Epistemic and Ethical Issues in Machine Learning Based Recidivism Risk-Assessment: Lessons from Philosophy of Measurement

Master thesis written by

Luca van der Peet

For the completion of the Master's degree in

MSc Philosophy of Science, Technology and Society (PSTS)

Under the supervision of

Dr. Koray Karaca (1st supervisor)

Dr. Yashar Saghai (2nd supervisor)

Faculty of Behavioural, Management and Social Sciences

University of Twente, the Netherlands

October 2023

Contents

Acknowledgements

I would like to express my deepest appreciation to Dr. Koray Karaca, for being an excellent first supervisor, giving precise feedback when I could not see the forest for the trees anymore, and strengthening my confidence in this thesis. I extend the same level of gratitude to Dr. Yashar Saghai for introducing ideas and concepts that broadened my field of view beyond machine learning and philosophy of science and for continuing his support despite his worsening health conditions. Thank you both for accompanying me up until the very end of this journey.

I am also deeply indebted to Eran Tal for making available an early version of his paper on target specification bias. This allowed me to discuss his ideas early on and incorporate them in my thesis. His warm words of encouragement in our correspondence gave me more motivation to pursue my ideas.

I would also like to extend my sincere thanks to my friends in the PSTS cohort of 2021/2022. Without this colourful mix of people, my time as a philosophy student would not have been half as fun as it was. In particular, I need to thank my fellow members of the 41st Ideefiks board, Roos, Juan and Víctor for their invaluable emotional support and constructive feedback when needed.

Last, but not least, I'm extremely grateful for the unwavering support of my family. To my mom and dad for always having my back no matter the decision I take and to my brother Mika for always magically being able to turn a heavy conversation about something that is troubling me into a lighthearted one.

Summary

Trade-offs between different definitions of fairness plague many machine learning applications, from mortgage lending, to hiring algorithms and recidivism prediction tools. However, the existing literature on fairness in machine learning resides predominantly in the computer science domain and struggles to offer solutions to these trade-offs. Fairness is mostly reflected in a formulaic and reductionist view and significant progress in statistical progress has stagnated for decades. In this thesis I revisit the discussion that erupted around one of the most notorious algorithms that was accused of persistent fairness issues towards black people: the COMPAS algorithm for recidivism prediction used by many courts in the US.

I analyse how recidivism prediction tools are constructed and whether their use can be justified from an epistemological perspective. I identify two main epistemological obstacles that stand in the way of using machine learning based recidivism prediction tools. Firstly, the criminological theory surrounding recidivism is insufficiently well founded and, secondly, machine learning tools notoriously entail problems regarding the opacity of their inner workings. By drawing from literature from philosophy of measurement I analyse how COMPAS fares at both *measuring* criminogenic needs and *predicting* recidivism. I conclude that it cannot satisfyingly fulfil either role in major parts because the tool is based on a criminogenic model developed using a *dustbowl empiricism* approach which is regarded as "atheoretical". I argue that so-called atheoretical approaches for scientific modelling both do not exist and are undesirable and conclude that the model in question harmonizes well – *for the wrong reason* – with the supposed atheoretical approach of machine learning based modelling argued for by Anderson (2008).

To address the fact that the literature surrounding fairness in machine learning is too reductionist and appears to have been stagnating for a long time, I investigate two major publications on more recently developed approaches, namely counterfactual fairness (Kusner et al., 2017) and counterfactual explanations (Wachter et al., 2017). Both stand out because they are built using a causal model as an underlying basis which, at least in theory, allows for explicitly modelling and correcting for systemic biases and fairness issues. I raise several shortcomings with these approaches and cast further doubt on the overall project of statistical fairness. By contrast, I introduce a recent publication from philosophy of measurement which addresses fairness issues by abandoning the commonly accepted benchmark for accuracy in machine learning for a metrological conception of accuracy (Tal, 2023). I conclude that the latter may be a suitable candidate for rethinking statistical fairness in a more fundamental sense.

Introduction

Algorithms play an increasingly prevalent role in decision making processes, including high-stakes areas like medicine, hiring, mortgage lending, and crime prediction. These algorithms garner a lot of attention both by philosophers and in the public eye. Especially the recidivism prediction algorithm COMPAS obtained so much notoriety that it remained a discussion topic for two decades. Early publications of COMPAS mostly focused on epistemological standards with respect to its measurement scales and predictive validity (f.ex. Eno Louden & Skeem, 2007) but the 2016 publication by ProPublica (Mattu et al., 2016) accusing the algorithm of systemic biases against black people shifted the discussion towards fairness in algorithms. In its wake, this article sparked a wave of publications in many academic areas – most notably computer science, philosophy, and law – that discuss definitions of fairness and how they can be implemented in algorithms, especially machine learning.

However, developments in algorithmic fairness have been widely dissatisfying. Machine learning ethicists proclaim severe shortcomings in the machine learning literature when it comes to various fairness definitions and benchmarks (Lee et al., 2021). In particular, there appears to be no consensus when to use which notion of fairness and how to choose among different fairness benchmarks. This lack of clarity gives the project of fairness in machine learning an air of arbitrariness and makes it appear as if no real progress is being made. In fact, as I will explain in the first chapter, contemporary discussions of fairness merely echo identical issues identified fifty years prior in the statistical literature regarding fairness in standardised testing. Perhaps even more troublesome is the fact that several publications demonstrate that there are mathematically incommensurable trade-offs between different fairness conditions (Chouldechova, 2017; Kleinberg et al., 2016). This means that, in practice, the mathematical conditions of the fairness definitions cannot be fulfilled at the same time because they oppose one another. Given that the current status of algorithmic fairness is in such disarray, I will not enter the debate on which fairness condition is most suitable for which circumstances, but rather take a step back by reconceptualising what machine learning based prediction actually entails.

For this purpose, my analysis will draw predominantly from recent publications from philosophy of measurement. This is because these publications identify an analogy between machine learning and measurement, and I investigate this analogy to see whether philosophy of measurement can provide some useful tools for addressing the fairness problems in question. My main research question in this thesis is therefore: **What kind of insights can philosophy of measurement provide about machine learning based recidivism prediction and how can these insights help us address the problem of apparently incommensurable fairness trade-offs?**

I break down the main research question into three sub-questions, each treated in an own chapter. Firstly, since I approach my analysis by focusing not only on the ethical problems of recidivism prediction but also on its epistemological conditions, the first chapter will answer the following sub-question: **What are the ethico-epistemological problems of machine learning based recidivism risk prediction?** This chapter has a more introductory function and provides the landscape of problems that I will treat throughout this thesis. Here, I identify two fairness related problems – regarding the reductionist nature of current fairness definitions and the overall lack of progress in algorithmic fairness – and two epistemological problems – regarding the definition of the measurand and the validity of the prediction tools. For this analysis I consult the criminological literature to situate COMPAS accordingly in the criminological field on the one hand and recent analyses on fairness in machine learning to sketch out the ethical problems of recidivism prediction on the other.

The second chapter focuses on the epistemological problems. I choose to focus on the epistemological part before discussing the ethical part because it appears more intuitive to me to first ask how recidivism risk prediction tools work before contending with the fairness issues they bring about. For this reason, the chapter will treat the following question: **What kinds of models is the recidivism risk prediction tool COMPAS based on?** As the epistemological problems identified in the first chapter hint at, this question can be further divided into a theoretical part (definition of the measurand) and a technical part (validity of the prediction tools). For the theoretical part I visit the psychological literature on recidivism and for the technical part I consult technical studies investigating the validity of COMPAS. I connect this literature with recent publications in philosophy of measurement that concern the analogy between machine learning prediction and measurement in order to cast a critical look at the modelling assumptions behind both the theoretical and technical parts. Rather than determining to which degree the analogy between machine learning and measurement holds, I conclude that this discussion informs best practices and helps specifying epistemological responsibilities.

In the third and last chapter, I investigate how we can approach the fairness issues identified in the first chapter from a new perspective. The reductionist definitions commonplace in the literature around fairness in machine learning have thus far not provided satisfying solutions and the trade-offs they entail can be hard to accept. I will therefore consider the following sub-question: **How can we incorporate fairness in machine learning based recidivism prediction without relying on reductionist definitions or succumbing to inevitable trade-offs?** To address this question, I firstly present the technical literature on fairness trade-offs in greater detail before introducing the notion of counterfactual fairness as an alternative conception. I contrast counterfactual fairness with a recent publication introducing counterfactual prediction (Tal, 2023) which draws from philosophy of measurement and breaks with certain practices common to machine learning. While I conclude that

this latter approach could be a promising candidate for overcoming the predicaments of algorithmic fairness, I note some reasons for caution.

To summarise, the first two questions are dedicated to the first part of the main research question; they relate to the epistemological basis of recidivism prediction. The third question relates to the second part of my main research question and attempts to conjure up a solution to the problem of inevitable trade-offs. I seek to address the fairness issues that plague many machine learning applications by focusing on the recidivism prediction algorithm COMPAS. Since I neither want to rely on the reductionist definitions of fairness common in the machine learning literature nor succumb to apparently inevitable trade-offs, I attempt to reconceptualise machine learning based prediction overall rather than remaining in the present debates of algorithmic fairness. These fairness issues cannot be neatly separated from epistemological conditions and my thesis presents therefore a joint ethico-epistemological analysis of machine learning based recidivism prediction.

Methodologically, I draw from a wide range of literature. In order to understand what recidivism is in itself and how one attempts to predict it, I consult criminological literature, most notably in the domain of psychology of criminal conduct. Since COMPAS as a recidivism prediction tool relies on machine learning, I furthermore consult the technical ML literature as well as studies concerning the epistemic validity of COMPAS. I accompany this analysis with philosophical literature on the epistemological challenges of machine learning, most notably relating to the opacity problem and mention some parallels to arguments made in the juridical context. In order to elucidate the fairness trade-offs, I translate the mathematical proofs for their incommensurability in such a way that it motivates the alternatives presented in the last chapter. I complement my entire analysis by invoking recent publications from philosophy of measurement. The reason I focus on these recent publications is because they contain a paradigm shift in philosophy of measurement away from a representational account of measurement and towards a model-based account of measurement spearheaded by Eran Tal (2012). It is this model-based account of measurement which connects measurement best to machine learning prediction and serves as a main entry point for my analysis.

Chapter 1: Contextualisation of Crime Measurement and Prediction

In this first chapter I will describe two large issues that concern machine learning based recidivism prediction tools. Firstly, there are serious epistemological questions both on the side of the construct "recidivism" on the one hand and with regards to the prediction algorithm on the other hand. What exactly is recidivism? And how are we to understand a person's likelihood to recidivate? Is the latter purely an educated guess about the future based on statistical inferences? Or is "likelihood to recidivate" something like a latent variable, a dormant potential in a person that can be triggered and cause harm given the right internal and external circumstances? Is this latent variable measurable? How then can we trust a machine learning algorithm to measure this variable? The latter question – whether machine learning algorithms can be said to *measure* – will be treated in detail in the second chapter of this thesis. For now, in this chapter, I will outline the conceptual issues relating to recidivism or *likelihood to* recidivate by providing an analysis over the field of criminology in general and the development of crime prediction tools like the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) specifically.

Criminology as a field of study has deep rooted issues regarding its internal coherence and the existence of its object of study. It is best understood as a 'rendezvous subject' where the expertise of different domains meet and come together, instead of a discipline itself (Newburn, 2018, pp. 2-3). This picture is rendered more complicated when looking at the history of crime prediction tools. Developers of the COMPAS algorithm see the history of crime prediction as *progressive* with COMPAS being one of the best available tools to date. They claim that it includes solid theoretical foundations, namely the *psychology of criminal conduct* (PCC), as opposed to previous generations of crime prediction, which were less structured, and, while empirical, more static or atheoretical (Andrews et al., 2006, pp. 7-8). However, it is also noted that there was not enough evidence to confidently state that newest crime prediction tools fare better than the best tools from previous generations of crime prediction tools. A note of caution is therefore warranted that only because a tool presents itself as the latest innovation and the most complex variant, it does not necessarily mean that it is the most preferable especially when considering a range of different evaluation criteria.

Secondly, besides the epistemological challenges, there are serious ethical issues relating to the fair treatment of different target populations by the COMPAS algorithm. Investigations by ProPublica (Mattu et al., 2016) have sparked an immense debate about the fairness issues in predictive tools, with rebuttals by the developers (Dieterich et al., 2016) and other statisticians and data scientists chiming in, among which one very notable contribution by Kleinberg et al. (2016) in which they demonstrate inherent trade-offs between different notions of fairness. To top it all off, studies into general issues regarding fairness have kept the statistics community (for example Cleary (1966, 1968) and Sawyer et al. (1976) busy since the 1960s and produced no progress on resolving the most fundamental

problems regarding trade-offs. In fact, Hutchinson and Mitchell (2019) lay out how the discussion surrounding fairness issues in COMPAS is merely echoing many of the exact issues statisticians have identified in the 1960s with respect to fairness issues in standardized testing and grading.

The analysis to be provided in this chapter will suggest that justifying the use of recidivism prediction algorithms is problematic both from epistemological and ethical standpoints. This will set the stage for the subsequent chapters where I will analyse whether reframing machine learning as measurement instruments is, firstly, justifiable, and, secondly, able to handle both the epistemological and ethical problems raised in this chapter. For now, however, I will proceed as follows. Section 1.1. will give a cursory overview about the field of criminology before transitioning in section 1.2. into an analysis of the development of crime prediction tools and how they are framed. Section 1.3. will then outline the ethical issues relating to fairness, how there are trade-offs between incommensurable definitions of fairness and how the literature has responded to these issues. The conclusion will tie these analyses together and set the stage for the subsequent chapters.

1.1. The construction of crime

The purpose of this chapter is locating our concept of interest, recidivism, within its surrounding field of study, criminology. By highlighting the theoretical and methodological difficulties within criminology, in particular with regards to the notions of 'crime' and 'criminal', we can discuss the domain-specific obstacles to the creation of an appropriate metric for recidivism prediction.

First versions of criminology emerged in the late 18th century as disconnected endeavours to collect data and analyze criminal behaviour (Newburn, 2018, p. 2). These local movements rarely, if ever, used the label "criminology" to describe their work, but understood themselves as branching off from other established sciences, like anthropology, sociology, psychology, or statistics (Newburn, 2018, p. 2). Nowadays, criminology is still best understood as a subject or a field of study, rather than a discipline (Newburn, 2018, pp. 2-3). This is because a discipline generally disposes over a more or less well-defined and unilaterally accepted methodology and theoretical foundation which are both lacking in criminology.

In terms of theory, it is striking that "[…] crime, the core subject of criminology, has no *ontological reality*." (Newburn, 2018, p. 17, emphasis in original). Crime is a very relative concept. The legal status of many acts has transformed over time. For instance, in the UK, the Abortion Act 1967 decriminalized abortion under certain circumstances. Another example is the Sexual Offences Act 1967 decriminalizing homosexual activities between adult males. Examples for acts that used to be legal but were criminalized subsequently include, for example, the use of Opium in the UK up until 1860, smoking cigarettes inside public spaces, and marital rape. Furthermore, it goes without say that the legal status of certain actions varies from one legislation to another (Newburn, 2018, pp. 9-14).

Another ontological challenge for the concept of crime is the aspect of *criminalization*, i.e., the idea that criminals are *constructed* by the fact that individuals are "processed" within a criminal justice system and subsequently obtain the label 'criminal'. This is not to say that the process is unjust or arbitrary *per se*. Instead, what is important is to bring attention to the notion of power in deciding what constitutes a crime and who therefore constitutes a criminal (Newburn, 2018, pp. 14-17). For instance, policing practices like 'stop and frisk' (in the US) or 'stop and search' (in the UK) disproportionally targeted ethnic minorities leading to a disproportionate processing of minorities by the criminal justice system and labelling as 'criminal' (Newburn, 2018, p. 15).

Deriving from both the relativity of the concept of crime and the socially constructed aspect of criminalization are challenges to define constitutional differences between criminals and non-criminals (Newburn, 2018, p. 18). The vast majority of people commit at least one of a range of minor offences like weed consumption, driving under influence, stealing from a shop, or illegally downloading music from the internet (Newburn, 2018, p. 19). The most important takeaway from this chapter so far is that the terms 'crime' and 'criminal' should be handled with extreme caution due to their contextual dependencies both in the sense that the socio-political context determines their definitions and in the sense that the context influences the occurrence of crime.

