An Analysis of Kantian Ethics on the Usage of Artificial Intelligence in Legal Sentencing

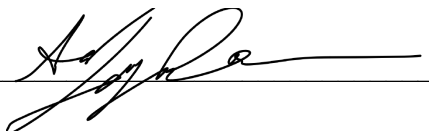
STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Sridhiraj Jayakumar

May 1, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: 

Approved: _____ Date _____
Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Statement of Intent

I am studying legal cases, specifically the actions the judge took in *Wisconsin v. Loomis*, in which the COMPAS algorithm, or the Correctional Offender Management Profiling for Alternative Sanctions algorithm, was used because I want to study the moral implications of using software to determine a sentence ruling in order to help relevant judging parties understand the moral obligation needed to interact with AI and contribute to the discussion of the interaction between humans and AI

Introduction

In 2013, Eric Loomis had been charged with six-years in prison for driving a stolen vehicle and fleeing from the police in the legal case *Wisconsin v. Loomis*. The proceedings for this case are normal, however, the usage of a risk analysis software, COMPAS, complicates the matter. This software aided the judge to make a decision on the defendant's sentencing, ultimately influencing his decision.

Many scholars have analyzed the limitations of COMPAS and discussed the software's vulnerability to bias. However, most of these analyses focus primarily on the legal aspect of using risk analysis software. This limited perspective fails to take into account the moral obligations the judge has to review the assessment fairly and provide a moral decision.

To further the understanding of using COMPAS, and other AI, in critical situations, I will use the duty ethics framework as a means to judge the actions performed by the relevant parties of the court system. Specifically, I will highlight that their actions were immoral as the judging parties did not place the same scrutiny on COMPAS as compared to other evidence and that the decision to allow the usage of artificial intelligence in sentence ruling sets a dangerous precedent.

Background

COMPAS, or Correctional Offender Management Profiling for Alternative Sanctions, is a machine learning software, developed by Northpointe Inc, used in the US court systems of New York, Wisconsin, California, Florida, and other jurisdictions. Its main job is to identify recidivism risk, or the potential to repeat a crime, to assist the court in sentencing. COMPAS does this by taking into account multiple factors of the defendant such as history of violence,

substance abuse, social environmental problems, and more including a 137-question survey that defendants must fill out.

In each of these categories, defendants get a score from 1-10 with 1 being the lowest risk and 10 being the highest. The main two scores that are significant in post-conviction sentencing is the general recidivism risk and violent recidivism risk (*Figure 2*). Each of these categories, have specific types (*Figure 1*) associated with them that help determine a defendant's risk. Usually, however, a judge will only get a limited view of the score, given only the general/violent recidivism risk score.

Table 2.1: Cutting Points for COMPAS Scale Types.

Type 1	Low (1-4)	Medium (5-7)	High (8-10)
Type 2	Unlikely (1-2)	Probable (3-4)	Highly Probable (5-10)
Type 3	Unlikely (1-5)	Probable (6-7)	Highly Probable (8-10)
Type 4	Unlikely (1-4)	Probable (5-7)	Highly Probable (8-10)

Figure 1: Risk Likelihood Based on Factor-Type. Adapted from 'Practitioner's Guide to COMPAS,' 2012, *Northpointe Inc.*, p. 11. Copyright 2011 by Northpointe Inc.

Table 2.2: COMPAS Core Scales and Types.

Scale	Scale Type
Violent Recidivism Risk	1
General Recidivism Risk	1
Pretrial Release Risk	1
Criminal Involvement	1
History of Noncompliance	1
History of Violence	1
Current Violence	1
Criminal Associates/Peers	4
Substance Abuse	2
Financial Problems/Poverty	3
Vocational/Education Problems	3
Criminal Thinking	3
Family Criminality	3
Social Environment Problems	3
Leisure and Recreation	3
Residential Instability	3
Social Adjustment Problems	3
Socialization Failure	3
Criminal Opportunity	3
Criminal Personality	3
Social Isolation	3

Figure 2: Risk Factors and their Types. Adapted from 'Practitioner's Guide to COMPAS,' 2012, *Northpointe Inc.*, p. 11. Copyright 2011 by Northpointe Inc.

