving the problem of mass incarceration and reducing criminal justice costs. This assumption doesn't take into account the nexus between criminal justice economics and detention and does not account for the complex dynamics behind the costs of criminal justice. Therefore, it fails to even clarify which costs will be reduced.

In the next section, I indicate why philosophy is needed for the understanding of ML in criminal justice, and I will outline the main research question and the sub-questions that I will answer in this thesis.

1.5 A philosophical enquiry: research questions and methodology

The assumptions presented in the previous section indicate that the current debate does not properly address: first, the meaning of terms such as fairness, predictive accuracy, and risk when it comes to ML; then, whether ML risk assessments provide evidence; last, what it might mean for ML risk assessments to contribute to the reduction of costs in criminal justice, and whether those assessments are fit to help addressing the problem of mass incarceration. Each of those shortcomings can be addressed by a branch of philosophy: the first refers to epistemological questions, and as such it is pertinent to philosophy of science; the second refers to a question about the nature of technology, and therefore it can be addressed by philosophy of technology; the third refers to the nexus between criminal justice reform, economics, and mass incarceration, which can be addressed by philosophy of criminal justice.

My main research question then is: *how can philosophy of science, philosophy of technology and philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?*

As I indicated at the beginning of this chapter, to ground my analysis in criminal justice practice, I decided to focus on a specific case: COMPAS in *State v. Loomis*. My analysis is based on the available literature and legal documentation. This main research question will be addressed by answering three sub-questions, each of which constitutes a chapter of the thesis:

Sub-question 1 - *how can philosophy of science contribute to the understanding of machine learning in criminal justice?*

This sub-question will be answered in chapter 2, where I will draw from philosophy of science to challenge the assumptions on which the current debate bases the characterization of knowledge associated to COMPAS. I will address four elements of that assumption (fairness, accuracy, prediction and risk) by addressing value-ladenness, induction, the difference between explanation and prediction, and inductive and epistemic risk.

Sub-question 2 - *how can philosophy of technology contribute to the understanding of machine learning in criminal justice?*

This sub-question will be answered in chapter 3, where I will draw on philosophy of technology to challenge the assumptions on which the current debate bases its characterization of COMPAS as a risk assessment tool for the provision of evidence. The perspective of technology that I will take is that proposed by Heidegger in *The Question Concerning Technology* [Heidegger, [1993], 2008). This choice was made because Heidegger challenges the characterization of technology as a tool, which is one aim of this chapter. For the analysis I will also draw from Foucault's book *Discipline and Punish* (1991), as a way to deconstruct the characterization of risk assessment as provision of evidence. I will address panopticism in relation to the structure of the assessment, normalization in relation to the rule for the assessment, and the cost-benefit analysis that risk assessment entails when the two come together in a form of *triage*.

Sub-question 3 - *how can philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?*

This sub-question will be answered in chapter 4, where I will draw on philosophy of criminal justice to challenge the assumptions about the contribution COMPAS is supposed to give to the efforts of criminal justice reform to reduce costs and solve the problem of mass incarceration. Specifically, I will draw from Angela Davis' insight on detention in the United States, as it is a critical perspective on criminal justice that matches COMPAS geographical and historical framework.

1.6 Conclusions

In this chapter I introduced COMPAS in *State v. Loomis*, and the current debate around it. COMPAS is characterized as contributing to the efforts of criminal justice to improve the evidence-based decision-making processes that are deemed optimal for criminal justice practices. By improving the sentencing process, COMPAS is characterized as contributing to the reduction of recidivism, crime, and prison population, and therefore of the costs of the criminal justice system, benefitting society at large. However, after careful consideration of the legal case and the current debate surrounding COMPAS, I have identified a series of assumptions which impact on all the three levels of operational function COMPAS is connected to: tool, sentencing, and criminal justice reform. The assumptions relate to: the characterization of knowledge when it comes to COMPAS; the characterization of COMPAS as an evidence provider; and the characterization of COMPAS as contributing to achieving the goals of criminal justice reform, namely solving the problem of mass incarceration and reducing criminal justice costs. Through my initial analysis I identified one main research question and three sub-questions, which I will address in chapter 2,3 and 4. In the next chapter I will begin by addressing the first sub-question by drawing from philosophy of science to challenge the assumptions about knowledge in ML risk assessments.

CHAPTER 2 - Philosophy of science for the understanding of machine learning in criminal justice

2.1 Introduction

Equivant describes COMPAS as a ML tool to perform recidivism risk assessment. When it comes to the quality of the knowledge associated to the assessments, accuracy, reliability, predictive validity, interpretability, fairness and lack of bias are sometimes defined and sometimes used as words to characterize COMPAS without a clear definition of what those words mean. Equivant refers to reliability to in terms of test-rates and internal consistency, meaning “that its scales measuring needs have construct validity and behave consistently and that its risk scales have predictive validity” (Equivant 2019, p. 12). Since no definition is given of what it is meant exactly for predictive validity, for the purpose of this thesis predictive validity will be understood to be predictive accuracy, as it seems to be the more reasonable concept in this context. Accuracy is understood to refer to the ability of COMPAS to predict recidivism, which is to say, to estimate for convicts subject to the assessment “the likelihood of reoffending” (Equivant, 2019, p. 33). Interpretability is defined as “the ability to explain or to present in understandable terms to a human” (Doshi & Kim, 2017) the results produced through ML. Fairness, on the other hand, is explicitly said by Equivant to be an element that is defined in different ways by the company and by others that have tested COMPAS (e.g., by Angwin et al., 2016). In the current debate, knowledge in COMPAS is at times described as providing an explanation,[7] and most of the time as providing a prediction.[8] The characterization of knowledge in the current debate does not however provide a clear explanation of what is meant with those terms. Philosophy of science provides insight into the value and characteristics of knowledge in scientific activity, which can be used as an analytical tool to better understand COMPAS. In this chapter I will address four elements that I identified as part of the assumptions on which the current characterization of the knowledge associated to COMPAS is based in the current debate: fairness, accuracy, prediction, and risk.

I will begin by looking at the technical features of machine learning (2.2), to introduce a few important elements that need to be considered when addressing questions of knowledge associated to ML. Then, I will challenge the assumption about fairness (2.3) by discussing the design process that precedes having something like COMPAS available, as a way to address values in COMPAS. I will then move on to challenge the assumption about accuracy (2.4) by discussing the problem of induction. I will address the third assumption, the understanding of prediction (2.5), by revisiting the analogy between COMPAS and medical practice. Lastly, I will address risk (2.6) as the last element of

[7] Equivant refers to COMPAS with a medical analogy intended to clarify what COMPAS does, and writes, “case interpretation involves connecting the dots to understand the relationship between a person’s criminal behavior and her history, beliefs, and skills” (Equivant, 2019, p. 3). This characterization of COMPAS seems to indicate an explanatory meaning to the COMPAS scores.

[8] In the words of Equivant, the purpose of the risk scales is prediction “the ability to discriminate between offenders who will and will not recidivate” (Equivant, 2019, p. 7).

the assumptions about knowledge. I will introduce two types of risk that are not considered in the current debate but belong to ML risk assessments: inductive risk and epistemic risk In the conclusions (2.7) I will make a few final considerations to bring together the elements discussed in the chapter and I will introduce how I plan to move the analysis forward in chapter 3.

2.2 Machine learning

Machine learning can be described as a way to find a solution to a problem. Abu-Mostafa et al. (2012), write that the problem in this case is that there is a function (f) that represents in truth a relationship that is the object of study, which cannot be reached because not all the possible variables that might be playing a part in the relationship can ever be known. The solution to the problem is finding the mathematical function (g),[9] that is the closest possible approximation to that target function. The relationship that is the object of study can be expressed as the relationship between an input and an output (Abu-Mostafa et al., 2012) For example, the input could be data about certain characteristics of the individual, and the output could be the risk score associated to that individual. In other words, the function (g), is a model of the relationship between the characteristics of the individual and their risk (Karaca, 2019).

Machine learning based software requires three elements to be able to find the function (g): a dataset, a set of hypothesis, and a learning algorithm. The dataset contains the available data about the individuals that is deemed relevant for finding the function representing the relation that wants to be addressed. The set of hypotheses is made of functions that might represent the relationship to be mapped. And the learning algorithm is an algorithm that has the ability to adapt according to the data it processes. The algorithm provides a risk score by multiplying each factor deemed relevant for the assessment by a weight that represents the relative importance of that specific factor for what is being addressed (for example, the importance of certain socio-economic factors and personal characteristics and recidivism). It is important to know that the dataset does not indicate the weights that need to be associated to the characteristics of individuals for what is being addressed such as recidivism (Karaca, 2019). Another thing that the dataset does not indicate is which of the hypotheses about the function that models the relationship between input and output is the right one[10]. How can ML work if neither the right hypothesis nor the algorithm is provided by the data? One way to make it work is to assign either value zero or random values to the weights applying to the factors that are deemed relevant. When the algorithm is provided with data, it starts to calculate the function (g). The training then consists of comparing the risk that the algorithm assigns to the dataset with that which is deemed true, and making the algorithms ‘learn’ what is right when it comes to assigning weights and tracing the relationship between relevant factors. Two things are relevant at this point: hypotheses are formulated based on data; and assumptions are necessarily made at the beginning of the process to make it possible for the algorithm to know how to go about finding the function (g) representing the

[9] When referring to functions, I will refer to functions as the mathematical expression of what can be referred to as patterns in the data (McAllister, 1997; McAllister, 2011).

[10] Karaca (2019) calls this problem the underdetermination of ML by training data.

relationship between input and output that is consistent with the data (Karaca, 2019). Translated for COMPAS, this means that the target is not the norm group, but the relationship between a certain type of data and a score. In other words, COMPAS target is the relationship between a set of inputs (the characteristics deemed relevant about certain individuals) and a set of outputs (the characterization of those individuals as high or low risk). This means that the target function (f) could be the function representing in truth the relationship between the characteristics that are deemed relevant about individuals for the purpose of predicting recidivism (this would be the set of inputs, which can be seen in the COMPAS questionnaire), and the scores provided by COMPAS for recidivism risk (this would be the set of outputs).

In this section I described the technical features of COMPAS as ML software. In the next section I begin by addressing the first of those assumptions, which is that COMPAS provides knowledge through an assessment that is characterized as fair.

2.3 Fairness: software design and value-ladenness

COMPAS is software, and as such it needs to be designed. Design involves a series of decisions that have to do with data, datasets and the algorithms, the technical features of ML that I described in the previous section (2.2). When it comes to data, decisions are made about what data matters, and how it is characterized in terms of the parameters that are used for that type of data. These decisions affect the way data is collected, store and processed. Talking about the processing of information, in the previous section I indicated that there are two elements in the workings of ML that are of particular interest when questioning the value of knowledge in that context: the hypotheses about the relationship that is being modeled in COMPAS, and the weights associated to the factors deemed relevant for recidivism prediction. The technical features of ML also show that these decisions are necessarily discriminatory, in the sense of making a difference between factors that matter and factors that don't, how much each type of data matters in its relative importance to other factors.[11] These choices are made, but not often discussed, which means that their relevance for the characterization of fairness is not acknowledged.

One of the matters raised in *State v. Loomis* is whether or not COMPAS discriminates based on gender, in the sense of assigning higher risk to male convicts. In the current debate, another example of the discussion of bias is about whether COMPAS assessments are biased on the basis of race in the sense of assigning more risk to African Americans (Angwin et al., 2016; Larson et al., 2016). In both cases, when the question on whether or not COMPAS presents racial or gender bias is raised, it is dismissed either as just a matter of difference in the definition of fairness, or on the basis of the lack of proof that bias was actually the case. There is one difference however between the two cases, which is that while COMPAS does not use race as an explicit criteria in the assessment, it does use gender as an explicit factor for assessing risk. In *State v. Loomis* (2016), when gender bias is discussed, the most interesting response is that COMPAS does indeed discriminate based on gender, and that gender discrimination is not only acceptable but necessary, for statistical reasons which

[11] Merriam Webster defines *Discrimination* as "the act of making or perceiving a difference" https://www.merriam-webster.com/dictionary/discrimination last visited on August 10th, 2019.

ultimately have to do with predictive accuracy. In other words, COMPAS does discriminate on the basis of gender, but because predictive accuracy is considered of paramount importance, the question of whether or not gender discrimination is acceptable in criminal justice is discarded. In the design process choices were made about what is acceptable and relevant, and what needs to be prioritized. In the current debate this prioritization seems to take the form of giving priority to questions of accuracy over questions of discrimination. The choice of accuracy over concerns for discrimination is given as a necessary choice dictated by ML as a technology but while it is true that for statistical accuracy gender is useful, it is not dictated by the technology that discrimination on the basis of gender should be therefore accepted in criminal justice.