The above critiques about the contextually dependent relativity of crime and criminals are most prominently provided by feminist and critical studies. These fields broadly emphasize the power dynamics associated with the act of defining and categorizing concepts like "crime" or "risk" and put special attention the social context in which these concepts are deployed. They often try to maintain that these concepts have no authority that extends beyond specific socio-cultural contextuality and, as such, have no claims to authority in an absolute sense. They would argue along the lines of stating that the criminal *per se* does not exist, but is only constructed within a particular setting and that there are no universal criteria that ultimately and indefinitely pin down the essence of crime, criminal, risk, etc.

Feminist and critical critiques have a strong point especially when considering non-violent offenses like weed consumption, or when considering the history of the criminalization of homosexuality[1]. The differences in legislations across different societies and times lends credibility to their being arbitrary and expressions of contingent factors, perhaps even Foucauldian desires for control and power. These critical points are also acknowledged in the literature about crime prediction. They are even invited and encouraged as assuring that developers of crime prediction tools remain sceptical of their practices and are committed to demonstrate the reliability and validity of their tools (Andrews et al.,

[1] As of writing this (30.05.2023), Uganda passed an Anti-LGBTQ bill endorsing the death penalty for "aggravated homosexuality".

2006, p. 21). Maintaining a commitment to the rationalist empirical tradition, they consider scepticism and aversion towards misuse as an integral part of their practice.

Variants of typical postmodern critiques towards criminology manifested themselves for example as *labelling theory*, which states that the law is created by the powerful and is as such a tool for power and controlling what they deem deviant. The criminal therefore does not exist but is created by the institution of law; the criminal is *labelled* as such, leading eventually to an identification of the labelled with the constructed deviant subculture and thus with their own label (Wellford, 1975, p. 333). One main counterargument is the widely established literature on cross-cultural studies about categories of criminal behaviour. This literature finds that most violent crime such as, for instance, murder, forcible rape, and robbery are condemned to relevant degrees in all cultures (Wellford, 1975, p. 334). The rational empiricist tradition maintains a tension against the most sweeping postmodern critiques and justifies their commitments with rising predictive power of their methods and cross-cultural validations when it comes to categorizations of crime.

A more thorough and complete analysis on the debate between the rational empiricist camp on the one hand and the feminist and critical studies camp on the other deserves its own thesis. A glance into this debate was necessary to clarify the constructed nature of the concepts “crime” and “criminal”, as this has repercussions for the justifications of a metric created for recidivist behaviour. For reasons of scope, I will limit my analysis to the cursory points made above. I also consider the rational empiricist argument for considering violent crimes like murder, forcible rape, aggravated assault, robbery, burglary, larceny, and auto theft (Wellford, 1975, p. 334) as *intrinsically criminal* as justified due to cross-cultural studies supporting their condemnation in all cultures studied. At the same time, the limitations of crime prediction tools with respect to non-violent crime like drug abuse or practices criminalized based on ideological grounds like homosexuality have to be acknowledged and require a more nuanced look at crime prediction tools which differentiates between different criminalized practices.

To the degree that we do distinguish between criminals and non-criminals, a range of ‘risk factors’ have been identified that indicate a higher propensity to committing crimes. Individual factors include low intelligence, low educational achievement, hyperactivity, impulsiveness, and childhood antisocial behaviour. Family circumstances include poor parental supervision, child physical abuse, child neglect, parental conflict, and delinquent siblings. Peers can contribute to a higher likelihood of criminal behaviour when they are themselves in trouble or rejected by other peers. Lastly, the community can raise the likelihood for crime when the individual is living in a high crime

neighbourhood (Newburn, 2018, p. 32)[2]. Regarding the impact of these influences, it needs to be said that "[t]he risk factors are important influences rather than distinguishing characteristics or determining features." (Newburn, 2018, p. 21). There appears to be a strange twilight space which predictive tools occupy. They identify factors that consistently, in terms of statistical relevance, predict certain events, but, as Newburn (2018) statement above appears to illustrate, proponents of these tools want to refrain from claims about determinism. Instead, one talks continuously of "likelihoods", "propensities". In the next chapter I will provide an introduction into how such likelihoods are estimated with recidivism prediction tools, most notably at the example of COMPAS.

1.2. The logic behind COMPAS?

Crime prediction tools are generally not used to predict *first* crime but rather *second* or *future* crime. With a bulk of interventions attempting to treat and prevent *re-offense* rather than *crime itself*, it appears as if the justice system treats first crimes as tragic, uncontrollable events, whereas reoffenders are immediately on the system's radar and the failure to control their reoffending is seen as particularly tragic. Indeed, Weisberg (2013) comments:

"First crimes are caused by inherent character or social conditions that are too complex to control. But once someone is identified as an offender, the system is on notice that he is prone to offend, and if he enters the system the failure to control becomes an especially lamentable and, in theory, avoidable failure." (Weisberg, 2013, p. 788)

This quote indicates a certain framing of crime prediction tools. Most people object in horror to an Orwellian system set out to monitor every citizen in order to prevent first crime. It is all the more striking that there is a system in place that intently studies the criminal population, which, at least in 2008 in the US, made up one percent of the entire population[3].

Vast studies on the US criminal population have, in a way, culminated in the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) recidivism risk assessment algorithm developed by Northpointe (now Equivant). This supervised machine learning algorithm is used as a tool to predict the needs and risk of recidivism of a defendant in order to inform a judge's decision about where to place, how to supervise and how to manage the case of a particular offender. Besides jailtime, drug courts and mental health courts have emerged to provide alternative sanctions to offenders comprised of a mixture of "sanction, supervision and therapy" (Weisberg, 2013, p. 797). Rather than exclusively providing motivations in favour of or against jailtime or bail, the COMPAS

[2] Note also that many of these proxies appear to track poverty, which is another problematic aspect of crime prediction tools.
[3] This has declined to 810 people per 100.000 in 2019 (Gramlich, 2021). For contrast, the incarceration rate in the Netherlands declined from 100 per 100.000 in 2008 to 66 people per 100.000 in 2021 (*World Prison Brief - Netherlands*, retrieved 1st June 2023)

tool is therefore additionally aimed at informing a judge's decision about the kind of sanctions to offer to an offender.

For a thorough overview of the recent history of crime prediction tools, I refer to Andrews et al. (2006)[4]. In cursory terms, they distinguish between four generations of crime prediction tools, where those of the first generation (1G) were mostly unstructured professional judgements, second generation (2G) tools began to include empirical bases but remained atheoretical and static, and third generation (3G) tools tended to include wider samples of dynamic variables and introduce theoretical foundations (Andrews et al., 2006, pp. 7-8). Fourth generation (4G) tools are described as the best currently available methods with wide applicability, able to handle multiple purposes, and accompanying assessments from the beginning until the end of an offender's time within the criminal justice system. Indeed, Andrews et al. (2006) describe their major goal as "to strengthen adherence with the principles of effective treatment and to facilitate clinical supervision devoted to enhance public protection from recidivistic crime" (Andrews et al., 2006, p. 8).

It is clear that the authors attach high hopes and expectations to this generation of instruments. Their confidence in "effective treatment" and "clinical supervision" stems in great deals from advances in the psychology of criminal conduct (PCC). In fact, they note that "theoretical, empirical, and applied progress within the psychology of criminal conduct (PCC) has been nothing less than revolutionary." (Andrews et al., 2006, p. 8). The main contributions of PCC to 4G instruments consists in a theoretical understanding of social learning and social cognition theory, an empirical understanding of the so-called risk-need-responsivity (RNR) model shared by several disciplines interacting in the criminal justice system, and empirical studies into the effectiveness of different treatment options (Andrews et al., 2006, pp. 9-12).

This risk-need-responsivity model played a central role in third and fourth generation crime prediction tools. In fact, Bonta and Andrews (2007) note that "third and fourth generation risk assessment instruments would not have been possible without the risk-need-responsivity model of offender assessment and rehabilitation." (Bonta & Andrews, 2007, p. 4). As the name suggests, the model is based on three principles. The *risk principle* states that an offender's likelihood of recidivism can be reduced if the level of the treatment intervention is proportional to the offender's risk of reoffending (Bonta & Andrews, 2007, p. 5). It entails therefore assigning more resources to the treatment of high-risk individuals than to low-risk individuals. The *need principle* requires that the treatment focuses on *criminogenic needs*. Criminogenic needs are dynamic risk factors that can be affected by treatment (as opposed to static risk factors like sex which are immutable) and are directly linked to risk factors.

[4] See also Barabas et al. (2018).

Each need individually requires a specific intervention that is defined by the criminogenic need[5]. For example, one of the major risk factors is a history of antisocial behaviour and its associated criminogenic need that should be targeted by the treatment intervention is to "build noncriminal alternative behaviour in risky situations" (Andrews et al., 2006, p. 11). Lastly, the *receptivity principle* states with respect to *general receptivity* that social cognitive learning methods are the most effective treatment option irrespective of the kind of behaviour being treated and with respect to *specific receptivity* that treatment should be catered to "personal strengths and socio-biological personality factors" (Bonta & Andrews, 2007, p. 7) in order to increase the success rate of the intervention. All in all, the RNR model centres around effective treatment, focusing on effective resource allocation and tailoring treatment interventions to the specific needs and circumstances of the offender.

As a stark contrast, a report by ProPublica revealed biased predictions against black defendants, disproportionally classifying them wrongly as high risk and disproportionally classifying white defendants wrongly as low risk (Mattu et al., 2016). The report sparked a huge debate about the fairness of algorithms, especially risk assessment algorithms with main contributions pointing out how large and variable different notions of fairness can be as well as that they often cannot be fulfilled at the same time (Kleinberg et al., 2016). Instead, they tend to trade off both against one another and against the accuracy of the algorithm. A huge wave of publications analysed the issues of fairness in predictions and manifold fairness frameworks and benchmarks for ML and AI were created[6].

Rudin et al. (2020) attributed most of the confusion around the COMPAS algorithm and its fairness issues to the lack of transparency of the algorithm. Two types of opaqueness are at play here. On the one hand, it is unclear how to explain or understand the way the different variables are weighted and connected inside the algorithm and how to provide a satisfying justification for a given output. On the other hand, the algorithm is protected by proprietary law such that a direct investigation into its functioning is impossible in the first place. Investigators and scientists can only indirectly scan given outputs for particular datasets to come to certain conclusions. This is why ProPublica could be accused of a faulty analysis of the COMPAS algorithm – the lack of transparency makes a clear investigation impossible, leading to cases where a biased algorithm can be made to appear just, or, conversely, where a fair algorithm can seem biased. Transparency in both the legal and the technically

[5] The most relevant factors are called the "central eight" (Andrews et al., 2006, p. 10). For an overview of risk factors and their associated criminogenic needs, see Table 1 in Appendix A (*reprinted from* Andrews et al., 2006, p. 11).

[6] See Berk et al. (2021) for an overview of different fairness definitions and how they trade off against one another and against accuracy, and Lee et al. (2021) for going beyond technical definitions of fairness and attempting to reconnect them to philosophical definitions.

interpretable sense is thus an overarching value that would enable the creation of fairness in these algorithms in the first place[7].

This problem of transparency is already apparent in reports and assessments on COMPAS that were published before ProPublica's investigation. In (Eno Louden & Skeem, 2007), the authors provided a scathing analysis of the COMPAS algorithm. Their analysis followed along the lines of predictive utility, construct validity, and reliability which they define in the following way:

- "Predictive utility: [the COMPAS] must contain a standard scale(s) that has been shown to predict *future* reoffending; its actuarial prediction formulae must be cross-validated with independent samples; and it must emphasize criminogenic needs that predict future reoffense;
- Construct validity: it must measure the criminogenic needs it purports to measure; for example, it should relate coherently to other measures of needs and capture change in risk state over time;
- Reliability: at the most basic level, it should produce scores that are consistent over short periods of time (test-retest reliability), different items (internal consistency within scales), and different evaluators (interrater reliability)." (Eno Louden & Skeem, 2007, p. 5, emphasis in original)[8]

They found only weak support for predictive utility and even claimed that only one of the eight factors used in the prediction was actually predicting recidivism and constituted only a limited correlation. In terms of construct validity, the COMPAS developers do provide strong evidence that their model is rooted in substantiated theory, but provide no evidence that the way the variables are implemented and linked in the model relates in any way to existing measures and scales. Lastly, in terms of reliability, no support was given neither for test-retest nor for interrater reliability. As such, the authors could not recommend the use of COMPAS for its intended purposes. In (Brennan et al., 2009), the COMPAS developers themselves directly respond to the report by Eno Louden and Skeem (2007). They mainly make assurances that their algorithms are tested to ensure adherence to the criteria Eno Louden and Skeem (2007) list, but since the algorithm is protected by proprietary law, such assessments and responses ultimately cannot lead to a fruitful result.

An assurance of transparency is therefore a necessary condition for making proper assessments of recidivism risk assessment algorithms in the first place and enabling the creation of fairer algorithms in the future. The prospects of the latter are also bound by obstacles and reasons for pessimism, as I

[7] See also Rudin (2019) for advocating against black box algorithms and in favour of interpretable models.
[8] Note in this quote that *predictive utility* refers to predicting events in the future while *construct validity* to measuring a criminogenic need currently present in the offender. It is not clear how future prediction and present measuring relate to one another and which of the two the COMPAS algorithm is performing in actuality. This conceptual confusion will be picked up in chapter 2.

will demonstrate in the part below by analysing how notions of fairness in statistics have (or have not) developed since early works from the 1960s.

1.3. No progress in statistical fairness

The COMPAS affair sparked a massive debate around notions of fairness. Particularly noteworthy is that much of the literature stems from data scientists and has an origin in computer science, rather than a sociological or philosophical background. Critics point out that these computational fairness frameworks are too constrained, unable to capture the "true" complexity of fairness (Lee et al., 2021).

This issue leads to an interesting tension between the "true" meaning or conceptualisation of fairness and the *operationalization* of fairness. Data science scholars focused more on how to effectively operationalize the notion of fairness, i.e., make it implementable for an algorithm. This is why there exists a massive literature on technical fairness frameworks and benchmarks. Such frameworks have been summarized elsewhere better than I can do here, see for example (Hutchinson & Mitchell, 2019; Lee et al., 2021; Verma & Rubin, 2018; Washington, 2018), but I provide an overview over the fairness conditions that were at the core of the COMPAS-ProPublica debate: *Predictive parity*, *false positive error rate balance*, and *false negative error rate balance*.

Given a protected[9] class, for example race R, a predicted decision d (for simplicity, 1 for high risk of recidivism and 0 for low risk), and the actual outcome or true value Y (whether a defendant *actually* has a high risk of recidivism or not), *predictive parity* requires that the different groups w (for white) and b (for black) in the protected class R have an equal likelihood of truly belonging to the positive class given that they received a positive decision. Mathematically, this is equivalent to the following equation:

$$P(Y = 1|d = 1, R = w) = P(Y = 1|d = 1, R = b)$$

The equation states that the probability P that the true label is positive ($Y = 1$) given that we have a positive decision ($d = 1$) is equal for both races ($R = w$ and $R = b$). This is the definition of fairness the COMPAS developers focused on. Intuitively, the need for this requirement comes down to the fact that if predictive parity were not respected, then white and black defendants would have different likelihoods of *truly* belonging to the high-risk category when given a positive prediction (this likelihood is called *positive predicted value* (PPV), also referred to as *precision*). A judge evaluating the prediction of the algorithm would then face the conundrum that a positive prediction by the algorithm essentially has a different meaning for white and black defendants. If black defendants had a lower PPV than white defendants, then a judge had reason to trust the predictions for white

[9] Protected in the sense that it is prohibited to discriminate based on the features that define this class, e.g., race, sex, gender.

defendants but not for black defendants. This effectively renders the algorithm useless, since the judge would disregard the algorithms prediction specifically for black defendants and rely instead on his personal judgement which constitutes a case of differential treatment based on race.