Literature Review

While there have been some scholars that analyzed COMPAS in *Wisconsin v. Loomis*, the majority of these analyses focus upon the legal aspect of using COMPAS to assist in sentencing. Specifically, these scholars consider if it conveys accurate information and whether

or not its usage violates due process, or a fair treatment through the judicial system. What these analyses fail to consider are the moral implications of using this sort of AI in court proceedings.

In *Machine Bias*, researchers Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner discuss the contrasting results COMPAS has on determining recidivism risk, based on race. Supposedly, COMPAS ranks recidivism by considering multiple factors of a criminal's life. It takes into account a criminal's personality, social isolation, substance abuse, and residence/stability. However, these researchers found that black criminals were twice as likely to receive high recidivism ranks even when white criminals performed more significant crimes (Angwin & Larson 2016). These scholars go on to analyze that the existence of bias in the software may be the result of a well-known concept in computer science called GIGO, Garbage-In Garbage-Out, which explains that given flawed input data the model produces flawed output data (Demming 2019). While these researchers do outline the dangers of using COMPAS due to implicit bias, they do not adequately consider the moral decisions humans made using COMPAS.

Similarly, in *Does the use of risk assessments in sentences respect the right to due process?*, Iñigo De Miguel Beriain performs a critical analysis on *Wisconsin v. Loomis* and its appeal to the Supreme Court of Wisconsin. First, Beriain highlights that the Supreme Court ruled the usage of algorithms in sentencing was acceptable. They also ruled that the defendants' right to due process was not violated by the mere fact that they could not access an explanation of the algorithm. Then, Beriain scrutinizes the reasoning used by the Supreme Court and argues the use of algorithms in sentencing is a violation of due process and that this case sets a significant precedent for future cases. He does this by arguing against a few points made by the Supreme Court. One of these points was that the software "correctly predicted an offender's recidivism

61% of the time, but was only correct in its predictions of violent recidivism 20% of the time" (Beriain 2018). He argued that the business logic of the software was misaligned with the court's true intentions, stating that if COMPAS recommended custody for a defendant that does not need it, it would hardly be detected. However, if COMPAS did not recommend custody and the defendant went on to perform a violent crime, the company and developers would face strong criticism. There is benefit from analyzing why the shortcomings of COMPAS make it unfit to use in a court setting, but examining the actions taken by humans due to COMPAS will further our moral understanding in this case.

Both these scholars identified important areas of scrutiny dealing with COMPAS, however, I want to shift the focus from a legal perspective to a moral perspective by focusing on the specific actions the judge in this case took because of COMPAS. By analyzing this other side of the case, I want to further understanding in how humans should act in conjunction with AI and highlight the moral obligations we have, to use it in the right way.

Conceptual Framework

I will use deontology or duty-ethics to analyze the actions the judge took when using COMPAS to determine sentencing. This will allow me to open a discussion about the actions and precautions humans must take when dealing with technologies such as COMPAS.

Deontology is the normative ethical theory in which the morality of an action is based upon whether or not the action itself is right or wrong, rather than weighing in the consequences of an action. Under deontology, I will be using Immanuel Kant's moral system, which has two main concepts: moral autonomy and the Categorical Imperative. First, moral autonomy is the notion that any human should be able to rationally determine what is or is not moral of their own accord, rather than heeding the injunctions of others (Cureton & Johnson 2004). This means that

people should be given the relevant information to make their own decisions and that people should practice some scrutiny when hearing the decisions of others (Dryden, n.d.). To find this moral reasoning, one can use the Categorical Imperative, which describes a universal principle from which all moral norms can be derived from. Additionally, "the first formulation of the categorical imperative, the universality principle, is as follows: 'Act only on that maxim which you can at the same time will that it should become a universal law' (van de Poel 2011). This principle simply states that an action is only moral if one can envision a world where any person can perform the same action and it is still a just world. For example, looking at the action of lying under this Categorical Imperative would be as follows. The university principle states that if lying is moral, a world where everyone can lie should be a just world. However, this is clearly not the case. In this scenario, there would be not trust among others and society would not function with the same level of productivity.