This rationale presented by the Supreme Court in *State v. Loomis* when it comes to discussing gender bias and discrimination, presents a logical fallacy. While bias necessarily requires making a difference, discrimination in the sense of making a difference does not necessarily mean bias in that negative sense that is commonly associated to the term. For example, in the case of the gender bias problem raised in *State v. Loomis,* gender is a discriminating factor because it is deemed a relevant factor in criminological theories on crime, recidivism, and the characterization of offenders. To the claim that COMPAS might be gender biased, the answer was that gender needs to be considered because it is a relevant discriminating factor when it comes to recidivism. Discrimination takes a negative connotation, and therefore becomes bias in that negative sense that is relevant to the claim, depending on why that difference is made, why it is considered a relevant factor, and what the making of that difference is used for. Gender bias in COMPAS is interesting especially when considered alongside racial bias. The part in the current debate that relates specifically to the discussion of racial bias in COMPAS (Angwin et al., 2016; Larson et al., 2016) shows that race is not, at this time, considered an acceptable criterion for discrimination in sentencing for parole. Race and ethnicity are still deemed relevant factors in different contexts, which include law enforcement, but also social policy implementation, in the United States (Abramovitz, 2006). What determines whether or not a factor is considered relevant and acceptable for discrimination depends on the historical circumstances, the context and the purposes to which the making of that distinction relates. Gender has undergone similar changes in its status as a discriminatory factor, and in relation to bias. When gender is used as a factor to discriminate in the sense of making a difference, gender is often regarded as no longer binary (Richards et al., 2016), and bias based on gender is now considered unacceptable in many contexts that were considered just a reflection of reality not too long ago. This example indicates that decisions that relate to what is relevant and what is not are decisions based on value judgments (Ananny, 2015; Gitelman, 2013; Kitchin, 2017; Zarsky, 2016). Those decisions both emerge from, and are the reflection of, societal values. COMPAS is designed and its design requires decisions making on what data is deemed relevant and the characteristics of that data that are based on value judgments.

In philosophy of science, these choices are indicated as based on epistemic and non-epistemic values. An example of epistemic value is accuracy, while an example of non-epistemic value is the belief that racial discrimination is unacceptable. Every time a decision is made, a prioritization takes place, either involving a choice of which values matter most between epistemic and non-epistemic, or even among epistemic values and among non-epistemic ones. Earlier in this chapter, I discussed how in COMPAS a choice has been made about the prioritization of concerns for accuracy over concerns for discrimination, but that choice has been described as an obligated choice. When it comes to looking at discriminatory factors such as gender and race in risk predictions, it should be remembered that although they are now included in software such as COMPAS, those factors had already been ruled out because of concerns for fairness in the 70s and the 80s (Tonry, 2014). The most interesting aspect of this change in what is relevant and what is not and whether that is acceptable in terms of fairness, is that not only the explicit factors 'gender' and 'race' were ruled out, but also education, family characteristics, employment and residential status, were considered unacceptable because they are strongly correlated with race (Hamilton, 2015; Tonry, 2014). Here is another element that is then missing from the discussion of COMPAS when it comes to bias: in addition to explicit bias on the basis of factors such as gender and race, there can also be implicit bias, based, for example, on other factors that could be proxies to race. COMPAS lacks race as an explicit criteria for the categorization and convicts and their relation to recidivism risk, but it uses socio-economic factors, therefore opening the possibility of racial bias by proxy. COMPAS uses socio-economic factors to assess recidivism and as such could still be considered as operating an assessment on the basis of race, even in the absence of the explicit criteria.

COMPAS knowledge is developed through a design process which is value-laden[12], which means that if bias and discrimination in COMPAS are to be discussed with a serious intent to address the issue, then their needs to be more attention for the complexity of the topic and the technical features of COMPAS. The discussion of value-ladenness in COMPAS shows two things: first, that there is a logical fallacy in the characterization of bias and discrimination in the current debate, and second, that there are inconsistencies in the characterization of COMPAS as fairer than judges at performing the assessments.

2.4 Accuracy: the problem of induction

Equivant characterizes COMPAS as accurately predicting recidivism. Earlier in this chapter, when describing the technical features of COMPAS (2.2), I indicated that ML requires a series of hypothesis to be formulated in order to provide the predictions. The current debate seems to suggest that the hypothesis is something that is known thanks to the data and the criminological theories on which COMPAS is built. In other words, the debate leaves the impression that COMPAS "knows" the criminal personality that is associated to a recidivist, and can identify it in the convicts that it assesses. However, I pointed out that the dataset does not provide which hypothesis is the right one which means that COMPAS does not know who a recidivist is and finds who matches that characterization. Rather, who a recidivist is emerges through the assessment. In other words, the technical features of ML discussed earlier (2.2) indicate that the one true hypothesis about who a recidivist is, is not indicated in the dataset. What ends up being the hypothesis that is considered true, depends on two

[12] Karaca (2019) argues that value-ladenness and bias are essential features of ML models used for societal applications.

things: the assumptions that are necessarily made during the design process, and the data that is given back to COMPAS for it to adjust its calculations based on whether the predictions turn out to be true.

In philosophy of science, going from data to formulating a hypothesis is known as "inductive inference" (Lipton, 2004). This matter has been discussed at length in philosophy of science, but here I will just recall a few aspects that should be sufficient for the purpose of this thesis. In philosophy of science inductive inference is connected to the so-called "problem of induction" (Hume, 1963; as cited in Ladyman, 2002; Lakatos, 1968; Lipton, 2004; Popper, 1959). The problem of induction (Hume, 1963; as cited in Ladyman, 2002) is that in an inductive argument, no matter how true all the premises might be there is still the possibility that the conclusions might be false. In the COMPAS case this means that no matter how true the criminological theories and the hypothesis that derive might be, the conclusions that are made about the individual might be false. Because there is no reason to believe that what happened in the past will happen in the future, there can be no justification for inductive practices (Hume, 1963; as cited in Ladyman, 2002). Hume's position on the matter has been challenged over the years, and induction is a very important aspect of ML that should be investigated further. For the purpose of this thesis is for now sufficient to consider the problem of induction as an additional indicator that the current debate does not do justice to the complexity behind the characterization of something as knowledge, and the practical consequences that doing so might have. For example, how does the problem of induction impact on the understanding of accuracy in ML risk assessments? Two considerations need to be made when considering what accuracy means in the case of COMPAS: first that the characterization of accuracy is dependent on what is considered proof of that accuracy, which does not take into account that only those convicts that are released actually contribute to the testing of the hypothesis; and second, that what is considered accurate depends on an a priori decision of what constitutes an acceptable level of error, which in the case of recidivism risk assessment correspond to the value of .70 (State v. Loomis, 2016; Baird, 2009; Brennan et al., 2009; Farabee et al., 2010; Skeem & Louden, 2007). Accuracy is then a much more complicated matter than the current debate recognizes.

In the debate accuracy refers to predictive accuracy. Equivant explains what it means for COMPAS to predict through an analogy to medical practice. I will analyze the analogy in the next section as a way to address the third element of the assumptions about knowledge: prediction.

2.5 Prediction: the analogy to medical practice

In the medical analogy that Equivant proposes, COMPAS ability to predict is associated to a model, a predictive model, that is compared to the "medical model of interpretation of information" (Equivant, 2019, p. 3) that assessed the symptoms, makes a diagnosis, and prescribes treatment. Philosophy of science indicates that models in science are something very specific. Models can be descriptive, explanatory or predictive, and it should be noted that for something to have explanatory power does not necessarily mean that it has predictive power as well (Shmueli, 2010). The difference between explanation and prediction and the complexity that it entails, is often underestimated and shortcomings in addressing the difference between the two has consequences for the characterization of scientific activity and of the knowledge it provides. Differentiating between explanation and prediction is necessary to understand what knowledge can be associated to COMPAS.

When it comes to discussing the difference between explanation and prediction, for most of the twentieth century, scientific explanation was philosophically understood to be as described by the so-called "covering law theory of explanation" (Hempel & Oppenheim, 1948; as cited in Godfrey-Smith, 2003, p. 191; Hempel, 1965; as cited in Ladyman, 2002, p. 200). This theory, simply put, states that explanations and predictions only differ in the fact that in the case of explanations the conclusion is known, while in the case of predictions it is not (Hempel, 1965; in Ladyman, 2002). The medical example involving symptoms and disease, in light on the covering law model of explanation, would see a symmetry in the understanding that explanation can be found both going from symptoms to disease, and going from disease to symptoms. Those addressing the shortcomings of the model[13] however, argued that symptoms do not explain the disease, meaning that explanation can only be found when going from disease to symptoms. In other words, explanation works only in a certain direction (Godfrey-Smith, 2003; Ladyman, 2002). This dynamic is known as "the asymmetry problem" (Godfrey-Smith, 2003, p. 193) and has to do with a misunderstanding about this idea that explanations and predictions are more or less the same thing, with the difference of time (whether the explanation/prediction is about the past or about the future). Understanding that there is an asymmetry problem that applies to the medical case, means that the symptoms do not explain the disease but they might be used to make a prediction, and that the disease can only explain why the patient is having symptoms. To see what this perspective can contribute to the understanding of COMPAS I now turn to the analogy between COMPAS and medical practice. Equivant writes:

> "A model that everyone can relate to is the medical model for interpretation of information gathered on a person. Think about the different steps taken in the medical field to find a solution to an illness or a problem. When you don't feel well and you go to the doctor, what is the first thing that the doctor does – Asks about symptoms: When did they start? How severe are they? She asks about your medical history: Are you taking any medications? Have you had this or a similar problem before? And, she runs tests, takes your temperature, takes your blood pressure, takes blood samples, orders MRIs, etc. What does she do with all of this information? She makes a diagnosis and prescribes an effective treatment."
>
> (Equivant, 2019, p. 3)

In the analogy proposed by Equivant, the symptoms are clearly indicated to be the elements considered in the questionnaire, and the diagnosis I take to be the assessment. The treatment that is prescribed to cure the disease I will take to be the decision on whether or not to release the individual. The analogy, however, does not clearly indicate whether the disease is meant to be the criminal

[13] For example, see Reilton (1978) and Kitcher (1981) for specific shortcomings debated in relation to Hempel's covering law theory. For a more general overview see Ladyman (2002), Godfrey-Smith (2003).

character of the offender, or the criminal behavior as the act that will make the former convict a recidivist. To make my point, I will discuss the analogy twice, once for each of the two possible characterizations of the disease. The analysis will involve looking at the assessment from the two directions of going from symptoms to disease, and from disease to symptoms, to see whether that assessment can be qualified as explanation or prediction, and what that means for the validity of the analogy to medical practice, and the characterization of knowledge in COMPAS.

C*riminal personality as "disease"*

This understanding of disease is that which recidivism risk refers since recidivism risk is defined as "the likelihood that a prisoner would reoffend after being paroled" (Farabee et al., 2010). The characterization of symptoms and diseases in the analogy to medical practice in this case can be seen as symptoms being the socio-economic conditions considered by COMPAS, and the disease being the criminal personality. Going from symptoms to disease, the socio-economic conditions assessed in COMPAS are not explanations of the criminal personality of the individual because while they might be characteristics that are found in individuals who committed crimes in the past, they do not explain why the person has a criminal personality. There might be predictive power when it comes to going from socio-economic conditions to criminal act, but those predictions might be based on an inaccurate understanding of the matter that is being addressed, as the explanation that is supposed to generate the prediction does not show to be justified for certain criteria considered in the COMPAS questionnaire. Going from disease to symptoms, there might be the explanation for some symptoms, but not for others. For example, having a "criminal personality" might explain past antisocial behavior, but it does not explain why the parents of the individual separated, or why the individual lived in indigence.

Criminal behavior as "disease"

This understanding of disease is that which recidivism refers to as recidivism is defined as "a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime" (National Institute of Justice [NIJ], n.d.). The characterization of symptoms and diseases in the analogy to medical practice in this case can be seen as symptoms being the socio-economic conditions considered by COMPAS, and the disease being the criminal act. Going from symptoms to disease, the socio economic conditions assessed in COMPAS might explain why someone went on to commit a criminal act. There might be predictive power, but those predictions might be based on an inaccurate understanding of the matter that is being addressed, as the explanation that is supposed to generate the prediction does not show to be justified for certain criteria considered in the questionnaire. Going from disease to symptoms, the criminal act committed at present time would not explain why the individual had certain socio-economic conditions in the past. For example, a crime committed today does not explain why an individual was having financial difficulties or why (s)he failed in school or his/her parents separated when (s)he was little.