False positive error rate parity requires that the false positive rate (FPR) for black and white defendants is equal. The false positive rate is the ratio between false positive predictions and all actual negative cases: $\frac{FP}{TN+FP}$ and reflects the likelihood that a defendant is given a positive prediction (i.e., high risk) when they are actually in the negative class (i.e., low risk). The intuition behind this condition is quite straightforward: a false positive prediction means that a defendant was wrongly identified as a high-risk individual leading to a harsher sentencing than they would actually deserve. Overall, one would like the FPRs to be as low as possible, but disparate FPRs additionally have the effect that one group disproportionately receives unjust harsher sentencing than another. In the case of COMPAS, the FPR for black defendants was higher than for white defendants, additionally leading to disfavouring a minoritized group. Mathematically, false positive error parity can be stated such that the probability P that a defendant receives a positive prediction ($d = 1$, i.e., high risk) when their true label is negative ($Y = 0$, i.e., low risk) is equal for both races ($R = w$ and $R = b$). It can be formulated as follows:

$$P(d = 1|Y = 0, R = w) = P(d = 1|Y = 0, R = b)$$

Lastly, *false negative error parity rate*, analogously to the condition above, requires that the false negative error rate (FNR) is equal between groups. It is defined as the ratio between false negative predictions and all actual true cases: $\frac{FN}{TP+FN}$. Again, the intuition behind this condition is quite clear. A false negative prediction would mean wrongly releasing a high-risk individual. Since, in the case of COMPAS, the FNR was higher for white defendants than for black defendants, it meant that white defendants disproportionally wrongly received milder sentencing than black defendants. Mathematically, this condition can be stated as:

$$P(d = 0|Y = 1, R = w) = P(d = 0|Y = 1, R = b)$$

Each of these definitions can be debated based on their merit and drawbacks for a given context. However, the key point is that these frameworks mostly function as technical approaches to an issue that is deeply social and philosophical. There is no clarity how to choose among the variety of different definitions of fairness nor how to handle their trading off against one another. Rather, because it is difficult to clarify what each of these benchmarks or technical definitions of fairness mean in relation to moral judgements and ethical schools of thought, it is also unclear how to justify the choice of a particular fairness definition or benchmark.

This problem that the choice of a particular technical definition of fairness is hard to justify lends them an air of arbitrariness and risks invoking the danger of developers picking those frameworks that require the least amount of effort in order to label their algorithms "fair". We should expect machine learning developers to reflect upon which kind of fairness issues are particularly relevant for a given application and reconnect their practice to the context their tools will be applied to. First attempts to stoke the discussion of values and fairness can be found in (Lee et al., 2021) in which the authors connect technical fairness definitions to schools of thought from ethical philosophy and welfare economics and suggest concrete steps on how to guide such reflections in development processes (Lee et al., 2021, p. 539).

Besides the problems regarding mathematical or technical definitions of fairness and benchmarks today, a historical argument is put forward by Hutchinson and Mitchell (2019) who demonstrate that the recent debate around COMPAS merely *echoed* a practically identical debate from the 1960s and 1970s which arose around test fairness at the time[10]. Most strikingly, the authors quote Sawyer et al. (1976) as eerily predicting the very same trade-offs that would occur in the COMPAS debate in 2016:

"A conflict arises because the success maximization procedures based on individual parity do not produce equal opportunity (equal selection for equal success) based on group parity and the opportunity procedures do not produce success maximization (equal treatment for equal prediction) based on individual parity. Such distinctions must be treated through utility statements." (Sawyer et al., 1976, p. 69)

Compare this statement with Kleinberg et al. (2016):

"Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal — that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously. Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to approximate versions of the conditions as well." (Kleinberg et al., 2016, p. 3)

Though using slightly different terminology ("group parity" is equivalent to the "calibration condition", and "equal opportunity" is equivalent to the "balance conditions for the positive and negative classes"), they identify the same kind of trade-offs, forty years apart from each other, one in statistical fairness for testing, and the other in machine learning applications.

[10] See for instance Cleary (1966) and Cleary (1968) for studies of fairness in college GPAs and SAT scores.

The implications of the arguments presented in this chapter are, firstly, that technical or mathematical definitions of fairness are too reductionist and insufficient to satisfy richer conceptions of fairness that are valuable and important, especially if one talks about high stakes domains like criminal justice. Relying too much on these definitions risks underestimating the importance of richer fairness definitions and may lead developers to resort to oversimplified, bare-minimum benchmarks to fulfil an arbitrary, not well justifiable benchmark for fairness.

Secondly, Hutchinson and Mitchell's (2019) historical analysis of the statistical fairness literature indicates that the solution to these problems are likely not to be found within the current practices and paradigms. The fact that identical issues are re-identified forty years apart implies that the way experts in these fields have been thinking about these issues may be inadequate and that new ways of thinking are required to overcome these problems. I will pick up arguments drawing from philosophy of measurement in the following two chapters and discuss if they can provide tools that would enable exactly such a rethinking of the issues presented.

1.4. Conclusion

Four arguments have been presented in this chapter, two of them epistemological, and two ethical. The first epistemological argument put into question the degree to which we conceive of a metric for measuring recidivism. Criminology itself is a young field that has not developed a unified theory or methodological agreement about its objects of analysis. If one narrows down the discussion to most violent crimes, for which there appears to exist a strong empirical support for cross-cultural agreement and therefore strong justification for categorizing as criminal behaviour, one avoids typical arguments from feminist and critical studies about the arbitrariness of crime labelling and their being mere expressions of power dynamics. However, crime prediction tools need to be designed so as to keep differences with respect to less violent types of crime in mind.

The second epistemological argument concerned the history and development of crime prediction tools which describes their iteration from first generation, unstructured professional judgements to fourth generation machine learning tools as a story of linear progress, vindicating the latter as the state of the art. While COMPAS as one of the prime representants of 4G tools is claimed to encompass theoretical foundations from the psychology of criminal conduct and well established empirical methodologies like the risk-need-receptivity model, their usefulness is put into question along the dimensions of predictive utility (weak evidence that variables used in assessments predict recidivism), construct validity (strong evidence for theoretical foundations, but no evidence for validating the way the factors are implemented and weighted inside the algorithms), and reliability (no evidence for test-retest or interrater reliability). Responses by COMPAS developers defending their algorithm cannot be substantiated because they are protected by proprietary law and therefore not required to disclose their exact functioning.

Despite the shortcomings due to these epistemological considerations, algorithms like COMPAS have been in use for years now and are bound by issues relating to fairness between different population groups. While direct analyses into the severity of these fairness issues is rendered close to impossible due to their lack of legal transparency, a wave of publications proceeded to address several different definitions of fairness that can be at play in machine learning applications, as well as how they trade off both against one another and against accuracy. The first ethical argument stated that many of these publications provided fairness definitions and benchmarks that were too simplistic and reductionist in nature, essentially disconnecting the computer science literature from more profound, context-dependent discussions of fairness.

The last ethical argument stated that the literature relating to statistical fairness has made no significant process within the last fifty years. Rather, scholars writing on the subject of test unfairness in the sixties and seventies *predicted* some of the exact same issues the fairness literature in machine learning has noted in the debate that was sparked around the COMPAS algorithm. This implies that current paradigms in statistical fairness are insufficient for addressing these problems and that a more profound rethinking of these problems is required to make progress.

In the following chapter I will suggest one such way of rethinking the field of machine learning by drawing from philosophy of measurement. Mussgnug (2022) suggested that, in the area of machine learning based poverty prediction, a "predictive reframing" has happened which led to machine learning developers to understand their task as a *prediction* task different from the original *measurement* task that measured the initial levels of poverty. This reframing came with an abdication of responsibility on the side of the machine learning developers, paying only little attention to the validity of the poverty metric they trained their models on and not taking into consideration anymore the scope of applicability in specific contexts. Reverting this reframing and considering machine learning tools as "automatically calibrated measurement instruments" would, according to Mussgnug (2022), reinstate these epistemological responsibilities. This account provides a useful starting point to address the ethical issue of the epistemological responsibilities of machine learning developers in recidivism risk prediction in the following chapter.

Chapter 2: COMPAS – Measuring? Predicting? Or None of the Above?

As has become clear in the first chapter, new concepts are required to stimulate a rethinking in the field of machine learning development. In this chapter I will focus on the epistemological issues concerning the use of COMPAS and will draw comparisons to recent publications in the field of philosophy of measurement (Mitchell, 2020; Mussgnug, 2022; Parker, 2017; Tal, 2023; Tal, 2012). The first epistemological issue identified in the first chapter relates to the definition of the measurand. In order to address this issue, however, it is necessary to clarify the purpose of COMPAS, because both the developers themselves and critics of the tool refer to COMPAS both as *measuring* criminogenic needs and as *predicting* an offender's likelihood of recidivism. After a brief motivation for the use of philosophy of measurement for my analysis, I therefore dedicate a section to an analysis of COMPAS as *measuring* tool and one to COMPAS as a *predictive* tool. The measuring part of COMPAS is predominantly concerned with the risk-needs-receptivity (RNR) model introduced by Andrews and Bonta (1998) which I problematize for its lack of theoretical structure. My main conclusion of this section will be that machine learning and the RNR model are well compatible with each other for the wrong reasons, namely that the RNR model was developed by relying on a *dustbowl empiricism* approach which understands itself as "atheoretical" and that this harmonizes well with the opaque structure of ML algorithms. The prediction part of COMPAS concerns the inference from the algorithm's output to an individual's likelihood of recidivism. My main contention of this inference is that it lacks most of its explanatory basis and is therefore hard to defend when challenged from an epistemological perspective. Subsequently, I contrast the notions of measurement and prediction using Mussgnug's (2022) account of how epistemological responsibilities shifted when poverty prediction algorithms were used in the place of original measurement procedures. My overall conclusion of this chapter will be that important decisions about the design and use of algorithms like COMPAS do not allow for considering epistemological and ethical issues separately.

2.1. Why Philosophy of Measurement?

With machine learning and AI technologies booming, many scholars, amongst which philosophers of science and philosophers of technology, are dedicating a lot of attention to both the epistemological basis of these tools and the ethical justifications and implications of their use. Of the few things that these scholars tend to agree on, one is that these technologies are quite powerful and hold a large chunk of the public imagination in their grasp. Fantasies describing AI as new forms of intelligence seem to presume a level of novelty and uniqueness that sets these technologies apart from everything that has come before, leading to some kind of enchantment of these technologies (e.g., Campolo & Crawford, 2020). Part of this enchantment is the fact that many of the more sophisticated AI and ML tools are opaque black boxes with internal structures that are so complex and convoluted that they

escape any meaningful interpretations. While claims of intelligence are difficult to demonstrate, the opaque nature of many AI and ML models also makes it difficult to disenchant such claims.

Part of my intention in this chapter is to take a step back from the enchanted discussion about ML and AI and cast a fresh look at them from a well-defined vantage point. While philosophy of measurement set itself apart as a distinct field after around 1850, discussions about magnitude and quantity reach back until the antiquity (Tal, 2020). As such, there is a vast literature available with different schools of thought that grapple with the fundamentals of measurement. It is from here that I would like to draw concepts to analyse the epistemological and ethical conceptualisations of machine learning. At first glance, there are several similarities between measuring and using machine learning (Tal, 2023, p. 317): Both are types of methods used to evaluate certain variables based on concrete input. They share a modelling phase (training phase for machine learning, and calibration for measurement instruments) where reliable data (a training data set or values corresponding to standards) are used to ensure a stable mapping between inputs and outputs. Both methods are supposed to be generalizable to new, unseen events or objects during application and they are optimized to predict the values of a target variable. Lastly, the evidence they provide for decision making is said to be reliable within certain limits. One can thus reasonably state, at least at a glance, that there exists some sort of analogy between measurement instruments and machine learning algorithms.

Knuuttila and Loettgers (2014) and Linnemann and Visser (2018) describe which role analogical reasoning plays in synthetic biology and emergent gravity, respectively. The purpose of analysing the analogy between measurement instruments and machine learning will follow in a similar vein: machine learning can be seen as a comparably new, complex discipline that is struggling with different paradigms stemming from engineering, science, and computer science. The way recent scholars from philosophy of measurement began writing on the analogy between machine learning and measurement instruments may be interpreted as attempts to clarify and solidify the paradigms that guide and should guide best practices in machine learning. Both Knuuttila and Loettgers (2014) and Linnemann and Visser (2018) describe analogies in their respective contexts not as being used for providing specific arguments but rather to come up with new concepts. Recent publications from philosophy of measurement similarly discuss the role that machine learning can play in science and science-based decision making.[11] As such, Parker (2017) discusses the epistemic justification of simulations in climate modelling by comparing simulation outcomes to measurement instrument outcomes; Mitchell (2020) compares machine learning tools to NMR spectroscopy and defends the use of ML from an instrumental perspectivist view; Mussgnug (2022) diagnoses an abdication of

[11] One thing I will not attempt to do is to *equate* machine learning and measurement instruments in order to justify calling machine learning tools "measurement instruments". As my analysis below will demonstrate, the differences between these two kinds of tools are prominent enough to prohibit such an equation. Instead, their comparison will serve to fix epistemic responsibilities.

responsibility by ML developers when compared to measurement experts in poverty prediction tasks; and Tal (2023) directly tackles the joint ethico-epistemological issue regarding fairness in machine learning aided decision making in health care. Each of these publications serves as a vantage point from which to analyse the use of machine learning in science and decision making and will be discussed in the following sections. Tal (2023) will be discussed in greater detail in the last chapter of this thesis.

2.2. COMPAS as *measurement* tool

In any kind of measurement activity, one resorts to inferences from observable variables to a target, referred to as 'measurand'. In order to establish this inferential structure between observable variables and measurand, one has to specify the relations between the two, and, in the case of more complex measuring tools, the relations between different parts of the tool deployed, as well as theoretical assumptions about the context of use. This process of specifying relations and assumptions is referred to as 'calibration'. While the term 'calibration' is used in many different contexts with slightly different meanings[12], Tal (2017) describes calibration as a modelling process involving the specification of said relations and assumptions aimed at establishing and justifying the inferential structure. In order to obtain justified inferences about the measurand, one requires justified specifications of relations and assumptions in the measurement process.

What then would be the equivalent of a measurand in recidivism risk prediction? Strangely, COMPAS seems to present itself both as a tool for *predictions about future outcomes* and about *measuring a latent, psychological construct* roughly circumscribed as "likelihood for recidivism" and connected to the risk-needs-responsivity (RNR) model. It is difficult to separate the two conceptions because the tool makes statistical inferences based on both contextual (social) and psychological variables of the defendant to determine the likelihood of this person offending in the future. Based on this inference, the offender is at the same time branded as embodying something akin to a latent potential of recommitting a crime, ready to emerge at any point. Before we can discuss the measurand COMPAS is targeting, it is necessary to untangle this conceptual confusion about prediction of future risk and measurement of psychological properties.

In my view, the reason why COMPAS appears to present a certain hybridity as both a prediction and a measurement instrument stems in part from the fact that its developers resort to concepts from psychology of criminal conduct as a theoretical underpinning for its model. However, these theoretical considerations do not appear to go beyond the selection of the input variables (of which there are 137 (Rudin et al., 2020) taken into consideration for the model. How these factors interrelate

[12] For example in Bayesian statistics to describe valid statistical inferences, or in common parlance when one gauges the kitchen balance (Tal, 2017, p. 33).

and determine the outcome of the algorithm is left to the training process of the model and hidden, first of all, behind the trade secret of the algorithm and, secondly and more to the point for the topic of this thesis, behind the opaque internal structure of the algorithm. One crucial problem with the opaque structure of machine learning based algorithms is that it prevents any meaningful explanation of the way by which the tool arrives from the input to the output.

Eno Louden and Skeem (2007) mention several times in their assessment of the COMPAS algorithm that the tool purports to *measure* certain targets[13]. These targets are *criminogenic needs* derived from the psychology of criminal conduct and prominently featured by Andrews and Bonta (1998). Characteristically, criminogenic needs are non-static attributes that, when influenced, contribute to a decreased chance of recidivism (Ward & Stewart, 2003, p. 127). They include, for example, "pro-offending attitudes and values, aspects of antisocial personality (e.g., impulsiveness), poor problem solving, substance abuse, high hostility and anger, and criminal associates" (Ward & Stewart, 2003, p. 127). These factors should be distinguished from static risk factors like gender, age, or criminal history, which, while important for initial risk assessments, are of lesser significance for treatment decisions (Ward & Stewart, 2003, p. 127). In this section, I will discuss the theoretical basis of these criminological needs and argue that their framing as theoretically well founded is not sufficiently justified. The latter entails that COMPAS cannot be justifiably regarded as a measurement tool for criminogenic psychometric properties. Instead, the lack of theoretically founded causal links between criminogenic needs and delinquent behaviour, as well as the confounding influences between different risk factors, make the RNR model a good candidate for machine learning techniques *precisely because* its lack of theory harmonizes with the atheoretical essence of machine learning tools. What this atheoretical essence consists in will be more closely described in the next section. For now, I focus on the lack of theoretically founded causal links underlying the RNR model.