Under Kant's theory, I can analyze the specific actions the judge took when working with COMPAS, and establish whether or not it was moral, which leads into a bigger discussion of the interaction between humans and AI in critical situations.

The main claims I will be identifying, later in the paper, are that COMPAS was not placed under the same scrutiny as other evidence would have been when imposing sentencing and that this action sets precedence for allowing similar software to be used in court systems.

Analysis

The actions the judge made, regarding sentencing, in *Wisconsin v. Loomis* were absent of the qualities needed to be morally acceptable under Kant's Theory. The judge decided on the sentence without having the precise knowledge of how the COMPAS score was determined and he set a dangerous precedent of allowing the use of artificial intelligence in the court systems. By

analyzing these issues through the moral lens of duty-ethics, I will highlight how they violate the two main ideas revolving Kant's Theory: moral autonomy and the university principle under the Categorical Imperative.

Moral Autonomy

In order to exhibit how the judge violated his moral autonomy in *Wisconsin v. Loomis*, it must be proven that the judge performed the sentencing without being given the accurate knowledge regarding how the recidivism risk score was calculated. This was the same concern Loomis had when learning how his six-year sentence was determined. So, Loomis filed a motion for post-conviction relief, arguing that the usage of COMPAS violated his due process rights ("Criminal Law", 2017). He had a couple of arguments to file this motion, but the significant complaint to focus on is that he argued because the methodology behind COMPAS reports are proprietary, he, as a defendant, is not provided with enough information to refute claims assessed by COMPAS. In response to this motion, the Supreme Court of Wisconsin denied Loomis' post-conviction relief however, Supreme Court Justice Bradley followed this ruling by enacting a rule that incorporated these five rules as warnings to PSIRs (Presentence Investigation Report) that include COMPAS reports:

> first, the "proprietary nature of COMPAS" prevents the disclosure of how risk scores are calculated; second, COMPAS scores are unable to identify specific high-risk individuals because these scores rely on group data; third, although COMPAS relies on a national data sample, there has been "no cross-validation study for a Wisconsin population"; fourth, studies "have raised questions about whether [COMPAS scores] disproportionately classify minority offenders as having a higher risk of recidivism"; and

> fifth, COMPAS was developed specifically to assist the Department of Corrections in making post-sentencing determinations. ("State v. Loomis," 2017)

These rules act as an indicator to show that Justice Bradley recognized the risks associated with using software similar to COMPAS, and emphasizes that judges must exercise caution when using them to determine sentencing. Additionally, it shows that the court systems have low confidence with this software and that sentencing should not be based solely on COMPAS (Forward, 2019). The most important rule that relates back to Kantian ethics is the first rule that indicates that judges and defendants will not have access to the methodology in which the score was calculated and because of the proprietary nature of COMPAS the relevant information on how the score was calculated cannot be disclosed. These rules are reasonable because after the COMPAS report showed that Loomis had a high risk of violence, recidivism, and pretrial risk, "[t]he judge agreed, telling Mr. Loomis that 'you're identified, through the Compas assessment, as an individual who is a high risk to the community'" (Liptak, 2017). This reliance upon the assessment suggests that the COMPAS score had an influence in the sentencing of Loomis, despite the fact that the judge was not presented with all the information regarding COMPAS's assessment. Now, considering moral autonomy can only be performed if the relevant parties have the necessary information to appropriately make decisions, the judge that ruled the six-year sentence against Loomis did not follow this concept of Kantian Ethics.