Bringing the two analysis together, the analogy to medical practice shows that in COMPAS explanation is found only in going from disease to symptoms in case of the disease being the criminal character of the offender, and going from symptoms to disease in the case of the disease being the criminal act. In both cases, explanation is not found in the opposite direction. The analysis of COMPAS confirms that the covering law theory presents an asymmetry problem, meaning that explanation and prediction need to be considered different from each other. Moreover, when there is explanation in COMPAS, it is not equally the case for all the characteristics considered by the assessment, which indicates that there might be a problem with the theoretical foundations of recidivism risk assessment. Considering that explanation is found only in the direction of going from disease to symptoms, either that is not the case in COMPAS (e.g., when the criminal act is considered the disease), or it is the case only for certain symptoms (e.g., when the criminal character is considered the disease). Strangely enough, however, contrary to what is found in the medical analogy (which is that there cannot be explanation going from symptoms to disease), in the COMPAS case there is the possibility of explanation going from some of the symptoms to the disease. Because symptoms are not supposed to explain the disease, either those characteristics that are connected to a criminal act as symptoms are not symptoms at all, or crime is not the disease.

The characterization of the ability of COMPAS to predict recidivism seems to be based on confusion on the difference between explanation and prediction. COMPAS risk assessments are supposed to support decision-making in sentencing by providing information that helps the judges make a decision on the treatment (which in that case is either detention or release) but it is not clear what disease that information refers to. When it comes to questioning whether the knowledge COMPAS provides is actually what it is said to be in the current debate, so far my analysis indicates that not only the characterization of accuracy in the current debate does not account for the problem of inductive inference (2.4), but also the characterization of prediction through the analogy to medical practice is misleading (2.5). A crucial element of that analysis is the difference between the understanding of what the "disease" is according to the definitions of recidivism and recidivism risk. Challenging the characterization of risk is the topic of the next section.

2.6 Risk: inductive risk and epistemic risk

There are two types of risk that appear mentioned in relation to COMPAS in the current debate: recidivism risk and the risk of bias in sentencing. Risk is commonly defined as the possibility of an undesirable event taking place in the future.[14] When it comes to the definition of recidivism risk in the current debate, recidivism risk is defined as "the likelihood that a prisoner would reoffend after being paroled" (Farabee et at., 2010). When it comes to risk of bias, in the debate, the risk of bias is characterized as a problem that affects judges. Earlier in this chapter I discussed how the matter of bias in COMPAS is more complicated than it seems to be described in the current debate. First of all, there is a difference between bias and discrimination when it comes to considering the technical features of machine learning (2.2), and as my analysis showed, discrimination is a necessary part of COMPAS as ML software. Moreover, it is then not sufficient to look for explicit bias such as based on race or

[14] Merriam-Webster dictionary online defines risk as "the possibility of loss or injury." Definition available at: https://www.merriam-webster.com/dictionary/risk

gender, in addition bias should also be carefully assessed in its implicit forms, the forms that can be implemented by proxy (2.3). In the discussion on accuracy (2.4) I pointed out that COMPAS is affected by the problem of induction, and that what is considered accurate depends, among other things, on the decision on what is considered an acceptable margin of error (State v. Loomis, 2016; Farabee et al., 2010; Baird, 2009; Brennan et al., 2009; Skeem & Louden, 2007), which in the debate is defined according to how COMPAS performs in comparison to other risk assessment tools and predictive models applied to criminal justice (Farabee et al., 2010; Baird, 2009; Brennan et al., 2009; Skeem & Louden, 2007).

In philosophy of science, risk is discussed in relation to two aspects that are relevant for this analysis: first, the risk of being wrong about the relationship between input and output, in other words the risk of being wrong about the hypothesis that indicates what leads to recidivism; and second, the risk of being wrong about the purposes COMPAS serves. The first type of risk is called "inductive risk" (Hempel, 1965; Rudner, 1953; Douglas, 2000; Biddle, 2016) while the second type is called "epistemic risk" (Biddle, 2016). I will discuss them in turn.

Inductive risk

Inductive risk is "the risk of wrongly accepting or rejecting a hypothesis H, given a body of evidence E that is taken to support H" (Biddle, 2016, p. 1). In the COMPAS case, the hypothesis seems to be that individuals presenting a certain mix of characteristics will recidivate. When the individual for whom the prediction had been made recidivates, that event is taken as a confirmation that the hypothesis was correct, and that the mix of characteristics that was used for the prediction is what leads to recidivism. Given that no hypothesis can be proven true through evidence,[15] then the matter that philosophers of science have debated on is twofold and pertains to the question of when it is possible to confirm and accept a hypothesis: what kind of evidence confirms a hypothesis? How strong the evidence that supports a hypothesis needs to be to make the hypothesis acceptable in terms of scientific knowledge? These questions, it turns out, place inductive risk (Hempel, 1965; Rudner, 1953; Douglas, 2000; Biddle, 2016) at any point in which decisions are made, from when there are choices to be made about the decisions of what matters and what doesn't, to the interpretation of data, to the development of a methodology (Biddle, 2016). As discussed earlier (2.3) the problem of induction (Hume, 1963, as cited in Ladyman, 2002; Popper, 1959; Lakatos, 1968; Lipton, 2004) relates to the background assumptions that are not justified and are impacting on the relevance that evidence has for a specific hypothesis. The current debate not only fails to address the problem of induction when characterizing accuracy in COMPAS; it also fails to address inductive risk as part of the elements that impact on the value that is associated to knowledge in COMPAS. In addition to inductive risk, there is another type of risk that is relevant for the discussion of the value of knowledge in COMPAS: epistemic risk.

[15] The possibility of proving a hypothesis true through evidence is a matter that has been discussed at length in philosophy of science. In addition to what has been discussed in this thesis as "the problem of induction" (Hume, 1963; as cited in Ladyman, 2002; Popper, 1959; Lakatos, 1968; Lipton, 2004) which questions what can be considered evidence. Popper (1959), for example, believed that observations cannot confirm theories.

Epistemic risk

Epistemic risk is “the risk of being wrong” (Biddle, 2016). This type of risk includes inductive risk, but also considers other things that are part of the process and are not covered by the notion of inductive risk, namely, the risk of wrongly accepting a definition, and therefore a policy based on that definition (Biddle, 2016). Biddle (2016) clarifies what it means to wrongly accept a policy: “if one affirms goal G, if one adopts policy P as a means of achieving G, and if pursuing P will not lead to the achieving of G, then one wrongly accepts P” (Biddle, 2016, p. 10). Applied to COMPAS this means that if COMPAS, as Equivant and the Court believe, aims at reducing recidivism, crime and prison population, and as a consequence reduce spending for the criminal justice system, the policy consisting in the enforcement of COMPAS as an evidence-based practice in sentencing would be wrongly accepted if it does not lead to the achieving of the goal. Another way in which the policy might be wrongly accepted is if the policy goes against the ethical principles that are supposed to be respected in that context. For the example Biddle (2016) uses in his argument, the ethical principles are those indicated by Beauchamp and Childress (2012) as “non-maleficence, beneficence, respect for autonomy, and justice” (Beauchamp & Childress, 2012; as cited in Biddle 2016, p. 10). In the case of COMPAS, the epistemic risk could be the risk of being wrong when adopting the policy introducing COMPAS risk assessments, for example, in case it turned out that the risk assessment was biased in a way that is considered unacceptable in that context.

In the current debate risk is taken into consideration in terms of recidivism risk and the risk of bias in sentencing. The debate doesn’t take into account the additional risks which I indicated as inductive risk and epistemic risk. Therefore, when calculating the benefits of introducing ML in sentencing, the current debate only considers part of the risk involved, and, I would argue, only part of the costs that derive from taking those risks, which ultimately impacts on the value associated to ML as a solution for reducing costs in criminal justice.

2.7 Conclusions

The sub-question that needed to be answered in this chapter was: *how can philosophy of science contribute to the understanding of machine learning in criminal justice?*

To answer this question I drew from philosophy of science and challenged the main four elements that make up the assumption about the nature and qualities of the knowledge that is associated to COMPAS as a case of ML in criminal justice: fairness, accuracy, prediction, and risk. When it comes to fairness, ML is described as an improvement to risk assessments because it is supposed to avoid the possibility of subjective bias and values taking part in the assessment, as it is said to be the case in risk assessments performed by judges alone. However philosophy of science allows to see that the characterization of ML risk assessments as fairer than judges is inaccurate ML necessarily involves decisions about what matters and what needs to be prioritized that are ultimately value-laden (2.3) which take place during the design process of the software. The discussion of value-ladenness in COMPAS shows two things: first, that there is a logical fallacy in the characterization of bias, and discrimination in the current debate, and second, that there are inconsistencies in the characterization of COMPAS as fairer than judges at performing the assessments. Philosophy of science also indicates that ML is affected by the problem of induction (2.4), which means that accuracy is a much more complicated matter than being able to agree on an acceptable margin of error. Philosophy of science also shows that when it comes to characterizing ML risk assessment scores as predictions, there is a misunderstanding about the difference between explanation and prediction which is crucial in determining the value which can be associated to the scores (2.5). When it comes to risk (2.6), looking at inductive risk and epistemic risk shows that there are risks that are relevant and not considered in the current calculations of the added value recidivism risk reduction is supposed to bring to criminal justice.

Philosophy of science can then be said to contribute to the understanding of ML in criminal justice by providing the analytical tools needed to challenge the assumptions about the qualities of the knowledge associated to the assessments on which the current understanding is based. By challenging those assumptions, philosophy of science elucidates important critical points that need to be addressed when trying to understand the value that can be associated to that knowledge, and the implications for the achievement of the goals that knowledge is expected to serve. So far I have addressed the first of the three main assumptions which I will discuss in this thesis. In the next chapter, I will focus on challenging the assumption that COMPAS is a tool for the provision of evidence.

CHAPTER 3 – Philosophy of technology for the understanding of machine learning in criminal justice

3.1 Introduction

> "At sentencing, the circuit court is to consider three primary factors: gravity of the offense, character of the offender and the need to protect the public."
>
> (State v. Loomis, 2016, p.1)

Criminal justice seeks to improve the ability of its courts to assess their convicts for recidivism risk. The idea is that there is a truth to be known about the individual, truth that can be established through science, specifically through the adoption of statistical methods and criminological theories. COMPAS scores are considered one of several pieces of evidence the judge needs to weigh in order to make a decision on whether or not to release the convict. The scores are characterized as providing knowledge about the "criminal personality" (Equivant, 2019, p. 39) of the individual. Equivant explains that the way COMPAS works can be understood by thinking about *triage*[16] takes place in medical practice (Equivant, 2019).

The word "*triage*" comes from the French *trier*, meaning to separate, sort, sift or select. Medical practitioners use triage to make decisions about how to separate people to avoid contagion and distribute resources according to a set principle, that could be, for example, the gravity of the conditions for which the individual seeks medical attention. *Triage* requires for a structure to be put in place which indicates both the elements that are relevant for the evaluation of the patient's condition, and the options available as categories that indicate the different degrees of urgency that are configured within that medical practice. For example, blood pressure might be relevant while one's height might not. In addition to the structure, a rule (or set of rules) is necessary to indicate how an individual's characteristics compare to the standards set for those characteristics. For example, in relation to blood pressure, a certain value for blood pressure might be considered normal, while another might be considered abnormal and as such indicate the individual as sick. The structure and the rule for *triage* contribute to the characterization of the patient, who is being evaluated in comparison to a certain understanding of what is considered healthy, and in comparison to other individuals who are competing for the available resources (in this case for medical care).

In COMPAS, the separation of individuals at sentencing is connected to two types of actions that criminal justice reform indicates as a solution to the problem of the excessive costs of mass incarceration: first, the separation of individuals to avoid the low-risk ones to be negatively influenced by the company of the high-risk ones; and second, the separation of individuals for a differential allocation of resources (American Bar Association, n.d.). When it comes to looking at the impact of the assessment on the ability of the judge to make a decision, COMPAS is characterized as not

[16] Equivant says that COMPAS is "particularly useful to agencies that apply a triage strategy as part of their risk and need assessment protocol to improve efficiency and reduce workload" (Equivant, 2019, p. 32).

affecting the decision-making process if not in positive terms by providing useful and accurate knowledge.

In the current debate, COMPAS is described as a tool, and as such it is characterized in terms of what it is used for by the people involved in the sentencing process. This characterization of COMPAS is based on a certain understanding of the nature of technology (what technology is) and a certain understanding of the relation between humans and technology. These two elements are objects of study of philosophy of technology, and that is why philosophy of technology fits the purpose of this chapter. Philosophers of technology have taken many different perspectives on the nature of technology and human-technology relations. For the purpose of this thesis, I will focus on Heidegger's contribution to the subject by drawing from *The Question Concerning Technology* ([1993], 2008), which addresses both the question of what technology is and what that means for human-technology relations. Heidegger ([1993], 2008) begins by presenting the most common understanding of what technology is: either "a means to an end" (p. 4), or "human doing" (p. 13). Understanding technology this way is pretty common because it is intuitive, and it is what can be referred to as the "instrumental view" of technology. The instrumental view of technology is the view of COMPAS as technology presented in the current debate. This view, however, is inaccurate, Heidegger says, because technology is neither means to an end nor human activity. Technology is not something that provides truth but instead it is something that emerges from truth. This means that COMPAS is not a tool that does something which has been decided by humans, and it is not something that provides truth about the individual that is being assessed. What is COMPAS, if not a tool? Heidegger says that technology is in fact "a way of revealing" (Heidegger, [1993], 2008, p. 12), which applied to COMPAS would mean that COMPAS is an expression of what it emerges from, the American criminal justice system as it is. Understanding that technology is not a tool is important according to Heidegger because technology presents a danger. The biggest danger, according to him, is that humans will misinterpret what they see when they come in contact with technology (Heidegger, [1993], 2008, p. 26). The misinterpretation consists in practical terms in humans not realizing that while they think they are using the technology for their own purposes, they are instead going in the opposite direction, actually making things worse. This would be the case, for example, if COMPAS was expected to contribute to solving the problem of mass incarceration, while in fact it automates the same dynamics that led to mass incarceration in the first place. Then, the danger could be, for example, that automation could not only make the problem worse, but it could make it even more difficult to address than it is now.