The original theory proposed by Andrews and Bonta (1998) does not contain an account of how the criminogenic needs interrelate and influence one another (Ward & Stewart, 2003, p. 130). Their "dustbowl empiricism"[14] approach identified needs and risk factors as merely contributing to or subtracting from the potential of criminal behaviour, but it is not clear how they stand in relation to one another. This point is picked up in greater detail by Walters (2017) where he criticises Andrews and Bonta's (1998) RNR model precisely for this shortcoming. As a way of establishing causal

[13] E.g.: "Construct validity: it must measure the criminogenic needs it purports to measure; for example, it should relate coherently to other measures of needs and capture change in risk state over time" (Eno Louden & Skeem, 2007, p. 5)

[14] "When a single theory fails to emerge (as is inevitable), empiricists tend to reject the value of theory entirely and focus energy exclusively on the collection of data. Declaring a moratorium on theory - Alfred North Whitehead's "dustbowl empiricism" - is a recurring phenomenon in the history of social science […]. Dustbowl empiricism is characterized by what Feyerabend (1975) described as the rhetorical bullying that is implicit in appeals to rationality and evidence." (Suddaby, 2014, p. 408)

connections between different risk factors, he suggests *Causal Mediation Analysis*, a statistical methodology that introduces a third, mediating variable between two correlated variables. As an example, one of the best predictors of future physical aggression is past physical aggression. However, there is no clear causal link between those two phenomena. How exactly does past physical aggression lead to future physical aggression? By introducing a third variable, one can test the statistical validity of intermediate relationships and discover potential causal links between different phenomena. By introducing, in this case, the variable "positive attitude towards physical aggression" into the past-future physical aggression axis, Walters (2017) was able to identify a statistically relevant mediating variable between the two (Walters, 2017, pp. 51-52). In this way, one can begin to build a theoretical framework that connects the criminogenic needs to one another and design intervention strategies that tackle specifically the causal links that lead to criminal behaviour.

The importance lies in the link between measurement and theory. The main shortcoming of Andrews and Bonta's (1998) RNR model for measuring criminogenic needs and developing treatment for these needs is its lack of theoretical grounding. This is mainly due to the fact that Andrews and Bonta (1998) applied a *dustbowl empiricism* approach for determining factors that predict recidivism. This approach understands itself as *atheoretical*, contending itself merely with the empirical identification of statistically significant variables for a given target. However, such a conceptualisation of atheoretical research and model construction is importantly ill-conceived. It is reminiscent of a publication by Anderson (2008) which declared the end of the scientific method in favour of powerful big data applications. Many scholars have responded critically to such viewpoints (Boon, 2020; Calude & Longo, 2017; Kitchin, 2014) and highlighted the fact that theory-agnostic approaches in scientific research do not really exist. Rather, claims of theory-agnosticism have the unfavourable side-effect of rendering implicit assumptions and value laden judgements hidden and opaque (Suddaby, 2014, p. 408).

Whether one relies on further statistical tests or builds upon other Need models (e.g., Deci et al., 2001; Ward & Stewart, 2003) or evolutionary theory (Barkow et al., 1995), the criticisms of the RNR model indicate that a more developed theory of criminogenic behaviour and risk factors is required. Relying on a machine learning tool like COMPAS when considering the current theoretical shortcomings of the risk-needs model is problematic because it conveniently hides the lack of theoretical foundations (and, therefore, explanations and understanding) behind a wall of statistical relations. What this wall of statistical relations looks like and how it appears to circumvent theoretical considerations will be presented in the next section.

2.3. COMPAS as *prediction* tool

According to Tal (2012), measurement accuracy is a special case of predictive accuracy (Tal, 2012, p. 177). This is the case because he conceptualises prediction – specifically, prediction from instrument

indications to the measurement outcome – as an integral part of a measurement process. In fact, the measurement process is, in this view, seen as the result of a calibration process, whereas calibration, in turn, is a modelling procedure which aims at modelling the measurement process. During this calibration procedure, varying assumptions and idealizations will affect the reliability and generalizability of the measurement process[15]. One special case of such a calibration process is "black-box calibration" where the measuring instrument is treated as a mere input-output device: its inner workings and environmental influences are either neglected or heavily simplified (Tal, 2017, p. 36). Rather, in such cases the focus lies on establishing stable correlations between the instrument indications and measurement outcomes. However, such a correlation is not necessary and sufficient for *all* kinds of measurement processes. During "white-box calibration", the calibration procedure, in addition to the correlation between instrument indication and measurement outcome, establishes a stable correlation between the instrument indications and "the *predictions of an idealized model* of the measurement process" (Tal, 2012, p. 160, emphasis in original).

Machine learning tools have been notoriously problematized for their "black-box" nature (e.g., Rudin, 2019; Rudin et al., 2020; Sullivan, 2022; Carabantes, 2020; Krishnan, 2020). Looking at the difference between black-box and white-box calibration, it becomes clearer why machine learning tools struggle to shed the black-box label. The internal workings of these types of algorithms are opaque and indecipherable for human agents to the degree that they necessarily have to be neglected or simplified when providing an explanation of their functioning. In other words, black-box calibration is the only possible kind of calibration feasible for machine learning algorithms. Abstracting away the inner working of the tool is not a choice by the developers but a necessity stemming from the structure of these algorithms.

Srećković et al. (2022) lay out the two ways machine learning algorithms ban explanations by distinguishing between two *explananda* without *explanantia*: the *process* and the *phenomenon* (Srećković et al., 2022, p. 161). By *process* they refer to the internal algorithmic pathway between input and output. The process of a machine learning algorithm has two distinct features that render it opaque: *semi-autonomy* and *complexity* (Srećković et al., 2022, pp. 162-163). Semi-autonomy refers to the idea that machine learning algorithms automatically adjust the weights attributed to the input variables according to patterns observed in the training data[16] (Srećković et al., 2022, p. 162). In the end, not even the engineers themselves are able to tell why a particular weight has a certain value. By complexity, the authors mean the complexity of the information paths inside the algorithm, which

[15] For instance, "one-way white-box calibration" assumes that the behaviour of the measurement standard used for the calibration is perfectly predictable and "black-box calibration" additionally assumes that the mapping from instrument indications to measurement outcomes is unaffected by fluctuations in external circumstances (Tal, 2012, p. 176).
[16] Note that such patterns may be biased or spurious correlations.

makes it generally impossible for any human agent to track the flow of information from layer to layer (Srećković et al., 2022, p. 163).

Besides the process, the authors also claim that the *phenomenon* that is being predicted with a machine learning algorithm is an *explanandum* without *explanans*. This is due to the *associativity* of machine learning techniques; the model a machine learning algorithm creates is based solely on the data and it is hard to determine whether the associations surpass mere correlations (Srećković et al., 2022, pp. 163-164). Fundamentally, it is typically considered that machine learning algorithms are not capable of identifying causal, or otherwise explanatorily valuable relations between data points. That is the reason why they are often described as atheoretical in sense Anderson (2008) envisions.

If we consider COMPAS therefore strictly as a predictive algorithm, it runs into the inherent problems of explanatory opacity that concern many machine learning algorithms[17]. Both the semi-autonomy during the training based on data from various US legislations and the complexity with which the 137 input values of the algorithm are weighted against one another render it opaque and constitute significant hurdles for meaningful explanations. The most famous court case around COMPAS was the *State of Wisconsin v. Loomis*, in which Loomis appealed against the use of risk assessment software in court on grounds of violations of due process (Washington, 2018). The court dismissed the appeal by arguing that all the parties involved had access to the input data for the algorithm and could assess its accuracy. However, critics subsequently remarked that the court had "ignored the computational procedures that processed the input data" (Washington, 2018, p. 134) and that the mere accuracy of the input data reflected too low of a bar to justify the use of such risk assessment algorithms (Washington, 2018, p. 159).

Washington's (2018) assessment that a more plausible and stronger appeal against the use of risk assessment algorithms in courts could have been made if Loomis had put more weight on the opaque way the data of a defendant is processed inside the algorithm is thus in line with what Srećković et al. (2022) have to say about the lack of explanation that goes hand in hand with the fundamental way machine learning algorithms are designed. Sullivan (2022), on the contrary, argued that it is possible to obtain explanations and understanding from machine learning algorithms by decreasing the *link uncertainty* between the target phenomenon and the model. By link uncertainty, she means "a lack of scientific and empirical evidence supporting the link connecting the model to the target phenomenon" (Sullivan, 2022, p. 21). On her account, a machine learning model obtains greater epistemic validity as it replicates empirical findings. The greater the correspondence between the ML model and the empirical findings, the greater the chance that we can obtain explanations from a machine learning model about the target phenomenon.

[17] Some Interpretable and Explainable AI techniques may be exceptions.

However, my main contention with her account is that one cannot create a connection between a machine learning model and the target phenomenon due to the conditions listed by Srećković et al. (2022). Semi-autonomy and complexity are conditions that prohibit a meaningful description of what the machine learning model *consists in in the first place*. Associativity entails that the patterns identified by the algorithm can be caused by biases in the datasets or be entirely spurious correlations that do not indicate any causal or otherwise meaningful relation. Only further investigation into correlations identified by algorithms[18], e.g., empirical experiments determining causal links between certain phenomena, may reveal whether it constitutes a meaningful piece of information. However, in that case, it is still not true that the ML model itself serves as the explanatory medium, but the scientific theory that is enriched by a new piece of evidence.

The main problem with predictions from machine learning algorithms, as opposed to, say, predictive methods from more traditional statistics is that they contain hurdles preventing meaningful explanations about both the internal process and the target phenomenon that arise precisely because of the way machine learning algorithms work: they are semi-autonomous and complex, and derive patterns from mere associations. In essence, they are *pseudo-atheoretical association-detectors* because they detect relevant patterns and associations in datasets and are presented as theoretically neutral tools while, in reality, their application implies tacit theoretical assumptions[19]. In some instances, it may not matter that a machine learning algorithm cannot provide explanations for its predictions, as long as the latter have a high degree of accuracy. For instance, in the case of COMPAS, one might argue that it is of primary importance to prevent further harm from potential reoffenders to society and therefore should resort to any tool that increases the chances of making more accurate predictions. However, philosophers of science (e.g. Boon, 2020) have provided similar accounts as Srećković et al. (2022) where they problematize the epistemic role of these algorithms in the context of scientific knowledge generation. Such accounts often invoke the value of explanations and understanding, as well as the atheoretical nature of machine learning algorithms. This value of explanation and understanding is precisely what is invoked by Washington (2018) in her analysis of the case *State of Wisconsin v. Loomis* where she argues that the opacity of the COMPAS algorithm should have been considered in the appeal on grounds of violation of procedural due process.

What the arguments above suggest is that machine learning algorithms are, due to their very structure, unsuited for high-stakes decision making processes that require meaningful explanations for the

[18] Sullivan refers to these as *how-possibly* explanations (Sullivan, 2022, p. 20).
[19] I call them "pseudo-atheoretical" because they are only seemingly devoid of theory, whereas, in actuality, the use and application of machine learning for a given purpose already implies a set of implicit, theoretical assumptions, like for instance that the target phenomenon *can be* reliably captured by a machine learning tool. Furthermore, training datasets may have structural properties (biases) due to, for example, sampling errors which transmit value laden assumptions into the encoding of the algorithm.

output they provide. The main issue with COMPAS as a predictive tool is that the inference from the output of the algorithm to the conclusion that a certain individual will with a certain likelihood reoffend is very difficult to justify because of the internal opacity of the algorithm. The fact that COMPAS is doubly opaque, both in terms of trade secrecy and algorithmic opacity makes the use of the algorithm, in my view, very hard to justify both from an epistemic and ethical perspective.

Reviewing the present section and the one above, we find therefore that COMPAS purports both to *measure* and to *predict*. Specifically, it aims to measure criminogenic needs of offenders and infers a risk of future reoffence based on these needs in connection with other relevant factors. However, as things stand now, COMPAS' measurement scales for criminogenic needs are not sufficiently validated with other existing measurement tools while the predictive inference lacks any causal explanation and rests solely on statistical justifications. Furthermore, the account of this sections highlights that both the internal process of ML algorithms and the phenomenon they attempt to predict remain opaque. Taken together, I find that COMPAS is an opaque tool that bars any meaningful explanation about the highly contentious topic of crime prediction and is developed on the basis of a theoretically lacking conceptual model of recidivism. It appears therefore that opaque machine learning algorithms and weak theory work well together for the wrong reason: machine learning conceals the lack of theoretical foundations and renders elucidating investigations close to impossible due to its very structure. In what follows, I discuss the connection between machine learning prediction and measurement along three publications from philosophy of measurement: Mussgnug (2022), Parker (2017), and Mitchell (2020). As announced in section 2.1., each account serves as a vantage point to discuss the epistemic role of machine learning in measurement processes and will serve to underpin my discussion thus far.

2.4. Machine Learning Prediction vs. Measurement

One way of approaching the difference between machine learning prediction and measurement is by highlighting epistemic responsibilities connected to these practices. Mussgnug (2022) investigated how machine learning developers reframed the task of determining a poverty distribution in a certain area from measurement to prediction. Important about this reframing is that it is neither ethically nor epistemically neutral. The original poverty measurement task required a well-defined and validated metric. Essentially, the designers of this metric had to be well-aware of socioeconomic circumstances and be mindful of not (re-)creating unfair measuring outcomes by defining the metric in an unsuited way. It required thus both ethical and epistemic contemplations and justifications. Meanwhile, Mussgnug's (2022) observation was that, the moment machine learning experts took over the original measurement task, a problematic reframing took place where the justification of said metric was pushed under the rug. More than mere terminological convention, what machine learning experts called *poverty prediction*, was in effect the prediction of a *metric* whose suitability for certain contexts was no longer discussed.

This problematization of the predictive reframing is closely related to the issues surrounding the construct validity of recidivism measurement and the lack of causal explanations behind predictions outlined in the previous sections. While we do not find a reframing in the sense that there was a previously well-established measurement that was subsequently taken over by a predictive task, the same ethical and epistemic issues nevertheless arise because COMPAS constitutes a predictive task without sufficient justification and validation of the metrics used.

Mussgnug (2022) proposes to restore a more thorough contemplation of the metrics used in machine learning prediction tasks by reconceptualising the latter as *automatically-calibrated measurement instruments* (Mussgnug, 2022, p. 10). He motivates this move with the expectation that the subsumption of machine learning tools under the umbrella of measurement instruments would introduce the epistemic virtues of measurement practices into machine learning. The move is further justified by drawing from Tal's (2017) model-based account of calibration and Boumans' (2007) description of model-based measurements in the social sciences.

Tal's (2017) model-based account of calibration makes a difference between instrument indication and measurement outcome, essentially construing the relationship between the indication and the outcome as a calibration process which is conceived of as a model describing the inferential step from the indication to the actual measurement outcome and quantifying errors and uncertainties along the way. The modelling of this calibration process has two steps. The *forward calibration* step iteratively establishes a relation between the instrument indications and reference procedures. For instance, during the forward calibration of a calliper, gauge blocks are placed between its jaws in order to develop a mapping between the size of the gauge blocks as references to the indication provided by the calliper. The *backward calibration* step happens when one infers the measurement outcome from an instrument indication. The latter does not constitute the measurement outcome in and of itself. Rather, a set of background assumptions and quantifications of uncertainties accompany the instrument indication to arrive at an evaluation of the object to be measured.

Mussgnug (2022) draws important analogies between this two-part model-based account of measurement calibration and the way machine learning algorithms are trained and tested. Firstly, survey-based data and empirical forms of "passive observation" (Mussgnug, 2022, p. 14) used for the poverty prediction algorithm are similar to datasets commonly used for measurements in the social sciences. Furthermore, the iterative nature of the calibration procedure that Tal (2017) describes is also to be found in the way the parameters of the machine learning algorithm are adapted. The principle that guides the adaptation of these weights is that of a maximum likelihood estimation, meaning that the weights are adjusted such that the ML model best predicts the data. Again, such maximum likelihood estimations find applications in other social science domains, like econometrics (Mussgnug, 2022, p. 14). This would conclude what one could call the forward calibration of the

machine learning algorithm. Subsequently, Mussgnug (2022) mentions the similarity that the machine learning model is tested on a previously unseen dataset while a calibrated measurement instrument is applied to other reference objects for evaluation. If machine learning developers or measurement experts are unsatisfied with the inferences made from the instrument, they go back to revise the forward calibration process.

These analogies constitute the most intuitive similarities between machine learning and measurement. Mussgnug's (2022) motivation of treating machine learning tools like *automatically-calibrated measurement instruments* is commendable because he seeks to imbue the practice of developing and applying these tools with a greater care regarding the choice and justification of the metric that is being used. However, some issues remain.