Categorical Imperative

Similarly, the action of continuing to allow the use of artificial intelligence to help support judges in making sentencing decisions in a court of law is immoral under the university principle because his decisions, when scaled, bring about an unjust world. This principle defines

if an action is moral/amoral after examining the effect on the world if everyone was allowed to freely perform the same action. By expanding this decision, court systems around the world are able to use COMPAS-like software in order to help judges make decision on sentencing. Would this be a just or safe world? Looking at Paul Zilly's case, might shed some light into this. Zilly was convicted of stealing a push lawnmower and some tools in Barron County. The prosecutor of this case recommended that Mr. Zilly get a year in county-jail and follow-up supervision. In response, his lawyer agreed to a plea deal. However, Judge James Babler had received a COMPAS risk score and it had given Mr. Zilly a high risk for violent recidivism. As a result, Judge Babler overturned the plea deal and sentenced Mr. Zilly to two years in county-jail followed by three years of supervision (Murphy, 2017). This sole risk assessment score strongly influenced Judge Babler and caused him to increase his original sentence, costing Mr. Zilly an extra year in prison. Recognizing this issue, Mr. Zilly appealed this decision and asked the score's creator, Brennan, to take the stand to testify about the usage of risk assessment score. Brennan testified that, originally, he did not design this software to be used in sentencing decisions. He specifically said, "I wanted to stay away from the courts[.] But as time went on […] I gradually softened on whether this could be used in courts" (Angwin & Larson 2016). After hearing Brennan's testimony, and realizing that the score was meant to be more of a guideline, the judge lowered the sentence to one year and six months. By allowing the use of this software to be used in all court systems, there may be many judges who take recidivism risk scores generated by artificial intelligence as gospel and blindly trust them without providing the proper scrutiny to them as they would to other evidence.

In a similar case of trusting risk assessment scores, a 19-year-old, Christopher Drew Brooks had consensual sex with a 14-year-old girl and was convicted with statutory rape in

Virginia. The sentencing guidelines suggested that Mr. Brooks be sentenced with a jail term of 7 – 16 months. However, after the recommendation of a risk assessment report (built by another company not affiliated with COMPAS), Mr. Brooks maximum sentence was increased to 24 months. Taking this into account, the judge sentenced him to 18 months in jail. However, there was a problem in how the software calculated the recidivism risk. The method in which the software calculated recidivism score took into account the age of the defendant, and since Brooks was convicted of a sex offence at a young age, even though he was closer in age to the victim, this particular algorithm weighed this as more negative. This means that "[i]n fact, had Brooks been 36 years old (and hence 22 years older than the girl) the algorithm would have recommended that he not be sent to prison at all" (Fry, 2018). This case raises a significant issue in trusting that the world should incorporate these algorithms to assist judges during rulings. It is hard to determine the accuracy of these complex systems and they may be misaligned in identifying what traits cause a higher risk in recidivism than others.

By allowing artificial intelligence-based risk analysis software to be used in every court proceeding, the legal system becomes liable to more cases like Mr. Zilly's and Mr. Brooks' where the judgements are based on incomplete information and the research and analysis that are supposed to be done by the judges gets offloaded to technology.

I have argued that the use of artificial intelligence in these critical settings should be subject to scrutiny and that when expanded to the whole world, the resulting world is unjust. However, some proponents to the usage of AI argue that this is the reality of the current world, where many algorithms and artificial intelligence dictate the next moves of an entity, whether it be related to investing in stocks or identifying health problems (Steiner 2013). However, this viewpoint fails to consider that this usage goes hand-in-hand with human involvement and, in