In this chapter I will challenge the characterization of COMPAS as a tool, as an evidence provider, by combining a critical analysis of the analogy to *triage* and the insight provided by philosopher Michel Foucault who studied matters of knowledge in criminal justice, in particular in his book *Discipline and Punish* (Foucault, 1991). The structure of this chapter reflects a three-step analysis found in Foucault's work (1991) which consists of: first, looking at the structure of the assessment (3.2); then, looking at the rule for the assessment (3.3); and last, questioning the assessment itself when structure and rule for the assessment come together (3.4). To conclude, I will make a few final considerations on this level of analysis, I will reconnect the analysis to the non-instrumental understanding of technology just introduced, and I will indicate how to move the analysis forward in chapter 4.

3.2 The structure of the assessment

Foucault, in his book D*iscipline and Punish* (Foucault, 1991), discusses criminal justice in France and looks at detention as a way to analyze power which he refers to as "discipline". The study of the prison is for Foucault a way to investigate the connection between knowledge and forms of social control. He sees this process of managing people as entailing the presence of two main elements: a structure in which to place the individuals that allows for them to be arranged in a specific way, and rules to determine how those individuals will be arranged. What the structure of the prison shows, in fact, is that the organization of an "analytical space" (Foucault, 1991, p. 143) takes place, which is the organization of a delimited space that allows for the identification, sorting and analysis of individuals according to certain parameters. For the sorting process to be possible, the individuals need to be characterized according to the categories that match the structure they will be placed in. That characterization is made possible by the deployment of "documentary techniques" (Foucault, 1991, p. 191), which could be, for example, the filling out of paperwork requesting biographical information about the individual. That information is then stored in what takes the form of "cumulative systems" (Foucault, 1991, p. 190), such as archives, in a way that allows for the information to be cross-referenced between individuals and at the same time between groups that are identified in relation to those characteristics deemed relevant for the purposes to which the information is gathered and stored. Foucault sees all of these processes in a specific kind of structure, a form of architecture designed in the 18th century by philosopher and social theorist Jeremy Bentham: the Panopticon (Bentham, 1995).

The Panopticon is a prison that takes the form of a circular building where there is watchtower at the center of the building and the cells for the convicts all around it. In the Panopticon, bodies are gathered and sorted in order to be studied and judged. The cells separate each individual from the others so they cannot see one another. Moreover, the convicts cannot see whether there is someone in the central tower, so they are always in doubt on whether or not they are being watched. Because of this uncertainty, convicts behave as if they are always being watched, and become what Foucault describes as "the principle of their own subjection" (Foucault 1991, p. 203), meaning that those that are subject to the panoptic structure cannot place responsibility for their condition of submission onto something (or someone) other than themselves. This mechanism that sees individuals conditioned in their behavior, is what automation is all about, as there is no need for actually providing surveillance at the center of the panopticon, since individuals themselves already behave as if surveillance is already in place. This condition in which convicts find themselves is made possible by two elements, according to Foucault (1991): for the individuals to be placed in a condition of "formal equality" (p. 189), and for them to be considered individually (p. 182). Formal equality can be seen in the individuals being in the same conditions of all the others gathered in the same space, and at the same time they are considered as individuals because they are each in their own cell, fully responsible for their condition. While in Bentham's Panopticon the structure that encloses and partitions is an actual building, in the case of COMPAS the structure cannot be seen as such. However, the structure that COMPAS presents has the same characteristics of a panoptic structure. Similarities between the two can be seen especially when COMPAS assessments are studied through the analogy to *triage* in medical practice, as provided by Equivant (2019).

COMPAS and panopticism

In medical practice, the structure of *triage* requires: first, the gathering of patients in a delimited space (for example a clinic or a hospital); and second, categories and criteria for the identification and classification of the individuals. In that context, the practices of gathering in a delimited space and sorting people according to certain criteria allow for managing both populations and groups (such as the ill in general, or according to specific diseases) and individuals (because each person is a medical case of his/her own). In that characterization, individuals and groups are organized in hierarchical order according to a predetermined principle (e.g., urgency). The information of the patients is gathered in some sort of written form that is then stored for future reference. Patients have little chances of successfully questioning the evaluation performed by the doctors, and cannot challenge their status neither in relation to the level of urgency that is assigned to them individually based on their characteristics, nor in comparison to the urgency of others in terms of the priority with which they will have access to medical care.

COMPAS structure appears to have the same characteristics as *triage* in medical practice. In chapter 2, I discussed how the software is designed (2.3) according to the categories and types of information that are deemed relevant to the assessment of recidivism risk, given the data available through the statistics that pertain to criminal justice. Criminological theories are the basis for identifying individuals according to their differences (Garland, 1992; O'Malley & Valverde, 2014), and sorting "for the purpose of assessment" (Lyon, 2005, p. 20) and judgment. Identification and sorting are activities that characterize a panoptic structure. COMPAS involves the gathering of data through a questionnaire, and the registration of that data in digital form for the assessment in a system that accumulates that information (for example in databases, datasets, and servers). COMPAS allows for both precision and flexibility as it addresses the "unique individual" (Equivant, 2019, p. 51), and at the same time it compares the individual to a generalized group of offenders that share with the individual certain characteristics.

Understanding that COMPAS presents a panoptic structure is important because it indicates that while in the current debate switching from risk assessment performed by the judge to risk assessment performed by COMPAS is characterized as a positive change, in practice that might not be the case. My analysis in the previous chapter indicated that COMPAS is value-laden (2.3), and that the assessment based on socio-economic factors might represent the possibility for COMPAS to be operating on the basis of bias, even if just by proxy. This means that the characterization of COMPAS as fair, or at least fairer than the judge, does not seem to find justification in practice. Moreover, the panoptic characteristics of the structure that is found in COMPAS indicate that the convict entering the "analytical space" (Foucault, 1991, p. 143) of the assessment has little chances of challenging the process through which the score is provided, just like a patient has little chances of challenging the assessment performed by a doctor at the emergency room. The element that determines the condition in which the patient, or the convict, is unable to challenge the assessment is not the lack of access to information; it is the structure that makes individuals, to say it in the words of Foucault, "the principle of their own subjection" (Foucault, 1991, p. 203).

In reference to *State v. Loomis*, this means that the dismissal of Loomis' claims (State v. Loomis, 2016) about the impossibility of challenging the assessment due to an asymmetry in the availability of information between himself and the judge, algorithmic opacity and lack of access due to the proprietary nature of COMPAS, tells only part of the story about how COMPAS might impact the ability of convicts to challenge the assessments. The Court in the Loomis case dismissed the claim Loomis made about not being able to challenge the assessment by indicating that both the judge and Loomis had the same information available. According to the Court this meant that there was no asymmetry and that therefore Loomis was in fact in the position to challenge the assessment. The panoptic features of COMPAS indicate, however, that characterization of the conditions Loomis was in, is inaccurate. It also indicates that because the problem does not seem to be limited to the impossibility of having access to information, eliminating the problem of opacity or the proprietary nature of the software might not increase the chances of convicts to challenge the assessment. In chapter 2, I discussed how a question of whether or not it is acceptable to have gender as a discriminatory factor in the assessment was discarded on the basis of a system that prioritizes statistical accuracy. While the right to be assessed in an accurate manner is prioritized, how that accuracy is reached might go to the expenses of the convicts having respected their right not to be discriminated on the basis of race or gender, even if just by proxy. When considering the panoptic structure of COMPAS, it becomes clear that also the right to due process, in the sense of having the possibility of challenging the assessment, is threatened in the process of prioritizing statistical accuracy.

In the next section I will move the analysis forward by looking at the second element that is part of COMPAS assessment as a form of *triage*: the rule on which the assessment is based. The analysis in the next section, combined with the one just performed, will allow for challenging the main assumption that is the focus of this chapter: the understanding of COMPAS as a risk assessment tool for the provision of evidence.

3.3 The rule for the assessment

Once again I will turn to Foucault for insight. Foucault discusses how criminology provided a way to explain the crime, but also to judge the quality of an individual. Criminology looks at crime as a form of deviance from what is considered normal in society. What is considered normal is then used as a 'norm,' meaning as a term of comparison for assessing the individual (Foucault 1991).

Comparing to a norm, Foucault tells the reader, is different to comparing to the law. The first fundamental difference is that while the comparison to a law involves a judgment of either compliance or not compliance, comparison to the norm, what Foucault describes as "normalizing judgment" (Foucault, 1991, p. 177), is a form of judgment by comparison that involves processes of homogenization, differentiation, and hierarchical distribution (Foucault, 1991). The characterization of the individuals is done in a away that identifies the individual only according to the characteristics that are deemed relevant, reducing them to a standardized version of themselves that matches one of the possible combinations of what is considered criminal. The individual is then assessed and classified according to how (s)he compares to the general idea of normality. The individual is therefore seen and assessed in a context of "formal equality" (Foucault, 1991, p. 189), as the characteristics that are relevant and the criteria for the assessment are the same for all the individuals, and so are the standards to which each individual is compared. The comparison to the norm "homogenizes" (Foucault, 1991, p. 183) because it standardizes in comparison to an idea of what it is to be normal, while at the same time it differentiates between individuals, because the information is taken about each individual separately. The comparison to the norm is not followed by a judgment of either being normal or abnormal. Instead, the individual is classified, ranked, according to how (s)he compares to others that are (or have been) assessed in the same context. The individual is, in other words, placed on a continuum between the two poles indicating the ideal normal and the extreme abnormal. The individual, in addition to being judged by him/herself is therefore also placed in a relative position to others (Foucault, 1991). This imposition of a homogeneous model, does not however get in the way of the possibility to consider an individual in isolation, as " within a homogeneity that is the rule, the norm introduces as a useful imperative and a result of measurement, all the shading of individual differences" (Foucault, 1991, p. 184). Equality and individualized judgments are not then to be considered positive traits of an unbiased and humane system; they are part of the factors that determine the condition of subjection of individuals in a panoptic structure (Foucault, 1991). In other words, equality and individualized judgments are part of necessary conditions that make it possible to automate a process that places the individual the position of subjection characterized by the structural impossibility of challenging the assessment (Foucault, 1991). This matters, Foucault tells the reader, because placing the individual on a continuum has an impact on the way criminal justice characterizes guilt.

Foucault recalls how in the eighteenth century there were many different types of evidence considered in criminal justice, just to name a few: the legitimate proof of the witness, the artificial proof made by argument, and the semi-full proof, which is proof not successfully proven wrong by counter evidence provided by the accused (Foucault, 1991). The interesting thing about this variety of evidence, he writes, is that the arrangement of those pieces of evidence in combinations is done according to "arithmetical rules" (Foucault, 1991, p. 36) that are known only to the experts that participate to that evaluation. Moreover, this is a process that finds full proof in partial proofs (Foucault, 1991), and makes it possible for different (combinations of) evidence to lead to different "judicial effects" (Foucault, 1991, p. 36). What happens with combinatorial evidence is a shift from a binary characterization of guilt, to the characterization of guilty as a matter of degree (Foucault, 1991). What can this possibly mean for the COMPAS case?

COMPAS, equality, individuality and automation

In *triage*, the rule that is applied is a comparison to what is considered healthy, which is then the norm to which all patients are measured against in order to determine, for example, their position in the hierarchy of coded levels of urgency (e.g., red for maximum urgency). Their code will then determine whether (and in which order) they will access the resources needed for treatment. The norm in this context, which is the rule that is used for judgment, is often characterized as a positive addition to the processes in which it takes part because it is supposed to equally assess all individuals without difference, and at the same time it is supposed to allow for a type of judgment that is specific for each individual. This understanding of equality and the individualized assessment are the two characteristics that the analogy then transfers onto the COMPAS case.