Firstly, a simple reconceptualization of machine learning as measurement will not change the fact that machine learning has developed in a discipline different from measurement, with different paradigms and different core understandings of their practices. Machine learning is a quite young field of study and if one follows the heated discussions, both in the field of computer science itself and in philosophy, surrounding the technologies that have emerged and are currently emerging from that field, one would rather conclude that the paradigms[20] guiding machine learning are heavily scrutinized and still under a process of formation and solidification.

Secondly, Mussgnug's (2022) reconceptualization needs to be justified beyond the ethical-epistemological demand that machine learning developers adopt some best practices form measurement experts. The analogy between machine learning and measurement, although present in the more or less superficial manner that Mussgnug (2022, pp. 14-15) describes, is problematic in multiple ways.

The most pressing difference between machine learning and how we commonly understand measurement instruments is the computational, as opposed to the physical, medium. The model of the ML algorithm exists purely in a digital, virtual form which additionally entails that the input fed into it cannot be a physical object as is the case for most common measurement instruments, but rather a *model* of the phenomenon in form of data. In fact, the way a data set about a desired phenomenon is created is accompanied by its own modelling assumptions regarding e.g., the collection, storage, and representation of information. The fact that the input is a mere digital representation of the

[20] "As a field of study, machine learning sits at the crossroads of computer science, statistics and a variety of other disciplines concerned with automatic improvement over time, and inference and decision-making under uncertainty. Related disciplines include the psychological study of human learning, the study of evolution, adaptive control theory, the study of educational practices, neuroscience, organizational behavior, and economics. Although the past decade has seen increased crosstalk with these other fields, we are just beginning to tap the potential synergies and the diversity of formalisms and experimental methods used across these multiple fields for studying systems that improve with experience." (Jordan & Mitchell, 2015, p. 256)

phenomenon, in turn, entails that the model the machine learning adopts cannot be a model of a physical object, but rather, a model of the *data* concerning that object (Tal, 2023, p. 317). One may object that the virtuality of the machine learning algorithm might not be sufficient to disqualify it as a measurement instrument on the whole. In fact, Parker (2017) investigates at length the role computer simulations play in measurement procedures. She concludes that "computer simulations can be embedded in measurement practices in such a way that simulation results constitute measurement outcomes" (Parker, 2017, p. 274). Important to note here is the word *embedded* – Parker (2017) notes further that "[m]easuring is an activity that involves, among other things, physical interaction with the system being measured" (Parker, 2017, p. 285) leading to the important statement that "[a] computer simulation, on its own, is not a process for measuring properties of the system being simulated." (Parker, 2017, p. 285).

So, even according to Parker (2017), a digital instrument cannot in and of itself fulfil the role of a measuring process. Nevertheless, it is interesting to pay closer attention to the role digital mediums *can* play in measurement processes. As an example, she raises the task of measuring the temperature of a cup of tea or coffee by inserting a thermometer into the cup. The problem is that the insertion of the thermometer will influence the temperature of the cup, leading to a discrepancy between the instrument outcome on the thermometer (neglecting, for the sake of example, other noisy interferences) and the actual property of interest. A corrective function is therefore needed to quantify the influence of the thermometer on the measurement, calculating the heat transfer that occurred between the instrument and the coffee in order to arrive at the measurement outcome. Sometimes, the function describing this corrective step is straightforward and can be calculated directly. In different, more complex circumstances, a computer simulation might be needed to arrive at this corrective step. In the latter cases, computer simulations therefore play the role of corrective error adjustments that bridge the gap between the raw instrument indication and the proper measurement outcome. The outcome of the simulation is, in effect, the measurement outcome.

At the same time, Parker (2017) notes that the output of computer simulations can also serve as raw instrument indications. For instance, she asks us to imagine a measurement procedure supposed to determine the positions of some celestial bodies in our solar system four or five months ago based on current measurements of their position, mass, velocity, etc., and a Newtonian model of motion. In such a case, the result of a simulation can be either an instrument indication or the measurement outcome depending on how exact the model and the corrections inside the simulations are. If the simulation is heavily simplified, numeric corrections are required afterwards to correct the output of the simulation. In this version, the simulation does indeed provide raw instrument indications, rather than a measurement outcome.

Looking both at the poverty prediction algorithm described by Mussgnug (2022) and at the COMPAS algorithm, we see that neither of them serve an error correction function for an instrument indication. Rather, what they provide can better be described as raw instrument indications which are unadjusted for any potential errors. I would further argue that it is a feature of machine learning algorithms that they are difficult to adjust for potential errors precisely because the statistical model constructed to map inputs and outputs is so opaque that the development of any meaningful corrective adjustment function is prohibited by the impossibility of understanding the machine learning model which led to the instrument indication in the first place. To illustrate through the example of the thermometer in the coffee mug: the process leading to the instrument indication of the thermometer is clearly understood through well-established laws of heat transfer between different mediums. It is practically feasible, therefore, to investigate the sources of errors and uncertainty and develop a corrective model for these. However, the process which led from the input parameters through the machine learning algorithm to its outputs is undecipherable. It is not clear why a particular weight has the value it has, and what that means, practically speaking. It is therefore also not clear, where and why the origin of potential errors in the outcome might arise. This renders the determination of a corrective function extremely difficult. It is because of the structural complexity of machine learning algorithms identified by Srećković et al. (2022) and outlined above – the semi-autonomy, complexity, and associativity – that machine learning algorithms cannot take the proper role of measurement instruments.

I agree with Mussgnug (2022) that in most machine learning applications, developers ought to be more concerned about the validity and justification of, firstly, the machine learning model in itself, and, secondly, of the conceptualisation of the phenomenon they predict. As measurement practices teach us, a measurand does not lie simply in the observable world, much less in a digital dataset, ready for the taking, extractable by something akin to a "statistical kraken". It has to be carefully constructed and validated; background assumptions have to be made explicit and associated uncertainties have to be quantified. The greater the stakes associated with the application, the greater the care that has to be dedicated to the design of the measurand. However, this normative maxim can be directed towards machine learning developers without resorting to a reconceptualization of machine learning as measurement. As the discussion above highlights, the effective analogies between machine learning and measurement instruments are not sufficient to warrant this equation. Instead, I would argue that greater epistemic responsibilities can be expected from machine learning developers simply by pointing at measurement practices as guiding examples.

Lastly, Mussgnug (2022) justifies part of the analogy between machine learning and measurement by referring to measurements in the *social sciences*. However, as I mention in the previous sections, the social sciences often struggle themselves with evaluative standards like construct validity and universalizable models. The risk-need-receptivity model used for COMPAS, while resting on an

empirical basis, lacks descriptions of the interaction between the different parameters that predict recidivism as well as a causal account of how the presence of a particular feature leads to recidivism. For this reason, I see the comparison of machine learning to measurement in social science less as a justification and more as a similarity between two domains that are riddled with similar core issues of validity and justification.

Mitchell (2020) investigates adjacent issues of trust and reliability by comparing machine learning tools to NMR spectroscopy in the context of knowledge generation. She adopts an instrumental perspectivist stance which states that instruments that are used in scientific research knowledge production, together with their associated models, constitute a certain perspective on the target phenomenon. She opposes views that idealize the representative aspect of models, criticizing the notion on the basis that the ultimate level of representation of a phenomenon would be an exact copy of that phenomenon, recreating the original problem regarding its accessibility. Models should not represent *as many aspects as possible*, but what is important, instead, is *which aspects* are represented. She concludes that different tools and models therefore constitute different perspectives based on the aspects they incorporate and that a scientist may therefore use many different ones for a given purpose as long as each individual one internally respects some standards of reliability and trustworthiness.

While Mitchell (2020) acknowledges that machine learning and especially more advanced AI applications differ from other tools in that the rules they establish internally to map the target phenomenon during their learning phase reflect an epistemology which is unintelligible for humans, she nevertheless maintains that their warrant is comparable to that of NMR spectroscopy in terms of reliability and trustworthiness, such that they constitute another perspective in the repertoire of any scientist. However, in line with the argumentation outlined in this chapter, I cannot accept machine learning application simply as another perspective on a phenomenon amongst others. I have to reject instrumentalism as a useful perspective to analyse machine learning tools. As Heidegger already so prominently stated "So long as we represent technology as an instrument, we remain transfixed in the will to master it." (Heidegger, [1993] 2008, p. 316). My concern is that instrumentalism is too uncritical of a lens, especially for newer technologies like machine learning, such that its proponents remain too fixed on how to make machine learning tools work rather than when, how, or if they should be used at all.

I do not state that Mitchell's (2020) thesis is *incorrect* per se. However, I will state that her frame of analysis does not allow for a proper engagement with the problems that underlie machine learning and AI applications and which I outlined thus far. Especially the issue that the latter are often framed as "atheoretical" is in contradiction with her own insistence that "just as there is no independent-of-theory test of a single measurement, neither is that [sic] any independent-of-theory calibration." (Mitchell, 2020, p. 19). As I outlined above, Andrews and Bonta's (1998) RNR model was

intentionally constructed with a methodology that was framed as atheoretical and many philosophers of science have criticized the "end-of-theory" mindset that went along with an overly enthusiastic and naïve view of machine learning technologies. That epistemological considerations in the development and application of technologies (and their models or underlying theories) go hand in hand with ethical choices is a theme that underlies most of the present chapter and which will be further reinforced in the next and last chapter where I consider the fairness problems that govern the COMPAS algorithm.

2.5. Conclusion

Both in terms of measuring psychological properties and predicting the likelihood of recidivism, the COMPAS algorithm exhibits epistemic problems that are tightly connected to ethical issues. The risk-need-receptivity model developed by Andrews and Bonta (1998) is supposed to capture attributes of defendants that indicate statistically relevant risk factors for increased recidivism. However, the *dustbowl empiricism* approach they deployed for the construction of the model was framed as an atheoretical methodology designed to identify supposedly objective statistically relevant influences. As many philosophers of science have argued in recent years, there is no such thing as an atheoretical lens. This renders the methodology of Andrews and Bonta (1998) problematic because it conceals implicit assumptions in their methodology rather than opening it up for scrutiny. Furthermore, focusing solely on a list of factors that indicated some predictive relevance for recidivism risk assessment should not be favoured over a holistic theory that contains information about causal mechanism and confounding factors. The present chapter argued that, from an epistemological and ethical perspective, understanding about how recidivism occurs is important, just as explanations about how an algorithm arrives at a certain output is important.

The fact that a machine learning algorithm like COMPAS was used in tandem with such a supposedly atheoretical framework is no coincidence. The "end-of-theory" mindset some enthusiastic and naïve proponents of machine learning and AI espouse features exactly this idea that all that is required for knowledge generation is heavy statistical models bolstered by the newest neural nets. I conclude therefore that the link between machine learning and the RNR model in COMPAS is a perfect match for the wrong reasons. The methodology framed as atheoretical goes hand in hand with the inherent nature of machine learning models as *pseudo-atheoretical association detectors*. Furthermore, the way the input features are linked amongst one another and mapped to the output does not reflect any humanly intelligible epistemology and serves as a convenient cover up for the lack of theoretical foundations with respect to the different risks and needs of offenders and how they relate to one another. From the present analysis, it already transpires that epistemology and ethics are tightly bound to one another within the development and application of technologies and their respective theories and models. In the next and last chapter, I will focus on the fairness issues that govern many of COMPAS' outputs and discuss existing counterfactual approaches to fairness and explanation in machine learning as well as a potential solution from the literature of philosophy of measurement.

Chapter 3: Counterfactual approaches to fairness problems

In this chapter I will delve more deeply into the fairness issues identified in the first chapter. As mentioned in chapter 2, defining the measurand of the COMPAS algorithm proved epistemically problematic both from the perspective of measuring and predicting. These epistemic concerns cannot be separated from ethical concerns regarding the definition of the measurand. The purpose of this chapter will therefore be to investigate how we can predict a property like the propensity to recidivate fairly in an unfair world and will first and foremost focus on counterfactual approaches to fairness. To address the argument that a more profound rethinking of statistical fairness is required, *counterfactual prediction* will be presented the way it is featured by Tal (2023) in the context of medical diagnoses. The main notion that Tal (2023) defends is that accuracy and fairness ought not to be considered orthogonal notions; rather, fair prediction is the truly desired goal of the tools in question which implies that unfair tools are simultaneously inaccurate because they miss their intended target. I compare his ideas to frameworks about counterfactual explanations and fairness from the machine learning literature.

3.1. Introduction to Counterfactuals

Generally speaking, the target function[21] in machine learning based risk assessment can be said to be implicitly based on a non-ideal point of departure. The trade-offs in recidivism risk assessment, for instance, occur because the base rate of recidivism differs from one social group to another (Kleinberg et al., 2016). In machine learning, a paradigm seemingly taken for granted is that the trained algorithm is supposed to reflect or represent aspects of the training data. In classification tasks, the aspects in question are relevant features of the data used for training the model such that target labels are correctly assigned to the inputs according to some metric of accuracy. There is therefore a condition which dictates that the algorithm should capture relevant features and their respective distributions in the way they are present in the dataset. It appears that it is because this condition in machine learning is taken for granted that we occasionally obtain a discrepancy between what the algorithm outputs and what stakeholders desire. Taking for granted that the ML model accurately captures the idiosyncrasies of the training data could potentially be entirely at odds with the actual desires of stakeholders because the feature distributions in the dataset may either reflect injustices one would not want the algorithm to exhibit or lead to the learning of false rules (for example, that belonging to a certain race leads to higher criminality).

Tal (2023) exemplifies this in the case of disease prediction in health care which is riddled with fairness issues similar to those that occur in the COMPAS algorithm – persistent, systematic, and disproportionate error rates for different identity groups, be it through markers of race or sex. These

[21] By target function I mean the intended prediction target that is operationalized by a machine learning model.

discrepancies arise due to features inherent in the datasets: As Kleinberg et al. (2016) demonstrated, it is mathematically impossible to respect the calibration condition[22] and the balance conditions for the positive and negative class[23] all at the same time *if the base rates of recidivism differ between groups* (Kleinberg et al., 2016, p. 5). I have provided the informal overview of the proof in Appendix B. For the detailed proof, see (Kleinberg et al., 2016, pp. 9-12). The discrepancies in the predictions are unacceptable as they, in a legal context, violate principles of equal treatment and justice, and, in the case of health care, could lead to dangerous health implications in particular for minority groups. Such negative consequences are, of course, not desired by the stakeholders using or affected by the algorithms in question. If these discrepancies arise due to unequal base rates in the dataset, ought one not perhaps design a tool based on an idealised approach where these trade-offs do not occur? Instead, it may be necessary to formulate *counterfactual scenarios* in order to correct for biases.

A counterfactual scenario is a hypothetical alternative to an observed scenario. There are multiple reasons to be interested in such hypothetical scenarios. As I will explain in greater detail in the next chapter, *counterfactual explanations* in machine learning aim at providing the user with explanations by describing how a given output would have changed if certain inputs had been different. In a similar way, *counterfactual fairness* approaches list as their primary fairness condition that a given prediction remain the same if one swaps out a sensitive, protected variable like race, sex, or gender. Counterfactuals also play a role in measurement: the definition of measurands often include idealized scenarios that are only approximately approachable in practice (Tal, 2023, p. 316). In the next section, I will introduce existing frameworks for counterfactual explanations and counterfactual fairness and subsequently contrast them to Tal's (2023) approach. I finish by discussing whether counterfactuals are able to address the fairness issues presented in the first chapter.

3.2. Counterfactuals – One Concept Among Many?

In recent years, counterfactuals have received significant attention in the domain of ethical machine learning, specifically in the form of counterfactual explanations and counterfactual fairness. While these two approaches can be technically succinct in the sense that a counterfactually fair algorithm does not have to offer explanations and an algorithm that offers counterfactual explanations does not require it to be fair, they are both based on Pearls (2000) theoretical work on causality. Prominent examples for each approach are Kusner et al. (2017) for counterfactual fairness and Wachter et al. (2017) for counterfactual explanations.