most cases, the information on how the algorithm makes decisions is available to the relevant parties. For example, there is an algorithm that diagnoses Pneumocystis Jirovecii Pneumonia (PJP) for seriously ill HIV-infected patients. PJP is a common cause of hospitalization for HIV-infected patients and, therefore, this algorithm is very critical in ensuring a patients health and safety. In order to diagnose these patients, this algorithm takes into account four predictive variables dyspnoea, chest X-ray, haemoglobin and oxygen saturation (Gary & Annemie, 2018). By compiling these four main areas and performing a type of regression analysis on the data, the algorithm can then identify whether or not a patient has PJP. However, the main advantage with this algorithm is that relevant professionals can read this data and come up with similar conclusions. This is what separates the use of AI in court systems versus other areas of society (Bench-Capon, 2020). The judges do not have the relevant information needed to assess how to scores were calculated and, therefore, cannot be expected to use these scores without giving up some of their autonomy.

Conclusion

By looking at *Wisconsin v. Loomis*, some knowledge can be gained in regards to the usage of complex artificial intelligence systems that aid a judge in determining a sentence. The actions taken by the judge in *Wisconsin v. Loomis* reveal a failure to uphold the judge's moral autonomy and an oversight in ruling that does not take into account the significant implications of setting precedent in using AI in court systems. Using the duty ethics framework, the actions taken by the judge are deemed to be immoral as he does not take into account the moral implications of his decisions, under Kantian ethics.

The usage of artificial intelligence in critical systems is an ever-increasing concern in the modern world. By analyzing these events under a moral framework, it can accentuate the

importance of the decisions and highlight that these decisions themselves have to be moral. By carefully considering these actions, it opens the door to understanding how humans should interact with AI and the moral obligations that come with it.

Word Count: 3068

References

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2019). Machine Bias. *Propublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bench-Capon, T. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence*. *281*, 1–18.

Beriain, I. D. M. (2018). Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. *Law, Probability and Risk*. *17*(1), 45–53.

Criminal Law - Sentencing Guidelines - Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. (2017). *Harvard Law Review*. *130*(5), 1530–1537.

Demming, A. (2019). Machine learning collaborations accelerate materials discovery. Retrieved from https://physicsworld.com/a/machine-learning-collaborations-accelerate-materials-discovery/

Dryden, J. (n.d.). Autonomy. *Internet Encyclopedia of Philosophy*. Retrieved from https://www.iep.utm.edu/autonomy/

Forward, J. (2017). The Loomis Case: The Use of Proprietary Algorithms at Sentencing. *Inside Track*. *9*(4), 1-4. Retrieved from https://www.wisbar.org/NewsPublications/InsideTrack/Pages/Article.aspx?Volume=9&Issue=14&ArticleID=25730

Fry, H. (2018, September 21). Can an algorithm deliver justice? *Science Focus*. Retrieved from https://www.sciencefocus.com/future-technology/can-an-algorithm-deliver-justice/

Johnson, R., & Cureton, A. (2016, July 7). Kant's Moral Philosophy. *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/kant-moral/

Liptak, A. (2017, May 1). Sent to Prison by a Software Program's Secret Algorithms. *New York Times*. Retrieved from https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

Maartens, G., Stewart, A., Griesel, R., Kengne, A. P., Dube, F., Nicol, M., … Mendelson, M. (2018). Development of a clinical prediction rule to diagnose Pneumocystis jirovecii pneumonia in the World Health Organization's algorithm for seriously ill HIV-infected patients. *Southern African Journal of HIV Medicine*, 19(1).

Murphy, B. (2017, May 1). Murphy's Law: State Justice System Uses Racially Biased Test. *Urban Milwaukee.* Retrieved from https://urbanmilwaukee.com/2017/01/05/murphys-law-justice-system-uses-biased-test/

Northpointe Inc. (2012). Practitioners Guide to Compas (pp. 7–13).

Shapiro, D. M. (2017, March 10). State v. Loomis. *Harvard Law Review*. Retrieved from https://harvardlawreview.org/2017/03/state-v-loomis/

Steiner, C. (2013). Pop Goes the Algorithm.