Equivant describes COMPAS as providing a score by comparing the characteristics of an offender to a representative criminal population. The criminal population is therefore taken as a reference, and represents the so-called "norm group" (Equivant, 2019, p. 5) which consists of former recidivists and it is supposed to indicate what characterizes an individual who recidivates. This means that the "norm group" (Equivant, 2019, p. 5) is already a construct based on an idea of what constitutes a "normal citizen" (Equivant, 2019, p. 54). What is considered normal, and which characteristics are deemed relevant, is based on the criminological theories and statistics about crime on which COMPAS is designed (2.3). Those characteristics are socio-economic factors and events in one's life that are beyond the control of the individual in relative terms at the time of the assessment, or in absolute terms because of the nature of the characteristic itself. The individual is reduced to a characterization of himself according to the categories deemed relevant when it comes to crime, and more specifically recidivism. A convict is not judged either criminal or not criminal; s(he) is placed on a continuum that goes from high to low risk according to the distance from the norm that has been associated to the combination of traits the individual presents. The individual is judged according to standards and placed on a continuum, but at the same time, thanks to the combinatory nature that is given to the criminal character of the individual, each person is also ranked according to the relative position (s)he acquires in relation to others, in a way that covers all possible combinations that being considered abnormal can take.

In this respect, equality is not the opposite of bias and discrimination, instead it is a necessary feature of the system that reduces individual differences to predetermined categories, a system that organizes individuals for the purpose of judging them according to a certain norm (Foucault, 1991; O'Malley and Valverde, 2014). The comparison to a norm is a comparison to a standard (Rose & Valverde, 1998). Similarly, individuality in judgment, even though it is described with the same terms, does not seem to be the same thing as the one intended by law as a right for the convict to be considered as an individual by the Court; it is one of the elements that determine the convict's condition of subjection. Both equality and individualized judgment are not positive traits of more just criminal justice system; they are necessary features of a system that tends towards automation. Automation, in turn, is not a solution to the problem of the subjective character of the judge's assessment; it is the feature of a process that involves the judgment of individuals as abnormal and dangerous because of events in that took place in their life they had no control over, or because of how they compare to certain socio-economic standards.

Now that I have addressed the structure (3.2) and the rule for the assessment (3.3) separately, it is time to see what Foucault's insight can contribute to the understanding of ML in criminal justice when the two come together.

3.4 Questioning the assessment in action

The combination of the panoptic structure (3.2) and the comparison to the norm (3.3) shows that there is a process in place that entails for individuals to be included and at the same time excluded from a certain "social body" (Foucault, 1991, p. 184). Inclusion is a form of "membership of a homogeneous social body" (Foucault 1991, p. 184) by matter of degree according to how close to the definition of normal each individual is found through the assessment. Exclusion is due to individual differences that are still being measured, so while everybody is equal within the enclosed space in which they are being assessed and belongs to a certain social group, they can, at the same time, be placed outside of another social ensemble.

Early on in criminal justice, Foucault tells the reader, when torture was still an acceptable practice and death penalty was also more often than not publicly executed. The public execution of the sentence served as a reminder for those watching that the law was being enforced and that crime was not tolerated. In other words, public display of the carrying out of a sentence served purposes of prevention of additional crime (Foucault, 1991). Preventive practices are directed especially at all the "potentially guilty" (Foucault, 1991, p. 108), and focus on crime that has not yet been committed. Targeting the "potentially guilty" (Foucault, 1991, p. 108) is made possible by a double shift that took place in criminal justice: first, a shift from criminal justice's interest in crime to interest in the criminal; and second, a shift from criminal justice's concern for committed crime to concern for future crime (Foucault, 1991). Moreover, when it comes to the preventive character of criminal justice, Foucault sees the same features associated to a panoptic structure in criminal justice and in the management of what he refers to as "the plague" (Foucault, 1991, p. 197). In the case of the plague, the available space is divided to isolate the town and contain the spreading of the disease, and a registration system is enforced to allow monitoring the conditions of the town. In that context, people are assessed and sorted according to the characteristics deemed relevant for the effective management of the situation. Although the preventive measures seem to be directed first and foremost at those already affected by the disease, in fact, those measures are being directed at the whole population. In addition to the preventive character of the process that take place in criminal justice, Foucault (1991) sees in those processes the goal of reducing the "inefficiency of a mass phenomena" (p. 219) which indicates that the management of the plague has an economic function consisting in a cost-benefit analysis that is taken into account for decision-making about how to manage the situation.

COMPAS, prevention and the distribution of resources

Triage in medical practice indicates that such an assessment has two purposes: having a system for prioritizing the allocation of resources, and preventing the spreading of diseases. Depending on the availability of resources, *triage* might also involve the need to make a decision about who will not get access to the scarce resources. That happens for example when *triage* is part of operations in conditions of emergency. In those cases, another aspect of assessment like *triage* emerges, which is that *triage* entails "probability estimates of outcome and benefits and burdens to the individual and to the system" (Hartman, 2003), estimates that support the decision-making process that needs to determine who will access the scarce resources available. In other words, *triage* involves a process of inclusion and exclusion, based on a series of principles that determine who will get what. *Triage* usually bases the assessment on one of three schemes: the utilitarian scheme, the clinical scheme, or a third scheme that brings some features of both together (Barilan et al., 2014). The utilitarian scheme is characterized by a well-defined structure and criteria for sorting people according to the priorities set in a specific context. Barilan et al. (2014) characterize the utilitarian model of *triage* as on in which the goal is to "maximize a pre-defined medical good among all relevant persons" (p. 53). The clinical scheme is less structured compared to the utilitarian scheme. Medical practitioners operating according to the clinical scheme do not aim for the maximization of health and because of that do not rank individual in comparison to a standard that places them in a position that is a relative position to other individuals; instead, practitioners operating according to the clinical scheme address each individual on a first come-first served basis (Barilan et al., 2014). The hybrid scheme brings the previous two together and allows for more flexibility in judging if prioritizing efficiency like in the utilitarian scheme, or the order of arrival at the medical facility like in the clinical scheme (Barilan et al., 2014). Every choice among these (or other possible) ethical frameworks for the allocation of resources presents challenges, possible downsides and consequences. For example, *triage* based on utilitarian principles might lack the means to measure its impact because "instruments of monitoring, follow-up and quality control" (Berilan et al., 2014, p. 53) are not available. Or again, it could be addressing the maximization of something that is not measurable (or hardly so), such as "the alleviation of suffering" (Berilan et al., 2014, p. 54). Utilitarian *triage* might be maximizing health but in a way that undermines and threatens other aspects that might be important such as the freedom an individual to choose which medical treatments to be subjected to. The characterization of this last aspect of *triage* will depend on the societal and cultural context in which it takes place.

In addition to providing a system for the distribution of resources, *triage* is also considered useful in medical practice because it serves a preventive function. *Triage* allows for keeping patients with different diseases separate so they do not infect one other, and it allows for the identification of who needs to be kept for treatment so that they don't infect others outside the hospital. This means that the assessment that leads to the decision of who will have access to the resources is also connected to the idea that the decision needs to prioritize the benefit of those that are considered healthier. This indicates that what is considered a cost and what is considered a benefit is based on the prioritization of the interests of some individuals over others, prioritization that is based on specific ethical principles, such as utilitarianism.

COMPAS scores are supposed to bring value to sentencing by improving the way decisions are made about who is released and who is detained, which indicates that COMPAS risk assessments present the same type of functions of *triage* seen in Foucault and in medical practice: prevention and prioritization in the distribution of resources. The prevention and reduction of crime indicates that COMPAS is connected to the so-called "need to protect the public" (State v. Loomis, 2016, p. 1). The interest of criminal justice for crime that has not yet taken place raises questions in relation to who can be sure to be included in that public that needs to be protected. Prevention is indicated in relation to the aim of criminal justice to reduce crime in two ways: first, in the sense of avoiding the placing of high-risk and low-risk individuals together, as detention tends to worsen the dangerousness of the least dangerous who get to spend time with the more dangerous ones (American Bar Association, n.d.); and second, in the sense of preventing future criminal activity that might be committed by those believed to be among the most prone to committing it (e.g., the high-risk individuals). This function of avoiding "contagion" between high and low-risk convicts it is the same function *triage* in medical practice has for the prevention of contagion among patients, and it is consistent with the characterization of recidivism as disease (I discussed this matter in chapter 2). The preventive function of COMPAS shows a concern in criminal justice for future crime, which seems to indicate that, as in the case discussed by Foucault, prevention targets first and foremost the "potentially guilty" (Foucault, 1991, p. 108). Moreover, the example of *triage* in medical practice shows that protection, safety and security are concepts that can vary greatly depending on the socio-cultural context. This means that what happens in criminal justice is not independent of what happens outside of the perceived perimeters separating the convicts from everybody else. It also means that the dynamics that can be recognized in COMPAS risk assessments might have an impact that goes beyond the effects envisioned in the current debate.

The other function that can be associated to COMPAS is connected to the declared goal of criminal justice reform to reduce costs. This means that COMPAS is understood to be serving the purpose of improving the prioritization in the distribution of resources. As Foucault (1991) explained, the assessments serve the purpose of reducing the "inefficiency of a mass phenomena" (p. 219), inefficiency that is understood in economic terms, and that involves the prioritization of the distribution of resources according to what seem to be utilitarian ethical principles (Garland, 1992). This aspect is clearly declared when it comes to the purpose of having ML risk assessments in sentencing, and it is a purpose that goes beyond the sentencing process itself. The characterization of COMPAS as a risk assessment tool for the provision of evidence does not seem to do justice to the dynamics that take place when COMPAS risk assessments are involved. COMPAS presents a process which is a cost-benefit analysis based on an assessment that judges individual's worth according to how their lives measure up to certain socio-economic standards. In the conclusions that follow I will bring together the different elements discussed so far, and reconnect them to the perspective on technology introduced at the beginning of this chapter.

3.5 Conclusions

The sub-question that needed to be answered in this chapter was: *how can philosophy of technology contribute to the understanding of machine learning in criminal justice?*

To answer this question I began by looking at the characterization of COMPAS as a tool for the provision of evidence. That characterization touches upon two important aspects that pertain to the field of study of philosophy of technology: the nature of technology and human-technology relations. For the purpose of this thesis, I decided to focus on Heidegger's perspective on the subject as explained in *The Question Concerning Technology* ([1993], 2008), which I briefly introduced in at the beginning of this chapter. The intention of that brief presentation of his work was to show why challenging the assumption of technology as a tool is important, and what doing so can contribute to the understanding of ML in criminal justice. Drawing from Michel Foucault's work in the book *Discipline and Punish* (1991), I articulated my analysis of COMPAS in three steps that reflect three elements studied in his book:

- First, I looked at the structure of the assessment and argued that COMPAS presents a panoptic structure (3.2). My analysis shows that philosophy of technology can contribute to the understanding of how ML impacts on the right of convicts to be able to challenge the assessments.
- Then I analyzed the rule for the assessment (3.3) and discussed normalization (Foucault, 1991; O'Malley & Valverde, 2014) in COMPAS assessments. My analysis showed that equality is not the opposite of a biased and subjective judgment, but it is instead a feature of ML which is necessary for the purpose of judging individuals in a panoptic structure. Similarly, individuality in judgment in ML does not reflect the right for the convict to be considered as an individual by the Court, but it is instead one of the elements that contributes to his/her condition of subjection and impacts on his/her possibility of challenging the assessment.
- Last, I looked into what happens when the panoptic structure and normalization come together. My analysis shows that ML for risk assessment entails a cost-benefit analysis that prioritizes the distribution of resources according to what seem to be utilitarian ethical principles (Garland, 1992).

Philosophy of technology can contribute to the understanding of COMPAS by providing the analytical tools to challenge the assumption that COMPAS is a tool for the provision of evidence. Challenging that assumption shows that ML assessments present a structure, dynamics and a rationale that is still the same that was in place before ML was introduced in criminal justice, and that would be in place without it. This is what Heidegger ([1993], 2008) indicated when saying that technology is not a tool but "a way of revealing" (Heidegger, [1993], 2008, p. 12), and that a misunderstanding on this point presents a danger. The danger can be seen in the potential for subjection, discrimination and exclusion that emerges when considering the mismatches between the current characterization of ML assessments provided in the debate and can be seen when looking at the assessments in practice.

So far I addressed the shortcomings of the current characterization of ML in criminal justice when it comes to the understanding of the knowledge provided by the assessments, by drawing from philosophy of science (in chapter 2), and the shortcomings of the current characterization when it comes to the role those assessments are supposed to play in sentencing (in this chapter). In the next chapter, by drawing from philosophy of criminal justice, I will address the third and final assumption that I will discuss in this thesis: the assumption that ML contributes to reducing criminal justice costs and solving the problem of mass incarceration.