In all brevity, a causal model based on Pearl's (2000) work features a *directed, acyclic graph* (DAG) with parameters (U, V, F), where U refers to unobservable background variables and V to observable

[22] The calibration criterion holds that for a given classification (e.g., "high risk of recidivism"), the probability of truly belonging to this classification should be equal for all groups.
[23] Equal false positive, resp., false negative error rates for different groups.

variables. It is important that no variable in V causes any variable in U. F is a set of *structural equations* (Bollen, 1989) for each $V_i \in V$ describing the relationship between parent nodes (i.e., nodes in the graph preceding or pointing at the variable V_i), background variables and V_i. Kusner et al. (2017) additionally distinguish between *A* and *X* where *A* are protected, and *X* are the remaining observable variables ($V \equiv A \cup X$). They then propose the following definition of counterfactual fairness:

"Definition 5 (Counterfactual fairness). Predictor $\hat{Y}$ is **counterfactually fair** if under any context $X = x$ and $A = a$,

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a), (1)$$

for all y and for any value a′ attainable by A." (Kusner et al., 2017, p. 3)

To understand this equation, note that the only difference between the expressions on either side of the equal sign are the terms $A \leftarrow a$ and $A \leftarrow a'$. It states that the probability P that the predictor $\hat{Y}$ outputs the classification y for a given individual, given that (indicated by the symbol "|") this individual has the feature vector $X = x$ and the protected attribute $A = a$ remains the same even if we change a for a' in the predictor. The authors note that the counterfactual definition of fairness sets itself apart from more conventional definitions like *fairness through unawareness* (simply leaving out the label of race at the moment of training), *individual fairness* (requiring that similar instances receive similar predictions), *demographic parity* (requiring that predictions are independent of the protected attribute), and *equality of opportunity* (requiring that probabilities are equal for different groups) because it enables the modelling of historical biases by explicitly incorporating them in the causal model (Kusner et al., 2017, p. 3). However, Rosenblatt and Witter (2023) prove that Kusner et al.'s (2017) definition of counterfactual fairness is equivalent to demographic parity and additionally demonstrate that Kusner et al.'s (2017) definition does not respect in-group ordering. The latter requires that, given a rank ordering of individuals along, say, their likelihood for recidivism in an unfair world, we would expect that the ordering would be preserved in the counterfactual scenario of a fair world since all individuals are equally affected by the unfair background conditions. However, all three levels of counterfactual fairness proposed by Kusner et al. (2017, p. 7) provide varying in-group orderings that differ wildly both from one another and from the initial ordering. This furthers the point, stressed in the first chapter, that any formulaic definition of fairness appears to invariably attract unwelcome side effects, increasingly putting into question the entire project of creating fair algorithms.

In order to further elucidate the idea of counterfactuals in machine learning, I want to discuss counterfactual *explanations* because they rely on the same model for causality as Kusner et al.'s (2017) counterfactual fairness approach, although counterfactual explanations are not strictly

necessary for fairness in the sense of equal predictions for different groups[24]. Rather, because counterfactual fairness relies on the same causal model as counterfactual explanations, it adopts the same assumptions about the causal structure of the target phenomenon. Further expanding on the counterfactual understanding of causality will provide better insides about the underlying assumptions of counterfactual fairness.

With counterfactual explanations one is more concerned with the explainability of machine learning models. Prominent in explainable AI (XAI), counterfactual explanations provide a sense about how the outcome would have changed in case some inputs had been different. They can take the form "Score p was returned because variables V had values $(v_1, v_2, \dots)$ associated with them. If V instead had values $(v_1', v_2', \dots)$, and all other variables had remained constant, score p' would have been returned." (Wachter et al., 2017, p. 848) Causal structures in the algorithm are highlighted in order to identify the set of input variables that led to the output at hand. In turn, one would then be able to state that, had (some of) these inputs changed, the output would have changed in a predictable manner. Identifying these relevant inputs requires comparing the given output to closely related scenarios (or *close-enough-possible-worlds*) with different outputs where the inputs were only minimally changed.

One problem with determining close-enough-possible-worlds is that the approach relies on a distance function to determine how closely related different worlds are. Defining such a distance function is difficult to justify (Kasirzadeh & Smart, 2021, p. 233). For example, it is conceivable that a recidivism prediction algorithm identifies the most salient factors for a high-risk prediction as the offender having had a low age at first contact with law enforcement and living in a high crime neighbourhood. It offers the counterfactual explanation that the offender would have received a lower risk classification if their age at first contact had been higher or if they had lived further away from the neighbourhood with a high crime rate. How can we decide which counterfactual world is "closer" to the original? Comparing the two proposed scenarios seems arbitrary, not only because it is difficult to define a metric by which one could compare the two, but also because the counterfactual scenario in which the offender had lived in a different neighbourhood could potentially have completely changed the offender's life to the degree that the counterfactual scenario is entirely unlike the original one.

Wachter et al.'s (2017) model of counterfactual explanation furthermore sidelines questions regarding the internal functioning of the algorithms and does not contend with the problem of opacity identified in chapter 2 (Wachter et al., 2017, pp. 845-846). Counterfactual explanations instead aim at providing information directly about the "dependency on the external facts that led to that decision." (Wachter et

[24] One may raise the point that explanations are necessary for fairness in the sense of due process before law, or in the context of medical advice for patients, but, while highly relevant, that is not the kind of fairness I discuss here.

al., 2017, p. 845) The authors acknowledge that opening and explaining the black box of machine learning algorithms is a large challenge and furthermore doubt that providing explanations about its functioning is helpful for a layperson. Counterfactual explanations are therefore presented as a means to reduce the regulatory burden for algorithms because they are simpler to obtain and give clear indications (Wachter et al., 2017, pp. 860-861). However, in this case it should be clearly distinguished between the causal model designed for an algorithm able to provide counterfactual explanations and the true causal model underlying the actual phenomenon of interest.

What becomes noticeable at this point is how causal structures in question are not equal for the phenomenon and the machine learning model supposed to provide explanations. Wachter et al.'s (2017) approach clearly targets the latter: the aim is to explain how differences in the input, which reflect real properties, would have changed the output of the algorithm. It is an attempt at rendering the algorithm more transparent without explaining the technical details of the algorithm itself. What has to be clearly highlighted, however, is that causal relations in reality do not necessarily follow counterfactual logic. Kohler-Hausmann (2019) explains at lengths how the factor race cannot be exchanged for a different value in real life situations while keeping all other factors constant. Such thinking treats race as an independent variable which relies on a biological conception of race rejected by constructivists (Kohler-Hausmann, 2019, p. 1169).

Constructivists argue that categories of race (at least in the US) have historically developed in contexts of domination and colonialism such that race as a concept has become a pervasive way of perceiving and experiencing the world. I hesitate to fully adopt the constructivist view at face value[25]. However, in my view, it demonstrates correctly that race is not subject to counterfactual logic since it does not function in the same way a treatment variable functions in, for example, medical studies testing the effectiveness of a particular intervention. In the latter case, the causal influence of the treatment is – provided the experiment setup allows for controlling for other factors – the difference in outcome between the group that received the treatment and the control group that did not. Such experiments are, according to Kohler-Hausmann's (2019) view not feasible with regard to race as a treatment variable. That means, it is simply not possible to create two comparable scenarios in which the only difference is race, because this would imply that race is reducible to factors like skin colour,

[25] While I do not have the space to expansively address the role of race in society, I hesitate to overemphasize its pervasiveness. This is because, in my view, it risks reinforcing the very differences constructivist scholars wish to resolve, since arguing that race is such a fundamental category of experience can easily lead to a version of "race determinism" which conceives individuals, and, in particular, black individuals, as mere pawns subject to the pervasive forces of a racialised society, denying individual agency and rendering a notion of progress in race relations close to impossible. I also believe that this view risks reducing explanations about disparate outcomes between social groups to a singular, universal factor of racial discrimination, where invoking other factors like, for instance, socio-economic status (although correlated) may be more illuminating.

whereas, according to the constructivist view, race influences the experiences and behaviours of individuals in many different ways[26].

The argument against the counterfactual aspect of race together with the criticism directed at the finding of a suitable distance functions for finding *close-enough-possible-worlds* and the flaws in the concept of counterfactual fairness all point at the difficulties of modelling social behaviour. Taking stock of the arguments presented in this chapter thus far, one may formulate the following tentative conclusions. Firstly, since fairness issues regarding the disparate treatment of different social groups by algorithmic decision making arise due to unequal base rates in these groups, the project of formulating mathematical definitions of fairness that could be implemented in algorithms are unlikely to deliver an ultimately satisfying solution. Rather, they are better conceived of as inherently flawed attempts at improving decision making processes in an unjust, complex world. The statistical fairness approaches reviewed in this thesis appear more like a drop of water on a hot stone. They merely contend with symptoms of underlying complex issues and will not solve the problem at the root.

The second conclusion I would like to draw is that counterfactual approaches struggle in particular when it comes to causality in the social sciences. For physical phenomena, one can usually draw concise causal relations. For instance, one can say that heat caused a metal bar to expand. With respect to social phenomena, whether macroscopic, historical events or individual behaviour, statements linking cause and effect are usually not as easy to formulate and are often heavily contested by different scholars (e.g., Chatterjee, 2017; Illari et al., 2011; Pearl, 2009; Salmon, 1998 for causality in more general terms, and Hedström & Ylikoski, 2010; Holland, 1986; MacIntyre & Korbut, 2013; Marini & Singer, 1988 for causality in the social sciences) . While an expansive review of the role of causality in social science is beyond the scope of this paper, suffice it to say that causality has been discussed by philosophers for centuries and that it poses particular problems to social scientists (Marini & Singer, 1988). It is further reinforced by Kohler-Hausmann's (2019) analysis of the factor race not following counterfactual logic. This problem poses a significant hurdle both to the development of a causal theory for recidivist behaviour and, by extension, for counterfactual models for machine learning like the ones introduced above. Turning to Tal's (2023) paper on counterfactual predictions I'm interested to see if his approach can handle these tentative conclusions and offer a way to address their problematic implications.

[26] Kohler-Hausmann (2019) discusses rebuttals to this view which, for reasons of space, I cannot go into in detail. Suffice it to say that, in her view, audit studies which attempt to control for the factor race in order to detect discrimination do not provide counterfactual explanations of race as a treatment variable, but rather "evidence of a constitutive claim that grounds a thick ethical evaluation" (Kohler-Hausmann, 2019, p. 1215). This means that audit studies do not demonstrably detect racial discrimination as a univariate explanation but rather provide empirical support for the existence of a complex, multifaceted ethical problem tied to race.

3.3. Counterfactual Target for Recidivism Prediction

Tal's (2023) point of departure is to compare the notion of accuracy commonly adopted in machine learning to the well-established metrological notion of accuracy. The former focuses on the rate at which the machine learning model correctly matches its predictions to the true label in the training dataset. The higher the rate at which the machine learning model assigns the correct label to a training instance, the more accurate it is deemed. Tal (2023) calls this notion of accuracy *label-matching conception of accuracy* (LMCA) (Tal, 2023, p. 313). The problem with using this notion of accuracy as a benchmark is that it may not agree with the what the stakeholders actually expect in a given scenario. For example, Tal (2023) presents the example where a machine learning model developed to help decide which pneumonia patients to hospitalize turned out to attribute a lower mortality rate to asthmatics. This was because the label the model was trained with was patient mortality, but asthmatics received more aggressive treatment early on, leading to a lower mortality rate for asthmatics. The model therefore picked up the rule that asthma *decreases* mortality rate (Tal, 2023, p. 314). Here, one can clearly see that stakeholders are not interested in a ML model that purely focuses on accurately matching the labels in the dataset – for that the model did well – but rather in predicting mortality rates *had patients received the same treatment*. They are interested in predicting patient mortality in a counterfactual scenario that does not correspond to the label the model was trained upon (Tal, 2023, p. 314). The predictions desired by stakeholders consist in predictions about an idealized scenario. The label is rather an *operationalization* of the ideal target in the sense that it is generally assumed that labels approximate the desired target sufficiently well. Tal (2023) refers to the bias in which the operationalized target is considered to be the ultimate benchmark for accuracy as *target specification bias* (Tal, 2023, p. 313). To remedy target specification bias, Tal (2023) suggests taking inspiration from the metrological conception of accuracy. In metrology, targets are generally idealized and considered inaccessible (Tal, 2023, p. 316). Realizations of measurands operationalize this ideal and are accompanied by information about possible biases and uncertainty to inform metrologists how reliable a measurement is with respect to the ideal target (Tal, 2023, p. 316).

It is worthwhile to mention at this point that the label-matching conception of accuracy was criticised for its insufficiencies in other contexts as well. Karaca mentions how cost-sensitive machine learning is used in the case of imbalanced classes (Karaca, 2021, pp. 13-14). Imbalanced classes are instances where the size of one of the classification categories vastly exceeds (i.e., by orders of magnitude) that of the other category. If uncorrected, such situations typically lead to significantly higher error rates for the minority classes because the model will adapt better to the majority class. In addition, minority classes are typically the more interesting cases in which a classification error weighs more heavily than for the majority class. For example, in cancer detection, patients with cancer are a minority class where a false-negative diagnosis can be far more harmful than a false positive diagnosis of a healthy individual. Cost-sensitive ML is therefore used to penalize errors for the minority class more heavily

such as to balance these discrepancies. Note however that in the context of COMPAS we don't find imbalanced classes in the sense of vastly disproportionate class sizes (both the number of black vs. white defendants and of recidivists vs. non-recidivists are of the same order of magnitude). The primary discrepancy between black and white defendants consists instead in different base-rates for recidivism. The kind of difference at the heart of the fairness issues discussed throughout this thesis is distinct from imbalanced classes.

In concrete terms, Tal (2023) is concerned with any machine learning application that exhibits unfair treatment for different groups. Unfairness is broadly conceived of as inequality in classification outcome along any significant metric and fairness as the remedy of these inequalities. Fairness does not have to consist specifically in equal false positive and false negative error rates, but, instead, different applications will weigh different kinds of inequalities in different ways. Verma and Rubin (2018), for instance, list twenty different fairness conditions in machine learning (Verma & Rubin, 2018, p. 2). The point that Tal (2023) makes is that machine learning applications are currently too transfixed on the label-matching conception of accuracy and he sees fairness issues as an at least partial result of the transfixion on this label-matching conception. This transfixion is the reason why a ML developer may state, for instance, that their application has fairness issues X, Y, and Z, but exhibits an eighty percent accuracy rate which makes it highly reliable, nevertheless. Tal (2023) rejects this two-part distinction between fairness issues on the one hand and (label-matching) accuracy on the other because it is *by accepting this distinction* that these fairness issues arise in the first place, and the field of machine learning fairness remains stuck.

In order to understand why this is the case, consider that the whole idea of training a machine learning classifier on a labelled dataset is that the dataset is representative of the target population and that its labels are sufficiently good approximations of the classification target. In other words, the dataset is an operationalization of the population and the labels an operationalization of the target function. The classifier receives its validity from the overall validity of this chain of operationalizations. For this reason, we can trust the label-matching conception of accuracy as long as it rests on a reliable chain of operationalizations. This view, however, misses the point that a representative dataset will necessarily include discrepancies and inequalities between different groups. As mentioned multiple times throughout this thesis, in the case of recidivism prediction, it is unequal recidivism base-rates between black and white defendants that unavoidably lead to unfair error rates during classification. It is therefore precisely because one accepts the dataset and labels as such as the operationalizations for the target that fairness trade-offs will become unavoidable.

To avoid the fairness-accuracy dichotomy, Tal (2023) invites us to consider what stakeholders actually expect from ML classifiers. Stakeholders do not want a recidivism estimation tool that exhibits different error rates for black and white defendants but one that treats each individual equally

and fairly. Focusing on the target the way stakeholders conceive of it, COMPAS already misses the mark widely. It is from this starting point that Tal (2023) develops the notion of *counterfactual targets* for machine learning applications. In short, this notion states that the desired target should be built from an idealized standpoint that does not allow fairness issues in the first place. It requires the developer to imagine an ideal scenario where the preconditions for fair classification (for COMPAS this means equal base-rates) are met.

The notion of counterfactual targets does not simply drop from the sky but builds on important parallels to ideal measurement targets common in the practice of complex measurement procedures. In measurement procedures, defining the quantity intended to be measured (the measurand) is distinguished from the task of realizing the measurement (Tal, 2023, p. 316). For instance, the SI unit for the second is defined as "the duration of exactly 9,192,631,770 periods of the electromagnetic radiation corresponding to the transition between two hyperfine levels of the unperturbed ground state of the cesium-133 atom" (Tal, 2023, p. 316) where it is assumed that the cesium atom is "unaffected by gravitational fields, magnetic fields, or thermal radiation, and to have no interactions with other atoms" (Tal, 2023, p. 316). Such conditions are practically unobtainable in any laboratory setting and are essentially counterfactual. Metrologists make use of a range of practical, theoretical, and statistical methods to approximately approach the ideal, for instance by practically setting the temperature as close to the ideal as possible or by developing statistical models that predict the frequency of the clock at zero density (Tal, 2023, p. 316).

By contrast, in machine learning, the transfixion on the label-matching conception of accuracy already maintains a focus on the operationalization of the measurand while its definition was not clearly developed. Believing in the dichotomy between accuracy and fairness already indicates a commitment to the LMCA and uncritically assumes it to be the best or only possible operationalization of the measurand. It represents a reversal of the procedural order compared to measurement: rather than departing from the ideal definition of the measurand towards realizations that approximate it, one is committed to a particular way of realizing the measurand and makes conclusions about the measurand on the basis of this operationalization. This is the reason why machine learning developers typically consider fairness and accuracy as orthogonal dimensions trading-off against one another: because the operationalization via the LMCA requires this conclusion.

Following Tal's (2023) suggestion to adopt a metrological conception of accuracy in machine learning recentres the practice of defining a measurand and loosens the grip that (seemingly) unavoidable trade-offs have on machine learning applications. In concrete terms, defining a measurand for recidivism prediction would consider that an ideal target does not allow for unequal outcomes for different social groups. Some predictive errors are practically unavoidable but that disparate error rates disproportionately affect disenfranchised groups is not acceptable. Simply put,

the desired target function is a reliable measure (implying calibration between groups) of an offender's likelihood of recidivism whose false positive and false negative error rates do not vary between groups. In other words, since disparate error rates are linked to unequal recidivism base-rates, the model has to provide predictions for an idealized, counterfactual scenario where the base-rates between groups are equal. The machine learning application needs to operationalize a realization of this measurand and enshrine these conditions.

Tal`s (2023) approach hence explicitly calls for value judgements in the formulation of the target function. At first glance, it could align itself well with a recent movement in philosophy of science that complicates the long purported epistemological objectivity of scientific practice (see for example Zecha (1992) for various formulations of the principle of value-neutrality) by demonstrating both the presence of and need for non-epistemic value judgements in science. Prominent examples of this movement include Douglas (2000) who argues that non-epistemic value judgements both do and should play an internal role in scientific research in order to mitigate inductive risk, and Karaca (2021) who adopts a similar line of argumentation for binary machine learning classifiers. Elliott and McKaughan (2014) demonstrate that non-epistemic value judgements may even override epistemic values and thus play a primary role in the adoption or rejection of theories and models, especially when it comes to considering the intents and purposes of users (Elliott & McKaughan, 2014, p. 4). Intemann (2015) goes a step further and argues that non-epistemic value judgements are "*legitimate* in climate modeling decisions insofar as they promote democratically endorsed epistemological and social aims of the research" (Intemann, 2015, p. 219, emphasis in original), thus concretizing the kinds of acceptable non-epistemic values in science.

Looking at the role that social values play in scientific practice, we can discuss the meaning of Tal's (2023) explicit call for non-epistemic value judgements for the target function and highlight some hurdles that need further exploration. One difference to be highlighted right away is that, in cases typically considered in the literature mentioned above, non-epistemic values play the role of mitigating inductive risk during the "choice of methodology, gathering and characterization of the data, and interpretation of the data" (Douglas, 2000, p. 565). That is, non-epistemic values help decide which methodological approach to adopt at certain stages because the problem of, for example, which model to choose for a particular application is generally underdetermined by the available data such that multiple competing models match the data equally well (Karaca, 2021, p. 5). Non-epistemic values can play the deciding role in such instances. Elliott and McKaughan (2014) exemplify this point at the hands of government agencies determining the carcinogenic properties of substances. They point out that researchers have to make a choice between an assessment method that is slow but accurate and one that is fast but less accurate. This choice reflects a non-epistemic value-judgement about the value of certain social costs (Elliott & McKaughan, 2014, p. 8).

Tal's (2023) metrological conception of accuracy in machine learning, by contrast, requires an explicit inclusion of social values into the very construct of the prediction target. Target specification becomes a task in itself which involves complex and difficult evaluations of values by many different stakeholders (Tal, 2023, p. 319). These stakeholders are typically interested in counterfactual scenarios that do not match the actual conditions under which the data for the prediction task was collected. For health outcome predictions, the learned rule that asthma *decreases* the mortality risk for pneumonia is technically correct if one solely looks at the training data, but entirely at odds with the target domain health professionals *actually* want predictions about. What they want to know is how efficient an intervention would be in a counterfactual scenario where an asthmatic had not received more intense treatment. This prediction target specified under counterfactual conditions is not a purely epistemological entity like the SI unit of the second. What I mean by this is that the latter refers to a purely physical phenomenon under ideal *physical conditions*. What Tal (2023) proposes for the definition of a counterfactual target function for health outcome prediction is predictions under ideal *socio-economic conditions*. It is necessary to make predictions under the assumption of such ideal social and economic conditions because it is precisely disparities in socio-economic makeup reflected in the training dataset that give rise to unfair predictive outcomes (Tal, 2023, p. 313).

By specifying the counterfactual target function under ideal socio-economic conditions, the work of the machine learning engineer becomes tightly entangled with understanding social problems and navigating ethical questions. The target function is no longer purely epistemological, like in metrology, but necessarily incorporates in its definition a vision of fairness and ideal socio-economic conditions. This goes well beyond what Douglas (2000) had in mind when it comes to values in scientific practice. Taking Tal's (2023) suggestion seriously, the target specification task requires machine learning developers to include the norms of the society affected by their prediction tools in the very aim of the tool. Tal (2023) highlights the need for a close collaboration with stakeholders in order to specify these conditions and clarifies that this task will very likely involve complex negotiations about different values[27]. In the example of the falsely learned rule that asthma decreases mortality chances for pneumonia, it is quite clear and uncontentious that fairness requires predictions about counterfactual scenarios where asthmatics would not receive more intense treatment. However, in more complex cases the task of target specification would require a clearer definition and methodology.

One way of concretizing the target specification task is through Richardson (1990) approach of specifying norms to solve concrete ethical problems. The strength of Richardson's (1990) approach is

[27] Karaca also clearly advocates for the inclusion of user values in the design of ML models (Karaca, 2021, p. 17). However, I see Tal's suggestion as more extreme since his suggestion is directed at the definition of the prediction target while I see Karaca's point as addressing the mitigation of inductive risk once a prediction target is chosen.

that it forms an alternative to dominant hybrid models of norm specification that include a core deductive element and a subsequent intuitive balancing element. The deductive element implies quite straightforwardly that a solution to a given ethical problem can be deduced from universal or general moral norms and the intuitive balancing element is supposed to provide the necessary flexibility to balance ethical conflicts between different general moral norms. Richardson (1990) points out that these approaches are weak because the complexity of ethical conflicts escapes any deductive general prescription, and the intuitive balancing approach devolves into arbitrariness without rational foundation; and the combination of both elements into a hybrid model does not solve either of these problems. Richardson's (1990) alternative rests on the assertion that the necessary action to be undertaken during a given ethical conflict will become sufficiently clear simply through the act of continued specification (Richardson, 1990, p. 294). It enables therefore certain actions to resolve a conflict without universalizing the norms arrived at through specification and avoids unfounded balancing approaches.

A presentation of how exactly such a specification task would look like in the context of COMPAS goes beyond the scope of this paper. However, in cursory terms, some of the norms in need of specification are the assertion (which I have been taking for granted throughout this thesis) that disparate predictive outcomes which disadvantage minority groups are unacceptable and the notion of equality (f.ex. of opportunity and of outcome) because they are tightly connected to the core fairness issues. Furthermore, since Equivant (formerly Northpointe), developed different variations of their COMPAS model for different states and legislations in the US, it is worthwhile to ponder whether the target specification task should be mandated at the federal or at the state level. Furthermore, comparing different versions of recidivism prediction algorithms through standardised benchmarks may become significantly more complicated because benchmarks would, similarly to the target function, be defined by concrete ideas of fairness that may differ in relevant ways from one legislation to another. A standardised system would therefore require a transparent overview of the values at play in the different benchmarks.

Lastly, on a practical level, it needs to be determined which level of metrological accuracy (as opposed to LMCA) recidivism prediction tools devised under Tal's (2023) suggestion can achieve. Right now, under the LMCA, COMPAS reaches *area under the curve* (AUC)[28] accuracy levels that hover around the seventy percent mark (Eno Louden & Skeem, 2007, p. 13). Since the inherent flaw of the LMCA is that it *overestimates* the accuracy of the tool (Tal, 2023, p. 315) because it wrongly

[28] AUC refers to the area under a *receiver operating characteristic (ROC) curve* which plots the true positive rate (tpr) against the false positive rate (fpr). Each predictor will have a characteristic curve and if one curve dominates (i.e., lies above) another curve, its associated predictor is said to be more reliable than the other. One way to determine whether one curve dominates another is to calculate the area under the curve, hence AUC has become a widely used heuristic to compare predictor performance (Powers, 2012).

operationalizes the target function, the metrological accuracy of the alternative tool will be lower than seventy percent. In addition to the questionable epistemological validity identified in chapter 2, a significantly lower metrological accuracy may, if no progress in performance is achieved otherwise, render the use of recidivism prediction tools entirely obsolete or unjustifiable.

3.4. Conclusion

In this chapter, I have focused on the fairness related problems identified in the beginning of my thesis. These problems stated, firstly, that existing fairness frameworks for machine learning are too formulaic and reductionist, and, secondly, that a rethinking of fairness in statistics is required since present debates merely echo past debates from over fifty years ago. As alternative frameworks I discussed counterfactual fairness approaches but highlighted several of their shortcomings. Kusner et al.'s (2017) proposed framework for counterfactual fairness introduced undesirable and unintuitive changes in rank orderings while Wachter et al.'s (2017) approach to counterfactual explanations struggled with counterfactual logic for social phenomena and an arbitrary distance metric to determine close-enough-possible worlds. As a way to rethink more fundamental axioms in fairness and accuracy in machine learning, Tal's (2023) framework was introduced and compared to the issues outlined above.

Tal's (2023) introduction of *target specification bias* for machine learning applications has the potential to cause a paradigmatic shift in machine learning because it fundamentally questions the reliance on the label-matching conception of accuracy. It also tightly connects the job of the machine learning developer with considerations of societal norms and makes the task of target specification an indispensable value-laden practice. As such, Tal's (2023) approach intensifies recent approaches that seek to promote a more thorough discussion of fairness in machine learning and reconnect it to philosophical concepts (Binns, 2018; Lee et al., 2021). A major bulk of work under this approach will involve consulting stakeholders in order to specify norms relevant to the particular application in question and Richardson's (1990) framework for the specification of norms was briefly introduced to concretize this practice. Whether and how machine learning developers can adhere to standards specified this way and operationalize them in prediction tools is a practical question that needs to be discussed in the technical literature. Especially the level of metrological accuracy obtained through Tal's (2023) approach may reveal recidivism prediction algorithms to be obsolete if they cannot meet a certain benchmark. Concludingly, while Tal's (2023) critique of target specification bias appears to strike at a core issue in machine learning that is responsible for what experts in the field identify as unavoidable fairness trade-offs, the practical implications need to be closer investigated. Furthermore, from a social perspective, the fact that the target function will, under Tal's (2023) approach, in large parts be defined by certain concrete concepts of fairness it may become difficult to inter-compare different machine learning models because the benchmarks used by different developers (derived

from the values of the stakeholders they consulted) may reflect altogether different values in a different legislation with different stakeholders.

The counterfactual approaches to fairness in machine learning base their superiority over other approaches on the adoption of causal models to model and correct for unfair conditions. One finds a reliance of ethical approaches on epistemological standards. This theme culminated in the discussion of Tal's (2023) approach on the relation between accuracy and fairness which explicitly rejects the position that these two notions form orthogonal dimensions and trade-off one another. The idea that accuracy and fairness trade off against one another is a direct consequence of the label-matching conception of accuracy which is a poignant example of how a particular epistemological commitment shapes one's view on fairness. Tal's (2023) suggestion that the task of counterfactual target specification should become an indispensable value-laden practice in the development of machine learning applications, especially those that exhibit fairness problems, is a hopeful message that these applications will, in the future, be developed transparently and that their values align with stakeholders' expectations.

Conclusion

In this thesis, I set out to give a broad picture about the epistemological basis of the recidivism prediction algorithm COMPAS and the fairness issues that plague it. In the first chapter I discussed the problematic basis for the development of this tool and in the second chapter I investigated COMPAS' framing both as a measurement and a predictive tool before arriving, in the third chapter, at a potential reframing of the very concepts of accuracy and its relation to fairness in machine learning. I divided my line of questioning into two parts. Firstly, I asked how these fairness issues arise and what we can do about them, and, secondly, how we can justify the use of recidivism prediction algorithms from an epistemological perspective. While the fairness issues were the primary point of interest, investigating the epistemological basis of predictive algorithms allowed me to determine to which degree they rest on a well-established foundation and whether their use could be justified or discredited before one even has to address the fairness related problems. Following this line of questioning, I identified two epistemological and two fairness related problems and each pair was subsequently addressed in an own chapter.

The epistemological problems consisted in the lack of well-established theory in the field of criminology and the weak support for the predictive utility of COMPAS and for the validity of its measurement scales. The lack of theory implies difficulties in supporting the very definition of recidivism as well as constructing an operationalizable measurand for measuring recidivism. The weak support for the utility and validity of COMPAS implied that the narrative of COMPAS being the state-of-the-art of recidivism prediction methods may overstate its usefulness. For these reasons, I analysed COMPAS in the second chapter both as a predictive and as a measuring tool, drawing from a range of recent publications from philosophy of measurement. I concluded that that the theoretical basis of recidivism prediction was insufficiently well-founded in major part because the primary model it was founded on, the risk-needs-receptivity (RNR) model, was developed using a *dustbowl empiricism* approach which is framed as an atheoretical, empirical approach. The developers of the RNR model specifically and intentionally avoided a theory-laden approach which made it a perfect match for machine learning models *for the wrong reason*. Philosophers of science have recently criticised the supposedly "theory-agnostic" conceptualisation of machine learning modelling, on the grounds that theory-agnosticism both does not exist (because value-laden assumption enter the modelling process either explicitly or implicitly) and is undesirable (implicit value-judgements in modelling can have potentially dangerous effects). I concluded therefore that the COMPAS algorithm was part of a trend gaining track in some scientific disciplines which hailed machine learning algorithms as superior techniques to modelling than existing, theory-laden approaches. Such approaches allude to a sense of objectivity by removing the human factor from the modelling process and avoiding potentially flawed human assumptions and biases. I argued that such an approach necessarily and implicitly imports value-laden assumptions that can have dangerous consequences if

not made transparent and fulfils neither epistemological standards of explainability nor legal standards of due process. I concluded that, on an epistemological and ethical basis, the use of COMPAS was hard to defend.

Turning to the two fairness related problems, I identified the first as the fact that problems regarding fairness trade-offs in machine learning are mainly addressed in the computer science literature and that the frameworks offered there are often too formulaic and reductionist to resolve the fairness issues. The second problem stated that many of the discussions around fairness trade-offs echo near identical debates about fairness that occurred fifty years prior in the context of fairness in standardized testing. The implication that statistical fairness has made no progress within the last five decades casts into doubt the entire project of resolving these issues. An overall rethinking of these fairness issues was therefore required. For this reason, I investigated recent publications regarding counterfactual fairness (Kusner et al., 2017) and explanations (Wachter et al., 2017). These approaches stand out from previous frameworks because they build on top of a causal model of the target phenomenon which, at least theoretically, allows for specifically modelling confounding links between variables and historic biases in order to mitigate them. However, subsequent publications demonstrated that the counterfactual fairness model (Kusner et al., 2017) demonstrated inconsistencies with respect to rank-orderings and was equivalent to the more simplistic model of demographic parity (Rosenblatt & Witter, 2023). Furthermore, the counterfactual explanation approach (Wachter et al., 2017) proved problematic because significant factors like race do not follow counterfactual logic since they do not act in the same way as treatment variables in, for example, clinical trials and cannot be trivially controlled for by keeping other variables constant. The latter is because race is confounded with many social, cultural, and economic factors that influence an individual's situation such that considering it as a treatment variable would amount to reducing the concept of race merely to superficial factors like skin colour. Counterfactual explanations furthermore rely on determining *close-enough-possible-worlds* to compare two predictions that are as similar as possible and in which only the variables that change the outcome differ. The comparison of the differences then serves as an explanation for the prediction of the algorithm. However, defining a metric for closeness between counterfactual worlds is hard to justify because different changes would presumably change the distance in worlds in different ways. Furthermore, some slight changes in variables could potentially change the world in a drastic way. Phenomena of interest therefore typically do not strictly follow counterfactual logic.