CHAPTER 4 – Philosophy of criminal justice for the understanding of machine learning in criminal justice

4.1 Introduction

COMPAS is supposed to contribute to the effective and efficient allocation of resources, because of the goal of criminal justice to reduce costs associated to mass incarceration. What "resources" mean in this case is financial resources. However, that discourse only addresses the economics of criminal justice in terms of costs, and does not address revenues, which are however also part of the economics of criminal justice. Considering only the costs hides the dynamics that relate to half of the process and need to be considered if the economics of criminal justice are to be addressed. Considering the revenues allows for commenting on two important aspects: the characterization of the relationship between COMPAS and the distribution of resources in criminal justice, and the characterization of the relationship between COMPAS, criminal justice and mass incarceration. In order to show why these two aspects matter for the understanding of COMPAS, I will draw from philosophy of criminal justice as a way to address the assumption about the contribution COMPAS is expected to give to reaching the goals of criminal justice reform. The reason for choosing philosophy of criminal justice is that it is the discipline concerned with criminal justice in a way that involves political and ethical considerations. COMPAS has been linked to political and ethical matters in the previous chapters, especially when discussing bias, discrimination and fairness (chapter 2), but also in relation to inclusion and exclusion from what is considered "the public" that needs to be protected (chapter 3). For the purpose of this chapter, I will focus on the work of Prof. Angela Davis (2003) who studied prisons, criminal justice reform, and the connection between the two in the United States. I chose her perspective on this subject because I find it to be very fitting for the analysis of COMPAS both for the adherence to the topic and for the geographical dimension of this analysis. Davis (2003), in her book *Are Prisons Obsolete?* challenges the assumption that incarceration is a necessary part of criminal justice, and she challenges the idea that criminal justice reform needs to be about prison reform. In her analysis she focuses in particular on gender and discusses individual rights and the profit dimension of criminal justice practices. Davis (2003) tells the reader that part of the reason for the acceptance of prisons as something that "works," something desirable and necessary, must have something to do with the relief knowing that criminals are physically separated from society. Moreover, Davis point out that looking at the costs of the justice system and what she refers to as "the prison industrial complex" (Davis, 1995; as cited in Davis, 2003, p. 12) without taking into account the larger economic and financial realm in which criminal justice operates, is leaving aside a crucial piece of the puzzle. Both Davis (2003) and Foucault (1991) discuss the economic value that is brought by these practices, economic value that cannot be understood without considering the role of the private sector that takes part in these operations. This chapter is organized in two parts in which I discuss: first, the economics of criminal justice (4.1), and then, detention (4.2).

4.2 The economics of criminal justice

Davis argues in favor of the abolition of prisons and in favor of a strategy that looks for new ways of thinking about criminal justice, ways that do not rely on incarceration (Davis, 2003). She asks the reader why we take prisons for granted, and she points out that there seems to be a mechanism in place that sees people that have never been in direct contact with the reality of prisons as thinking of it as disconnected from their own life. Prison then becomes something that is only relevant for the "other" that is identified with the "criminal". The prison, Davis writes, "functions ideologically as an abstract site into which undesirables are deposited" (Davis, 2003, p. 16). Those deemed undesirable by society are gathered and relegated in a place that is separate from, and more often than not literally out of sight for, the rest of society. Davis (2003) writes that it is important to remember that "imprisonment was not employed as a principal mode of punishment until the eighteenth century in Europe and the nineteenth century in the United States" (p. 42). Knowing that detention wasn't always part of criminal justice practices allows for considerations on the processes through which it has become part of those practices, which also open to the possibility of questioning the current characterization of criminal justice practices themselves.

When discussing detention, Davis uses the term "prison industrial complex" (Davis, 1995; as cited in Davis, 2003, p. 84), which was first introduced by historian Mike Davis who studied the political and economic dimensions of detention in California's penal system (Davis, 1995; as cited in Davis, 2003, p. 84). This notion goes beyond the focus on the criminal individual that is the focus of the mainstream narrative about crime, and takes into account the economics of detention. Detention as punishment is often calculated in terms of time, and sometimes in terms of money (for example when bail is set). It is no coincidence that this calculation took form in criminal justice at the same time when such calculations took place in relation to labor. As Davis (2003) observes "the computability of state punishment in terms of time-days, months, years-resonates with the role of labor-time as the basis for computing the value of capitalist commodities" (p. 44). Initially, the commodity in prison was labor, but as the economy became less industrialized, the commodity in prison became the body of the prisoner from which value is extracted by the industries that profit from the management of the correctional facilities themselves. More specifically, it is a value to be extracted from those individuals that are considered not productive, a burden to society, or how Davis characterizes them, "human surplus“ (Davis, 2003, p. 91). In my understanding, that means that convicts are considered unproductive in terms of participation to the general economy because they neither work nor consume, and detention is a way to make them productive. What can be defined as the "prison industrial complex" (Davis, 1995; as cited in Davis, 2003, p. 12) indicates that criminal justice reflects dynamics for the pursuit of profit that have detained bodies as the source.

Davis indicates that one feature that characterizes criminal justice economics is that there has been a "privatization and corporatization of services that were previously run by government" (Davis, 2003, p. 91). Just to provide an example, there are 78 businesses listed[17] as providers of goods and services, and therefore profiting from detention practices, in the State of Wisconsin alone. Criminal justice and detention might represent a cost for the state (if that's where the money comes from), but that money is profit for the companies and businesses providing all the goods and services required for the system to run. Davis recalls how these dynamics of privatization also can be seen in medical practice, or what she calls "the medical industrial complex" by recalling the work of Doctor Arnold S. Relman who wrote about it almost 40 years ago (Relman, 1980). It is not a coincidence that criminal justice is often compared to medical practice.

COMPAS and criminal justice costs

The connection that is traced between convicts and the economics of criminal justice is that convicts represent a cost. Criminal justice has limited resources that need to be optimized according to need, meaning that only those that "need" detention the most should be detained, while the others should be released and left be cared for by other services available outside of the detention facilities. As mentioned earlier, this narrative does not take into account that what is a cost for someone is revenue for someone else. This applies to costs in criminal justice as well, and I believe it is what Davis refers to when she discusses the dynamics of the "prison industrial complex"(Davis, 1995; as cited in Davis, 2003, p.12) and points out that in criminal justice the detention is a source of profit. This also reminds me of what Foucault refers to when he says that "the accumulation of men and the accumulation of capital" (Foucault, 1991, p. 221) cannot be considered disjoint. Considering that where there are costs there are also revenues allows to see that the cost-benefit analysis that can be seen in ML in criminal justice is an analysis that serves the pursuit of profit and it is likely to be the same cost-benefit analysis that led to the current conditions of the criminal justice system in the first place.

In the current debate risk assessments are described as a positive part of cost reduction in criminal justice (Tallarico et al., 2012; Klingele, 2016), but the current debate does not consider that there are costs that can be associated to the reduction of risk itself (Simon, 1987; Kemshall, 2003). These costs are not acknowledged in the current debate but do have an impact on the understanding of the added value of having ML in criminal justice. It is important to acknowledge the costs of risk reduction because when calculating the costs and benefits of risk reduction, a trading of risks (Bannister, 2005) takes place. The trading can be seen in the lack of concern for the questions raised when it comes to inductive and epistemic risk in the debate. The costs that are often not accounted for when it comes to calculations of the benefits and costs of reducing risk are non-financial costs (Bannister, 2005).[18] Those costs could be, for example, the social impact of having a criminal justice

[17] The list is provided by corrections.com, a website providing, among other things, lists of vendors working for corrections, by state. http://corrections.com/vendor/result_by_state?page=1&state=WI Last visited on July 27th, 2019.

[18] Bannister (2005) recalls the work of B. Schwartz in the book *The Paradox of Choice* (Schwartz, 2004) and discusses how many non-financial costs of risk reduction are often left aside in the assessment of whether or not risk reduction is desirable and for what reasons.

system that perpetrates a disparity in the recognition of individual rights on the basis of socio-economic factors. Ultimately, this perspective on the economics of criminal justice raises questions about whose costs criminal justice is aiming to reduce, and whose costs might increase in the process. I'll discuss this further in the next section within the discussion on detention.

4.3 Detention

Davis recalls how there was a gender difference in the way detention served the function of punishment, and how that difference had to do with the recognition (or lack thereof) of individual rights. Detention, as the deprivation of freedom, requires for the right to that freedom to be recognized to the individual in the first place. Moreover, in case detention is considered punishment for the infringement of the rights of others, those rights also need to be recognized, for detention as punishment to be possible. In light of this consideration, Davis indicates that in the past, for example, because women were denied public status, they were not often detained. Once women were rendered equal to men in their independent role in society, they were available for detention as punishment as much as men. However, a new difference emerged within detention between men and women: men were recognized a set of rights and liberties that they temporarily gave up but could eventually recover. Because of that framing of men's rights and liberties, "male punishment was linked ideologically to penitence and reform" (Davis, 2003, p. 69). On the other hand, women were not recognized those rights and liberties in the first place, which means that they were considered "with no possibility of salvation" (Davis, 2003, p. 70). This difference took extreme turns when, for example, eugenics became involved in the scientific approach to the individual, and when a certain category of people was considered genetically inferior, detention became a way to isolate the individuals deemed inferior so for example in the case of women, they could not have children (Zedner, 1995; as cited in Davis, 2003). That is an example of the turn from detention for rehabilitation to detention for incapacitation, or, in other words, from an approach to detention with a priority of getting people to be fit again for re-entering society, to an approach to detention that considers containing the high-risk individuals and preventing them from reentering society for as long as possible (Reichman, 1986; Pratt, 1995). The reason I think it is useful and interesting to keep in mind Davis' insight into individual rights and detention practices is because it offers an opportunity to look at what happens in the criminal justice system, and what that might mean for the society in which that system operates. Moreover, it offers an opportunity to address COMPAS in relation to criminal justice policy and the discussion on whether detention should be based on the concept of rehabilitation or incapacitation (Davis, 2003). The example that the author proposes is useful because it shows that disparity of treatment in the system is connected to a disparity in the status of the individuals in terms of the recognition of their individual rights at societal level. In this section, I will use the insight provided by Davis (2003) to see what the treatment within the system can tell about the recognition of individual rights in society in the COMPAS case. The connection between Davis' example and COMPAS is that access to the resources that are associated to the improvement of the individual in terms of being fit for society is based on discriminatory criteria, that criteria being gender (in Davis' example) and socio-economic conditions (in COMPAS).

COMPAS, the individual rights of prison population, and socio-economic factors

The analogy to medical practice frames recidivism risk assessment as a way for the judges to be able to decide which cure is the most effective for each convict, the options being incarceration or release. Now, even the current debate recognizes that it is doubtful whether incarceration makes convicts less dangerous (American Bar Association, n.d.). For the purpose of this analysis, I will therefore consider that the resources that contribute to the rehabilitation of an individual are found outside of the detention, and therefore correspond to the granting of the release to the convict that is assessed as eligible for parole. Another clarification is necessary in relation to what will be referred to as "individual rights". Davis (2003) writes that the difference between men and women in relation to criminal justice and prison was that men committing crime were considered as having temporarily given up their "rights and liberties" (Davis, 2003, p. 69), and therefore were considered as capable of redemption through self-reflection, study, work and other activities that were therefore made possible through the provision of certain resources in prison. Women, on the other hand, were not "acknowledged as securely in possession of these rights" (Davis, 2003, p. 70) and therefore were not framed as participants in a process of redemption similar to that of their male counterparts. Davis does not specify what rights exactly she is referring to, and the discussion of what individual rights are concerned when it comes to COMPAS and criminal justice is a very complex matter that I will reserve for another time. Therefore, for the purpose of this argument, I will refer to individual rights as a general category of rights that it is plausible to associate to an individual that is not being detained.

Davis (2003) begins with observing that men and women in prison are treated differently when it comes to their access to the available resources for rehabilitation. In the COMPAS case, the observation that matches Davis' is that low risk and high-risk convicts are treated differently in relation to their access to the available resources for rehabilitation. In Davis' example, the criterion for discrimination is gender, while in the COMPAS case it is the level of risk, which is based on socio-economic factors. The result of that differential treatment, in Davis' example, is that men are provided the means to rehabilitate, while women are not. In the COMPAS case, low-risk convicts are released and therefore given access to resources available for their rehabilitation, while high-risk convicts are kept in prison and therefore denied access to the resources available for rehabilitation. Davis then provides an interpretation of what she sees, and writes that what that differential treatment reveals is that men are considered capable of rehabilitation, and because of that, their deprivation of individual rights is only temporary. In other words, men are considered as *having done* something wrong. On the other hand, women are not considered capable of rehabilitation, and because of that their deprivation of individual rights is terminal. In other words, women are considered as *being* wrong. What Davis' interpretation points to, when applied to COMPAS, is that convict labeled as low-risk (which means convicts having 'normal' social and economic characteristics) are considered as capable of rehabilitation and/or deserving access to the resources that come with the recognition of their individual rights. In other words, low-risk convicts are "wrong" only to a degree that is still considered fit for society. On the other hand, convict labeled as high-risk (which means convicts having ‘abnormal’ social end economic characteristics) are considered incapable of rehabilitation and/or undeserving of access to the resources that come with the recognition of their individual rights. In other words, high-risk convicts are “wrong” to a degree that is considered unfit for society.

What is then the meaning of this analysis when it comes to individual rights? Davis’ (2003) identifies that women were found to have less individual rights than men at societal level, and that the positive characterization of equality for men and women, in terms of being independent subjects, needs to be reassessed considering that equality is a necessary condition for women to be eligible for detention. In the COMPAS case convicts are characterized according to how they measure up to certain standards for chosen social and economic conditions. Individuals who score poorly in those socio-economic measures are considered having less individual rights than those that are in better socio-economic conditions at societal level. This analysis contributes to what I introduced in chapter 3 as the question of whom is included in the notion of public that needs to be protected. Machine learning in the COMPAS case shows that the process of inclusion and exclusion is based on socio-economic factors and it is connected to the recognition of less individual rights to those that are in more difficult socio-economic conditions. This framework and the processes that derive from it reflect the conditions that have led to mass incarceration in the first place (Klingele, 2016; Kehl & Kessler, 2017).

4.4 Conclusions

The sub-question (3) that needed to be answered in this chapter was: *how can philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?*

To answer this question I looked at the characterization of COMPAS as a way to reach the goals of criminal justice reform. My analysis consisted of two parts: the first (4.1) focused on the economics of criminal justice, as a way to challenge the assumption made in the current debate about the connection between COMPAS assessments and the reduction of criminal justice costs; the second (4.2) focused on detention, as a way to challenge the assumption about the connection between the assessments and a solution to the problem of mass incarceration.

Considering the economics of criminal justice allows to see that when calculating the costs and benefits of risk reduction, a trading of risks takes place (Bannister, 2005). In addition to inductive risk and epistemic risk, that I argued are important risks that go unacknowledged in the current debate (chapter 2), there are also the risks associated to the decisions made about what constitutes a cost and what doesn’t. Specifically, some of the costs that are not considered in the current cost-benefit that takes place according to the debate are the costs of reducing risk, which are often non-financial costs (Bannister, 2005). Davis’ insight on detention showed that ML in the COMPAS case involves process of inclusion and exclusion based on socio-economic factors. Those processes are connected to a differential recognition of individual rights that sees less rights recognized to those individuals that are in more difficult socio-economic conditions. Whose costs are being reduced and at what cost? How can the rationale that led to the problem of mass incarceration in the first place, be considered a solution to that problem?

Philosophy of criminal justice contributes to the understanding of ML in criminal justice by providing the analytical tools to challenge the assumptions on which the current debate bases its characterization of ML as contributing to the achievement of the goals of criminal justice reform. In other words, it provides the framework to be able to use the insight provided by philosophy of science and philosophy of technology, to discuss the political and social implications of having ML in criminal justice.

FINAL CONCLUSIONS

Machine learning was introduced fairly recently in criminal justice to substitute the judge in assessing the recidivism risk of convicts eligible for parole. To be able to assess whether the introduction of ML represents a positive change in criminal justice, it is first necessary to have a good understanding of what ML in criminal justice actually entails. I decided to focus my analysis of a practical case, which I introduced in chapter 1: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in *State of Wisconsin v. Eric L. Loomis*. Through a review of the legal documents for the case and the available literature on COMPAS, I identified three levels of analysis, which I called: tool level, sentencing level, and criminal justice reform level. I also identified three main assumptions on which the characterization of COMPAS is based in the current debate, one for each level of analysis. The assumptions indicate that the current debate fails to properly address: first, the meaning of terms such as fairness, accuracy, prediction, and risk, when it comes to ML; second, whether the risk assessments provide evidence; third, what it might mean for the assessments to contribute to the reduction of costs in criminal justice, and whether those assessments are fit to help addressing the problem of mass incarceration. Each of those shortcomings can be addressed by a branch of philosophy: the first refers to epistemological questions, and as such can be addressed by philosophy of science; the second refers to a question about the nature of technology, and therefore it can be addressed by philosophy of technology; the third refers to the nexus between criminal justice reform, economics, and mass incarceration, which can be addressed by philosophy of criminal justice.

The main research question was: *how can philosophy of science, philosophy of technology and philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?* To answer this question I drew from each of the three branches of philosophy to analyze COMPAS and address one main assumption for each level of analysis. I dedicated one chapter to each branch of philosophy.

Chapter 2 asked: *how can philosophy of science contribute to the understanding of machine learning in criminal justice?* To answer this question I focused my analysis on COMPAS at tool level. My analysis indicates that COMPAS can be confirmed to be based on a predictive model, but that either the analogy to medical practice is inappropriate, or there is a problem of characterization of what constitutes a symptom, what a disease, and the relation between the two. My analysis also shows that COMPAS is value-laden, which highlights inaccuracies in the characterization of fairness in the debate and in *State v. Loomis*. In addition to being value-laden, COMPAS is also affected by the problem of induction, which means that accuracy is a much more complex matter than how it is presented in the debate. When it comes to questioning what risk ML in criminal justice refers to exactly, an analysis involving the study of inductive risk and epistemic risk can contribute to addressing questions about the operational function COMPAS is described as having in sentencing. Therefore, I argue that philosophy of science, through the discussion of value-ladenness, induction, the difference between explanation and prediction, and inductive and epistemic risk, can contribute to the understanding of ML in criminal justice by providing the analytical tools needed to challenge the assumption about the nature and qualities of the knowledge that is associated to ML risk assessments:

fairness, accuracy, prediction, and risk. By so doing, it contributes to the understanding of the value that can be associated to that knowledge, with important implications for the way ML is understood to improve criminal justice practice.

Chapter 3 asked: *how can philosophy of technology contribute to the understanding of machine learning in criminal justice?* To answer this question I focused on COMPAS at sentencing level. My analysis indicates that the problem with having the possibility of challenging the assessment ,that is recognized by the Court as limited to the issues of transparency, opacity and the proprietary nature of the software, is actually a structural problem associated to the panoptic features of COMPAS. Together with the analysis performed in chapter 2 about bias and fairness in COMPAS, my analysis in chapter 3 clarifies that the characterization of COMPAS assessments as fairer than those performed by the judge alone does not find justification in practice. The analysis clarifies that the characterization of COMPAS when it comes to the added value it is supposed to have for the convict, in terms of him/her being assessed objectively, is misleading because those categories and standards that are supposed to represent that objectivity in fact place the individuals in a position of subjection that leaves them incapable of challenging the assessment. My analysis also shows that the rule for the assessment is a comparison to a norm that impacts on the characterization of equality, individuality and automation. Moreover, recidivism risk assessment as a form of *triage* entails a cost-benefit analysis based on context dependent considerations and characterizations of what represents a cost and what represents a benefit. Therefore, I argue that philosophy of technology, by taking a perspective that sees beyond an instrumental view of technology, and by developing a Foucauldian argument addressing panopticism and normalization in ML, can contribute to the understanding of ML in criminal justice by providing the analytical tools to challenge the assumption that ML risk assessments are tools for the provision of evidence. By so doing, it provides insight into the cost-benefit analysis that risk assessments entail which impacts on the understanding of the role ML is expected to play in sentencing.

Chapter 4 asked: *how can philosophy of criminal justice contribute to the understanding of machine learning in criminal justice?* To answer this question I focused on COMPAS at criminal justice reform level. My analysis highlights a lack of consideration in the current characterization of COMPAS of both revenues and costs that are actually part of the economy of which detention practices are part. Looking at detention through the lens of Angela Davis' study on detention and criminal justice in the book *Are Prisons Obsolete?* made it possible to consider ML in light of the connection between the recognition (or lack thereof) of individual rights and criminal justice practices. Therefore, I argue that philosophy of criminal justice can contribute to the understanding of ML in criminal justice by providing the analytical tools to challenge the assumption that ML risk assessments can help achieve the goals of criminal justice reform. By so-doing, it contributes to a discussion of the political and social implications of having machine learning in criminal justice.

The three branches of philosophy address each one important assumption on which the current discourse around ML in criminal justice is based. In addition to contributing to the understanding of ML that each branch can provide individually, I believe the most important contribution is provided when the three perspectives come together. Philosophy of science contributes to the understanding of the technical features and the claims made about the knowledge that can be associated to them, and it is the perspective that comes closer to the framework of the current debate, as it focuses on predictive accuracy and fairness. As such, it is a good entry point for opening a conversation about challenging the current understanding of ML in criminal justice with the companies developing this type of software. In order to be able to bridge the gap between software developers and legal practitioners, the insight from philosophy of science needs to connect to the perspective and concerns of criminal justice in practice, and that, I believe, is what philosophy of technology can be useful for when combined with philosophy of science. An empirical study engaging with legal practitioners and software developers that takes into account the interaction between them and ML from a non-instrumental view of technology can both contribute to answering the epistemological questions asked by philosophy of science and raise awareness in legal practitioners and software developers about the hidden complexity of their work. Then, philosophy of criminal justice can contribute to both philosophy of science and philosophy of technology by providing a framework both for the discussion of fairness and values in ML and by making sure there is a historical critical perspective on criminal justice reform and criminal justice practices including detention. In turn, philosophy of criminal justice can benefit from the other two branches of philosophy by using their insight to ground future considerations on the nexus between criminal justice and detention, and contribute to policy making in terms that can be based on an understanding of ML that has bridged the gap between software developers and legal practitioners.

The analysis performed in this thesis is intended to be a first attempt to perform a philosophical enquiry into ML in criminal justice that combined insights from philosophy of science, philosophy of technology and philosophy of criminal justice. Because of the necessarily limited size of this project I made choices about both which elements to address when it came to COMPAS, and which analytical tools offered by those three branches of philosophy to apply in my analysis. Below, I will indicate a few elements that I have not addressed and I would suggest for future enquiries (divided by branch of philosophy).

For philosophy of science:

- Underdetermination, theory-ladenness and confirmation: when it comes to addressing the decisions made in designing COMPAS, one of the first questions that came to mind was: what are those decisions based on? Equivant (2019) indicates clearly that the software's theoretical basis includes a series of criminological theories,[19] and statistics.[20] The elements that are included in the design of the software will strongly depend on the characterization of recidivism as developed in the theory providing the scientific basis for the assessment. Underdetermination, for example, refers to the fact that what is considered evidence is often

[19] Equivant (2019) lists the following criminological theories in the *Practitioners Guide to COMPAS*: social learning theory; sub-culture theory; control/restraint theory; sociopathic/socialization breakdown theory; criminal opportunity theory; social strain theory.

[20] Equivant (2019) indicates that "the COMPAS Core normative data were sampled from over 30,000 COMPAS Core assessments conducted between January 2004 and November 2005 at prison, parole, jail and probation sites across the United States" (p. 11).

compatible with more than one theory, explanation, or law (Ladyman, 2002), which in the case of the COMPAS scores could mean that the scores are insufficient to determine which theory and hypothesis is true, among those included in the theoretical basis on which COMPAS is built. Studying underdetermination, theory-ladenness and confirmation would contribute to the understanding of ML in criminal justice by addressing the assumptions on which the current debate bases its characterization of the knowledge associated to ML.

- Causality and responsibility: the analogy of COMPAS to medical practice that I have discussed in chapter 2 traces a relation between symptoms, disease and cure. That relation is about a belief on "how people become involved in criminal behavior" (Equivant, 2019, p. 5), which is ultimately a form of understanding causality when it comes to crime. Philosophy of science offers useful analytical tools for the analysis of the relation between the knowledge ML is expected to provide in the form of risk assessments as evidence to be weighed by the judge, and the underlying causal claims.[21] As the analysis of COMPAS indicates, ML risk assessment are characterized as supporting the judge in the decision-making process, but the judge still holds full responsibility for the decision made. Philosophy of science provides once more useful insight into the relation between knowledge and responsibility, especially in medical practice[22], which can be a useful context from which to draw for attempting an analysis in criminal justice. Addressing this matter would contribute to the understanding of ML in criminal justice in practice, in terms of interaction between technology and the human beings involved, in particular judges, and in understanding the impact of ML in criminal justice when it comes to their professional autonomy.

For philosophy of technology:

- Human-technology relations when it comes to ML in criminal justice: in my analysis I just referred to the fact that the current debate does not properly address the impact that COMPAS has on both convicts and judges. A more thorough investigation of the relation between ML in this context could provide insight useful for policymakers and practitioners.
- A reflection on Foucault's insight in *Discipline and Punish:* it could be interesting to see what practical cases of ML in criminal justice can show about how fitting his account is for studying normalization in criminal justice today.

For philosophy of criminal justice:

- Recidivism as a concept: a philosophical enquiry into ML and criminal justice (perhaps once again through the analysis COMPAS) that focuses on questioning the characterization of recidivism as a concept in criminal justice, I think would contribute greatly in addressing the connection between socio-economic inequality at societal level and crime in the United States.

[21] For example, Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International studies in the philosophy of science*, *21*(2), 157-170.

[22] For example, van Baalen, S., & Boon, M. (2015). An epistemological shift: from evidence−based medicine to epistemological responsibility. *Journal of evaluation in clinical practice*, *21*(3), 433-439.