I contrast these counterfactual approaches with Tal's (2023) framework regarding counterfactual predictions which approaches fairness problems in a different way by drawing from philosophy of measurement. Specifically, Tal (2023) contrasts the commonly accepted benchmark of accuracy in machine learning – the label-matching conception of accuracy – with the metrological conception of accuracy and identifies an inferential gap between the measurement target that stakeholders actually desire and the target that labels in a training dataset operationalize. To mitigate this *target*

specification bias, Tal (2023) suggests adopting aspects of the way metrologists define measurands, namely by formulating an idealized, generally not directly accessible, target and finding ways to operationalize targets with respect to this ideal. Tal (2023) therefore calls for abandoning the automatic reliance on the label-matching conception of accuracy and instead proposes to design a suitable target function, in tandem with stakeholders, which specifies the counterfactual conditions of the predictions the stakeholders are interested in. Viewed from this perspective, the commonly held idea that accuracy and fairness trade off against one another in prediction algorithms is ill-conceived because it misunderstands that stakeholders actually expect a fair prediction tool such that fairness becomes an integral component of the definition of the ideal target. This approach explicitly rejects the idea that labels in a training dataset – however reliably the data was collected – automatically reflect the best operationalization of the target function and instead recentres rendering expectations and values of stakeholders explicit by making the task of target specification an indispensable, value-laden practice of developing machine learning based predictive algorithms. Viewing the labels in datasets as suboptimal operationalizations, one could conceivably manipulate datasets and training methods such that they operationalize a counterfactual vision of the target function.

How counterfactual targets can be operationalized in practice will need to be investigated by the technical literature. In particular, two major challenges I identified for future work will be, firstly, that the construction of standardized benchmarks for machine learning applications designed for counterfactual conditions will require transparent specifications of values in order to make different applications comparable at all. This is because the counterfactual conditions stakeholders are interested in may vary from legislation to legislation such that different applications reflect different sets of values and visions of fairness. The key issue is to agree on a way of comparing non-epistemological standards like definitions of fairness when even epistemological standards of performance evaluation appear to produce failures across many subfields of machine learning (Liao et al., 2021). A promising way of approaching such benchmarks may by following the work of (LaCroix & Luccioni, 2022) who consider the design of 'ethical' AI models impossible but suggest shifting the evaluation standard towards the degree of value alignment for different stakeholders.

The second challenge will be that, by adopting the metrological conception of accuracy, some machine learning applications will reveal themselves to be unable to meet the benchmark for reliability. Recidivism prediction algorithms like COMPAS produce, under the current label-matching conception of accuracy, accuracy levels that hover around seventy percent. Since applications designed under the LMCA *overestimate* the accuracy of the application since they fail to consider the counterfactual condition stakeholders are actually interested in, their metrological accuracy will likely be (much) lower. Without performance increases, recidivism prediction algorithms like COMPAS could likely turn out to be too unreliable to use. A way of safeguarding against such a scenario would be by better models of recidivist behaviour and a more thorough understanding of the causal links

between different risk factors. At the same time, directly addressing the socio-economic circumstances that cause base rate disparities in recidivism between different races will also reduce the disparities in the training datasets for recidivism prediction algorithms, thereby also reducing their biases and fairness issues. This last suggestion should be the overall goal of policy makers: the main focus should lie on working on socio-economic issues underlying criminal and recidivist behaviour while prediction tools only play – at best – a complementary role, provided their reliability and trustworthiness was sufficiently demonstrated.

Wordcount: 23720

Bibliography

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine June 23, 2008. Retrieved from: https://www.wired.com/2008/06/pb-theory/

Andrews, D. A., & Bonta, J. (1998). The Psychology of Criminal Conduct (Second Editions). *Cincinnati: Anderson Publishing Co*.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, *52*(1), 7–27. https://doi.org/10.1177/0011128705281756

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 62–76. https://proceedings.mlr.press/v81/barabas18a.html

Barkow, J. H., Cosmides, L., & Tooby, J. (1995). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation*, *6*(1), 1–22.

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation*, *6*(1), 1–22.

Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. In M. Bertolaso & F. Sterpetti (Eds.), *A Critical Reflection on Automated Science: Will Science Remain Human?* (pp. 43–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-25001-0_4

Boumans, M. J. (2007). *Invariance and Calibration* (SSRN Scholarly Paper No. 1434797). https://papers.ssrn.com/abstract=1434797

Brennan, T., Dieterich, B., Breitenbach, M., & Mattson, B. (2009). *A Response to "Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions*

(COMPAS) (Brennan, T., Fretz, R., & Wells, D.(2003). COMPAS Users Case Management Guide). Traverse City, MI: Northpointe Institute for Public Management.

Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, *22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

Campolo, A., & Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*. https://doi.org/10.17351/ests2020.277

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. *AI & SOCIETY*, *35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Chatterjee, A. (2017). Causality: Physics and Philosophy. *European Journal of Physics Education*, *4*(1), 1–5.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cleary, T. A. (1966). Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges. *ETS Research Bulletin Series*, *1966*(2), i–23. https://doi.org/10.1002/j.2333-8504.1966.tb00529.x

Cleary, T. A. (1968). Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement*, *5*(2), 115–124.

Deci, E. L., Ryan, R. M., Gagné, M., Leone, D. R., Usunov, J., & Kornazheva, B. P. (2001). Need Satisfaction, Motivation, and Well-Being in the Work Organizations of a Former Eastern Bloc Country: A Cross-Cultural Study of Self-Determination. *Personality and Social Psychology Bulletin*, *27*(8), 930–942. https://doi.org/10.1177/0146167201278002

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc. Research Department.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science*, *81*(1), 1–21. https://doi.org/10.1086/674345

Eno Louden, J., & Skeem, J. L. (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.].

Feyerabend, P. (1975). *Against method*. New York: New Left Books

Gramlich, J. (2021, August 16). *America's incarceration rate falls to lowest level since 1995. Pew Research Center*. *https://www.pewresearch.org/short-reads/2021/08/16/americas-incarceration-rate-lowest-since-1995/*

Hedström, P., & Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, *36*(1), 49–67. https://doi.org/10.1146/annurev.soc.012809.102632

Heidegger, M., ([1993] 2008). The Question Concerning Technology. In Scharff, R. C., & Dusek, V. (Eds). *Philosophy of Technology: The Technological Condition: An Anthology (pp. 305-317)*. John Wiley & Sons.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

Illari, P. M., Russo, F., & Williamson, J. (2011). *Causality in the Sciences*. Oxford University Press.

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, *5*(2), 217–232. https://doi.org/10.1007/s13194-014-0105-6

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Karaca, K. (2021). Values and inductive risk in machine learning modelling: The case of binary classification models. *European Journal for Philosophy of Science*, *11*(4), 102. https://doi.org/10.1007/s13194-021-00405-1

Kasirzadeh, A., & Smart, A. (2021). The Use and Misuse of Counterfactuals in Ethical Machine Learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 228–236. https://doi.org/10.1145/3442188.3445886

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (arXiv:1609.05807). arXiv. https://doi.org/10.48550/arXiv.1609.05807

Knuuttila, T., & Loettgers, A. (2014). Varieties of noise: Analogical reasoning in synthetic biology. *Studies in History and Philosophy of Science Part A*, *48*, 76–88. https://doi.org/10.1016/j.shpsa.2014.05.006

Kohler-Hausmann, I. (2019). Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. *Northwestern University Law Review*, *113*(5), 1163–1228.

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, *33*(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

LaCroix, T., & Luccioni, A. S. (2022). *Metaethical Perspectives on 'Benchmarking' AI Ethics* (arXiv:2204.05151). arXiv. http://arxiv.org/abs/2204.05151

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics*, *1*(4), 529–544. https://doi.org/10.1007/s43681-021-00067-y

Liao, T., Taori, R., Raji, I. D., & Schmidt, L. (2021, August 29). *Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning*. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://openreview.net/forum?id=mPducS1MsEK

Linnemann, N. S., & Visser, M. R. (2018). Hints towards the emergent nature of gravity. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *64*, 1–13. https://doi.org/10.1016/j.shpsb.2018.04.001

MacIntyre, A., & Korbut, A. (2013). *A mistake about causality in social science*. https://philpapers.org/rec/MACAMA-4

Marini, M. M., & Singer, B. (1988). Causality in the Social Sciences. *Sociological Methodology*, *18*, 347–409. https://doi.org/10.2307/271053

Mattu, J. A., Jeff Larson, Lauren Kirchner, Surya. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Mitchell, S. D. (2020). Instrumental Perspectivism: Is AI Machine Learning Technology Like NMR Spectroscopy? In M. Bertolaso & F. Sterpetti (Eds.), *A Critical Reflection on Automated Science: Will*

Science Remain Human? (pp. 27–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-25001-0_3

More than One in 100 Adults Are Behind Bars, Pew Study Finds. (2008, February 28). http://pew.org/1Q6ZxUK

Mussgnug, A. M. (2022). The predictive reframing of machine learning applications: Good predictions and bad measurements. *European Journal for Philosophy of Science*, *12*(3), 55. https://doi.org/10.1007/s13194-022-00484-8

Netherlands | World Prison Brief. (n.d.). Retrieved 1 June 2023, from https://prisonstudies.org/country/netherlands

Newburn, T. (2018). *Criminology: A Very Short Introduction*. Oxford University Press.

Parker, W. S. (2017). Computer Simulation, Measurement, and Data Assimilation. *The British Journal for the Philosophy of Science*, *68*(1), 273–304. https://doi.org/10.1093/bjps/axv037

Pearl, J. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*(2), 3.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Powers, D. M. W. (2012). The problem of Area Under the Curve. *2012 IEEE International Conference on Information Science and Technology*, 567–573. https://doi.org/10.1109/ICIST.2012.6221710

Richardson, H. S. (1990). Specifying Norms as a Way to Resolve Concrete Ethical Problems. *Philosophy & Public Affairs*, *19*(4), 279–310.

Rosenblatt, L., & Witter, R. T. (2023). Counterfactual Fairness Is Basically Demographic Parity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), Article 12. https://doi.org/10.1609/aaai.v37i12.26691

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.6ed64b30

Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press.

Sawyer, R. L., Cole, N. S., & Cole, J. W. L. (1976). Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection. *Journal of Educational Measurement*, *13*(1), 59–76.

Srećković, S., Berber, A., & Filipović, N. (2022). The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines*, *32*(1), 159–183. https://doi.org/10.1007/s11023-021-09575-6

Suddaby, R. (2014). Editor's Comments: Why Theory? *Academy of Management Review*, *39*(4), 407–411. https://doi.org/10.5465/amr.2014.0252

Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, *73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Tal, E. (2012). The epistemology of measurement: A model-based account. University of Toronto (Canada).

Tal, E. (2017). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A*, *65–66*, 33–45. https://doi.org/10.1016/j.shpsa.2017.09.001

Tal, E. (2020). Measurement in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Tal, E. (2023). Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 312–321. https://doi.org/10.1145/3600211.3604678

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Walters, G. D. (2017). Beyond Dustbowl Empiricism: The Need for Theory in Recidivism Prediction Research and Its Potential Realization in Causal Mediation Analysis. *Criminal Justice and Behavior*, *44*(1), 40–58. https://doi.org/10.1177/0093854816677566

Ward, T., & Stewart, C. (2003). Criminogenic needs and human needs: A theoretical model. *Psychology, Crime & Law*, *9*(2), 125–143. https://doi.org/10.1080/1068316031000116247

Washington, A. L. (2018). How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *Colorado Technology Law Journal*, *17*, 131.

Weisberg, R. (2013). Meanings and Measures of Recidivism. Southern California Law Review, 87, 785.

Wellford, C. (1975). Labelling Theory and Criminology: An Assessment*. *Social Problems*, *22*(3), 332–345. https://doi.org/10.2307/799814

Zecha, G. (1992). Value-neutrality and criticism. *Journal for General Philosophy of Science*, *23*(1), 153–164. https://doi.org/10.1007/BF01801800

Appendix A – Risk/Needs factors

TABLE 1 Major Risk and/or Need Factors and Promising Intermediate Targets for Reduced Recidivism

Factor	Risk	Dynamic Need
History of antisocial behaviour	Early and continuing involvement in a number and variety of antisocial acts in a variety of settings	Buil noncriminal alternative behaviour in risky situations
Antisocial personality pattern	Adventurous pleasure seeking, weak self-control, restlessly aggressive	Build problem-solving skills, self-management skills, anger management and coping skills
Antisocial cognition	Attitudes, values, beliefs, and rationalizations supportive of crime; cognitive emotional states of anger, resentment and defiance; criminal versus anticriminal identity	Reduce antisocial cognition, recognize risky thinking and feeling, build up alternative less risky thinking and feeling, adopt a reform and/or anticriminal identity
Antisocial associates	Close association with criminal others and relative isolation from anticriminal others; immediate social support for crime	Reduce association with criminal others, enhance association with anticriminal others
Family and/or marital	Two key elements are nurturance and/or caring and monitoring and/or supervision	Reduce conflict, build positive relationships, enhance monitoring and supervision
School and/or work	Low levels of performance and satisfaction in school and/or work	Enhance performance, rewards, and satisfactions
Leisure and/or recreation	Low levels of involvement and satisfaction in anticriminal leisure pursuits	Enhance involvement, rewards, and satisfactions
Substance abuse	Abuse of alcohol and/or other drugs	Reduce substance abuse, reduce the personal and interpersonal supports for substance-oriented behaviour,

		enhance alternatives to drug abuse

NOTE: The minor risk and/or need factors (and less promising intermediate targets for reduced recidivism) include the following: personal and/or emotional distress, major mental disorder, physical health issues, fear of official punishment, physical conditioning, low IQ, social class of origin, seriousness of current offense, other factors unrelated to offending. [Note in original]

Note: Reprinted from "The Recent Past and Near Future of Risk and/or Need Assessment" *by* Andrews, D. A., Bonta, J., & Wormith, J. S., 2006, *Crime & Delinquency*, *52*(1), 7–27.

Appendix B – Proof of mathematical incommensurability of fairness definitions

Given the number N_t of members in group t and the number μ_t of people in this group that belong to the positive class, let us call $x \in [0;1]$ the average score attributed to the members of the negative class and $y \in [0;1]$ the average score attributed to members of the positive class. Remember that we are looking here at the training dataset and therefore know all the true class labels. The balance conditions for the positive and negative class require that the scores x and y are the same for both groups. The number of people assigned to the negative class should be $(N_t - \mu_t)$ and the number of people assigned to the positive class should be μ_t (if the algorithm assigns more or less people to the positive class than there are in the training dataset, it is over- or underfitted). The calibration condition requires that an x fraction of the people assigned to the negative class, $(N_t - \mu_t)x$, belong to the positive class and a y fraction of the people assigned to the positive class, $\mu_t y$, belong to the positive class. To understand this, remember that the score here denotes the estimated likelihood that a member of this class will recidivate. Assigning, for example, an average risk score of 70% to the high-risk (positive) class containing 100 people, we expect 70 of these people to truly recidivate if the algorithm is well calibrated. Analogously, assigning an average score of 30% to the low-risk (negative) class also containing 100 members, we expect 30 of these members to truly recidivate. Since, in this example, we assume that there are only those two classes (positive and negative, high-risk and low-risk), the two terms add up to the total number of members of the positive class, μ_t, which results in the following equation for each group t:

$$(N_t - \mu_t)x + \mu_t y = \mu_t$$

Dividing by N_t, we obtain:

$$(1 - \rho_t)x + \rho_t y = \rho_t$$

Where $\rho_t = \mu_t / N_t$ designates the relative base rate of group t. We have therefore two linear equations in x-y-space that align perfectly only if the base rates are equal: $\rho_1 = \rho_2$, or which, if the base rates are not equal, only intersect at the point $(x, y) = (0,1)$. The latter condition amounts to perfect prediction: the average score to assigned to members of the negative class is 0 and the average score for members of the positive class is 1, meaning that no members assigned to the negative class recidivate and all members of the positive class do. This proof demonstrates therefore that the calibration condition together with the balance condition for the positive and negative class can only hold all at the same time if the base rates of the different groups are equal (which in the case of recidivism is not the case) or if we can make perfect predictions (which is practically infeasible). The immediate conclusion is that we cannot simply have totally fair prediction algorithms but rather have to contend with inherent trade-offs and discrepancies.