- Policy: a philosophical assessment of policy related to criminal justice reform, to address the nexus between ML risk assessment, criminal justice costs reduction and detention practices, would be beneficial for the understanding of ML in criminal justice if performed focusing on the available philosophical literature on labor (perhaps from political philosophy). I think such an assessment would contribute to addressing inconsistencies in the characterization of reform and criminal justice when it comes to the debate over rehabilitation v. incapacitation in detention practices.

A general limitation of this research is that it did not include empirical research in the sense of fieldwork, interviews, observations and data collection (other than desk research). Moreover, making all three branches of philosophy come together, in a way that traced the connections between the topics discussed in the different chapters was certainly one of the greatest challenges of this project. In a future project I would revisit those connections to further investigate how they can contribute to the understanding of ML in criminal justice.

Considering the limitations of this work, and acknowledging that a lot more study on this subject needs to be done, I still hope with this thesis to have provided some insight that can contribute to efforts of theory development on ML and criminal justice from a humanities and social science perspective. More specifically, I hope that my analysis can contribute to the efforts of rethinking the narrative behind the introduction of ML in criminal justice and to challenge the status of the nexus between criminal justice and detention. Mass incarceration and discrimination in criminal justice are societal challenges not only because detention is part of criminal justice in many countries but also because the impact of what happens in criminal justice goes beyond the delimited spaces of prisons and courts of law, and goes beyond the institutional delimitations of what we call criminal justice.

REFERENCES

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data vol. 4*: AMLBook New York. *NY, USA*.

Abramovitz, M. (2006). Welfare reform in the United States: Gender, race and class matter. *Critical Social Policy*, *26*(2), 336-364. https://doi.org/10.1177/0261018306062589

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93-117. https://doi.org/10.1177/0162243915606523

American Bar Association. (n.d.) Policy Implementation Project, 18. Retrieved from www.americanbar.org/content/dam/aba/administrative/crimin al_justice/spip_handouts.authcheckdam.pdf

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica.* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baird, C. (2009). A question of evidence: A critique of risk assessment models used in the justice system. *Madison, WI: National Council on Crime and Delinquency*. Retrieved from https://core.ac.uk/download/pdf/71341860.pdf

Bannister, F. (2005). The panoptic state: Privacy, surveillance and the balance of risk. *Information Polity*, *10*(1, 2), 65-78. https://doi.org/10.3233/IP-2005-0068

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). *Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. arXiv preprint arXiv: 1702.08608v2*

Barilan, Y. M., Brusa, M., & Halperin, P. (2014). Triage in disaster medicine: ethical strategies in various scenarios. In *Disaster bioethics: Normative issues when nothing is normal* (pp. 49-63). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-3864-5_4

Bavitz, C., Bookman, S., Eubank, J., Hessekiel, K., & Krishnamurthy, V. (2018). Assessing the Assessments: Lessons from Early State Experiences In the Procurement and Implementation of Risk Assessment Tools. *Berkman Klein Center Research Publication*, (2018-8). https://doi.org/10.2139/ssrn.3297135

Beriain, I. D. M. (2018). Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. *Law, Probability and Risk*, *17*(1), 45-53. https://doi.org/10.1093/lpr/mgy001

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-3085-8

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, *27*(4), 222-228. https://doi.org/10.1525/fsr.2015.27.4.222

Borden, H. G. (1928). Factors for predicting parole success. *American Institute of Criminal Law and Criminology*, *19*, 328. https://doi.org/10.2307/1134622

Beauchamp, T., & Childress, J. (2012). *Principles of Biomedical Ethics*. New York: Oxford University Press.

Bentham, J. (1995). , In Božovič, M. (ed.). *The panopticon writings*. London: Verso.

Biddle, J. B. (2016). Inductive risk, epistemic risk, and overdiagnosis of disease. *Perspectives on Science*, *24*(2), 192-205. https://doi.org/10.1162/POSC_a_00200

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, *36*(1), 21-40. https://doi.org/10.1177/0093854808326545

Brennan, T., & Oliver, W. L. (2013). Emergence of Machine Learning Techniques in Criminology: Implications of Complexity in Our Data and in Research Questions. *Criminology & Public Policy*, *12*, 551. https://doi.org/10.1111/1745-9133.12055

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, A. J. Harno, E. W. Burgess & J. Landesco (eds.), *The workings of the indeterminate sentence law and the parole system in Illinois* (pp. 221–234). Springfield, IL: Illinois Parole Board.

Davis, A. (2003). *Are Prisons Obsolete?* New York: Seven Stories Press.

Davis, M. (1995). Hell factories in the field: a prison-industrial complex. *Nation*, *260*(7), 229-234.

Discrimination [Def. 3a]. (n.d). *Merriam-Webster Online*, in Merriam-Webster. Retrieved on August 10th, 2019, from https://www.merriam-webster.com/dictionary/discrimination

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of science*, *67*(4), 559-579. https://doi.org/10.1086/392855

Equivant (2019). *Practitioner's Guide to COMPAS CORE*. Retrieved from http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

Farabee, D., Zhang, S., Roberts, R. E., & Yang, J. (2010). *COMPAS validation study.* Retrieved from https://jpo.wrlc.org/bitstream/handle/11204/1121/COMPAS%20Validation%20Study_Final%20Report%20(California).pdf

Fass, T. L., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior*, *35*(9), 1095-1108. https://doi.org/10.1177/0093854808320497

Foucault, M. (1991). *The birth of the prison.* London: Penguin.

Freeman, K. (2016). Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75. Retrieved from http://ncjolt.org/wp-content/uploads/2016/12/Freeman_Final.pdf.

Garland, D. (1992). Criminological Knowledge and Its Relation to Power-Foucault's Genealogy and Criminology Today. *British Journal of Criminology*, *32*, 403. https://doi.org/10.1093/oxfordjournals.bjc.a048248

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. Cambridge, MA: The MIT Press.

Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science.* Chicago: University of Chicago Press.

Hamilton, M. (2015). Risk-needs assessment: Constitutional and ethical challenges. *American Criminal Law Review.*, *52*, 231. https://doi.org/10.2139/ssrn.2506397

Hartman, R. G. (2003). Tripartite triage concerns: Issues for law and ethics. *Critical care medicine*, *31*(5), S358-S361. https://doi.org/10.1097/01.CCM.0000065130.18337.05

Heidegger, M. ([1993], 2008). The Question Concerning Technology. *Basic Writings*, rev. edn., pp. 311-341. New York: HarperCollins in Scharff, R. C., & Dusek, V. (Eds.). (2013). *Philosophy of technology: The technological condition: An anthology*. John Wiley & Sons.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and OtherEssays in the Philosophy of Science*. New York: Free Press.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, Vol. 15, No. 2. (Apr., 1948), pp. 135-175. https://doi.org/10.1086/286983

Hyatt, J., & Chanenson, S. L. (2016). *The use of risk assessment at sentencing: Implications for research and policy.* Villanova Law/Public Policy Research Paper, (2017-1040). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2961288

Hume, D. (1963). An Enquiry Concerning Human Understanding. La Salle, IL: Open Court.

Karaca, K. (2019). A Methodological Framework on Accountability in Societal Applications of Machine Learning. Paper presented in the Philosophy Colloquium at the University of Twente, The Netherlands, 16 May 2019.

Kehl, D. L., & Kessler, S. A. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing.* Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School. Retrieved from https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf?sequence=1&isAllowed=y

Kemshall, H. (2003). Understanding risk in criminal justice. McGraw-Hill Education (UK).

Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, *48*(4), 507-531.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20:1, 14-29. https://doi.org/10.1080/1369118X.2016.1154087

Klingele, C. (2016). The promises and perils of evidence-based corrections. *Notre Dame Law Review*, 91, 537. Retrieved from https://scholarship.law.nd.edu/ndlr/vol91/iss2/2

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lakatos, I. (1968). *The problem of Inductive Logic*. Amsterdam: North Holland Publishing Co.

Ladyman, J. (2002). *Understanding philosophy of science.* London: Routledge. https://doi.org/10.4324/9780203463680

Lipton, P. (2004). *Inference to the best explanation*, 2d Ed. London: Routledge.https://doi.org/10.4324/9780203470855

[List of vendors participating to Corrections Procurement in the State of Wisconsin] (n.d). CorrectSource: Corrections Procurement Directory. Retrieved from http://corrections.com/vendor/result_by_state?page=1&state=WI Last visited on July 27th, 2019.

Lyon, D. (2005). Surveillance as social sorting: Computer codes and mobile bodies. In *Surveillance as social sorting* (pp. 27-44). London: Routledge. https://doi.org/10.4324/9780203994887

McAllister, J. W. (1997). Phenomena and patterns in data sets. *Erkenntnis*, *47*(2), 217-228. https://doi.org/10.1023/A:1005387021520

McAllister, J. W. (2011). What do patterns in empirical data tell us about the structure of the world?. *Synthese*, *182*(1), 73-87. https://doi.org/10.1007/s11229-009-9613-x

NIC. (2010). Achieving, measuring, and maintaining harm reduction and advancing community wellness. A Framework for Evidence-Based Decision Making in Local Criminal Justice Systems (pp. 22). Retrieved from https://info.nicic.gov/ebdm/node/78

National Institute of Corrections (n.d). Evidence-based decision making. Retrieved from https://nicic.gov/evidence-based-decision-making

National Institute of Justice (n.d.). Recidivism. Retrieved from https://nij.ojp.gov/topics/corrections/recidivism

O'Malley, P., & Valverde, M. (2014). Foucault, criminal law, and the governmentalization of the state. *Foundational texts in modern criminal law*, 317-34. https://doi.org/10.1093/acprof:oso/9780199673612.003.0017

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Pratt, J. (1995). Dangerousness, risk and technologies of power. *Australian & New Zealand Journal of Criminology*, *28*(1), 3-31.

ProPublica (n.d.) Sample-COMPAS-Risk-Assessment-COMPAS-"CORE". Retrieved from https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

Reichman, N. (1986). Managing crime risks: Toward an insurance based model of social control. *Research in Law, Deviance and Social Control*, *8*, 151-172.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science*, *45*(2), 206-226.

Relman, A. S. (1980). The new medical-industrial complex. *New England Journal of Medicine*, *303*(17), 963-970. https://doi.org/10.1056/NEJM198010233031703.

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry*, *28*(1), 95-102. https://doi.org/10.3109/09540261.2015.1106446

Risk [Def. 1]. (n.d). *Merriam-Webster Online*, in Merriam-Webster. Retrieved on August 10th, 2019, from https://www.merriam-webster.com/dictionary/risk

Rose, N., & Valverde, M. (1998). Governed by law?. *Social & Legal Studies*, *7*(4), 541-551. https://doi.org/10.1177/096466399800700405

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of science*, *20*(1), 1-6. https://doi.org/10.1086/287231

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International studies in the philosophy of science*, *21*(2), 157-170. https://doi.org/10.1080/02698590701498084

Schwartz, B. (2004). *The paradox of choice: Why more is less.* New York: Ecco.

Simon, J. (1987). The emergence of a risk society: insurance, law, and the state. *Socialist Review*, (95), 60-89.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310. https://doi.org/10.1214/10-STS330

Skeem, J., & Eno Louden, J. (2007). Assessment of evidence on the quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). *Unpublished report prepared for the California Department of Corrections and Rehabilitation. Available at:* https://webfiles.uci.edu/skeem/Downloads.html

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803. Retrieved from http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2014/04/66_Stan_L_Rev_803-Starr.pdf

State of Wisconsin v. Eric L. Loomis, 881 N.W.2d 749 (Wis. 2016). Opinion of the Supreme Court retrieved from https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690

Tallarico, S., Cheesman, F. L., Kirven, M. B., & Kleiman, M. (2012). Effective Justice Strategies in Wisconsin: A Report of Findings and Recommendations: a Report to the Director of State Courts and the Wisconsin Supreme Court, Planning and Policy Advisory Committee, Effective Justice Strategies Subcommittee. National Center for State Courts, Court Services Division. Retrieved from the National Center for State Courts (NCSC), Court Services Division website https://www.wicourts.gov/courts/programs/docs/ejsreport.pdf

Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter*, *26*(3), 167-176. https://doi.org/10.1525/fsr.2014.26.3.167

van Eijk, G. (2017). Socioeconomic marginality in sentencing: the built-in bias in risk assessment tools and the reproduction of social inequality. *Punishment & Society*, *19*(4), 463-481. https://doi.org/10.1177/1462474516666282

van Baalen, S., & Boon, M. (2015). An epistemological shift: from evidence‐based medicine to epistemological responsibility. *Journal of evaluation in clinical practice*, *21*(3), 433-439. https://doi.org/10.1111/jep.12282

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132. https://doi.org/10.1177/0162243915605575

Zedner, L. (1995). Wayward sisters: the prison for women. *The Oxford history of the prison*, 328-361.