UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Mundus Intelligibilis:
Mitigating the Ethical Considerations of Artificial Intelligence through Visual Analytics

by Brett Anthony Hansard

Thesis submitted in fulfilment of the requirements for
the degree of

Master of Philosophy (Research)

under the supervision of A/Prof Jianlong Zhou

University of Technology Sydney
Faculty of Engineering and Information Technology

September 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP I, Brett A. Hansard declare that this thesis, is submitted in fulfilment of the requirements for the award of Master of Analytics (Research), in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology, Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 25 September 2024

Abstract

The *mundus intelligibilis* or 'intelligible world', is an omnipresent and almost inescapable phenomenon. An evolution where human intelligence is being supported, supplemented, or superseded by Artificial Intelligence (AI). Decisions once made by humans are now made by machines, learning at a faster and more accurate rate through algorithmic calculations, and those decisions are then made intelligible through interpretive measures such as visual analytics. Resolutely, this proposition in machine learning is explored in this thesis through predictive modelling on 101 bail decisions. Indicatively, the models' statistical performance and accuracy based on the nine predictor variables proved effective, with the more accurate of the logistic regression models at 78 percent and performance value 0.845 (AUC) and the classifier model accuracy at 72.5 percent and performance value 0.702 (AUC). By virtue of these results, this thesis explores how AI-generated bail decisions would be received by those who would be affected, giving prominence to the ethical principles of fairness and explainability. A user-study was subsequently undertaken within court environs surveying relevant stakeholders through a series of questions, vignettes and visualisations. Resultative exploratory data analysis found perceptions were weighted positively on visual analytics as an ethical mitigator to AI-generated bail decisions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Australia's criminal justice systems, as yet, have not transitioned to the 'intelligible world' – *mundus intelligiblis* – where Artificial Intelligence (AI) has a role in making decisions [15] and any cause to query the moral and ethical implications of AI has been inconsequential [16]. However, the literature reflects that the Australian justice system has recognised the benefits and limitations of AI [17][18][19]. Adversarial criminal justice systems similar to Australia's, such as the United Kingdom (U.K.) and the United States (U.S.), are or have transitioned to a *mundus intelligiblis* where AI has supplemented or superseded conventional decision-making in some capacity [15][20]. The transition and implementation has fostered debate on the ethicality and morality of AI: the pessimist's submit that AI cannot act or reason morally or ethically [21] thus impacting decision-making capabilities; contrastingly, the pragmatist/optimist submit that AI can act ethically and morally [22] and would benefit decision-making [23]. Scholars from Australian law and justice disciplines have given their support for AI-based decisions to be piloted in conjunction with conventional processes in bail, sentencing and parole, but emphasised it should be undertaken cautiously and empirically [15][24].

Machine Learning (ML), a sub-discipline of AI, is materialising to become an effective forecasting mechanism in offender risk assessments and predicting recidivism, ultimately supporting decision-making. For example, a meta-analysis of 11 research papers related to criminal justice and ML, found the models were robust in predicting recidivism [24].

Putting statistic vigour and value aside, AI supported decision-making across Australian criminal justice systems ought to be established with the *Ethical Considerations* (herein abbreviated to "ECs") of *fairness* and *explainability*, which are the foundational principles of this thesis (see Figure 1.1), as without these, foreseeable injustices are surely to manifest.

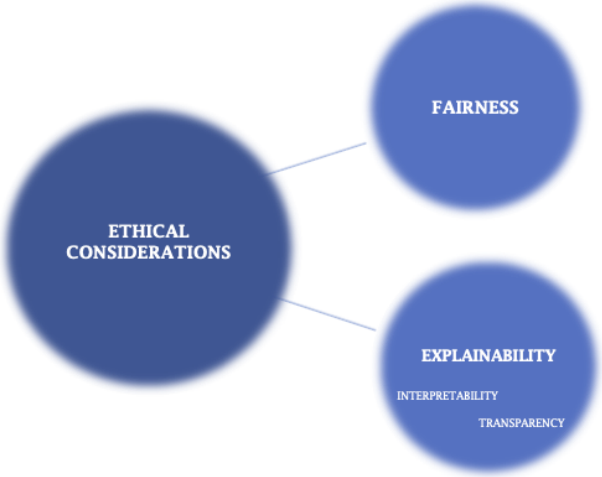

Figure 1.1: AI-related ethical principles this thesis will focus on are Fairness and Explainability, herein referred to as Ethical Considerations or ECs. Explainability also includes the AI-related principles of Interpretability and Transparency.

It is postulated then by query: can AI-generated decisions be mitigated, and if so, how? The position that will be taken in this thesis – one that has been developed by [25] – is that mitigation can occur through the application of Visual Analytics (VA). Fischer et al. [25] enlightened further on the dynamic between human (user) and machine (AI) – a position similarly expressed by other scholars – where decision-making is best achieved through a human and machine synthesis [25][26][27]. If it is postulated that such a synthesis can mitigate the ethical issues associated with AI-generated decisions, then an empirical undertaking to resolve this is to be pursued.

1.2 Aim

Bail decisions in NSW criminal courts are determined by a judge or magistrate. Despite sections within the legislation guiding the decision-makers, it is arguable as to whether or not exceeding that of human subjectivity, disparity, inefficiency, inconsistency, incomprehensibility, and unfair, could be better met in the alternative. A reasonable assumption inferred upon the alternative is for decisions to contain objectivity, and predictability, efficiency, consistency, explainable and fair. The two latter points are of interest in this thesis, accordingly, the research aim is to conceptualise AI-generated decision-making, by way of predictive modelling being applied to determine bail, and explore visual analytics as the ethical mitigator to these decisions being fair and explainable.

1.3 Objectives

The following section lists the three research objectives for this thesis.

Research Objective 1 (RO1)
Given the recency of AI-based technology being applied in criminal proceedings in jurisdictions outside of Australia, and the literature on Australian jurisdictions is theoretical and analytical, **RO1 is to identify and explore fair and explainable AI-based techniques in Machine Learning and Visual Analytics to use in bail decision-making.** This objective will be met through ongoing review of the literature and empirical undertakings.

Research Objective 2 (RO2)
After identifying and exploring fair and explainable AI-based techniques, **RO2 is to examine the effectiveness of the selected AI-based techniques at mitigating the AI-generated bail decisions in relation to the ethical principles of *fairness* and *explainability*.** This objective will be met by: (i) conceptualising a ML prototype to predict bail decisions in NSW criminal courts; (ii) facilitate VA using the ML model and data; (iii) survey VA mitigating effectiveness through participant engagement; (iv) provide a detailed analysis of the results.

Research Objective 3 (RO3)
While the ethicality of fairness and explainability will be resolved through RO2, **RO3 sets out to resolve a question on morality upon whether AI techniques applied to bail decision-making would place a greater value on (i) deontology or (ii) teleology:** (i) a singular justness/individual rights; (ii) a collective justness/public good.

1.4 Significance

Arguably, matters in AI are complex. Explaining and comprehending algorithms and computational data analyses – which are the essence of ML – may prove difficult to a layperson [28]. Compounding these difficulties are legalities, for instance, intellectual property rights and non-disclosure agreements [28] that regulate some ECs. These legalities will preclude some stakeholders from knowing how a decision was made by a court [28] such as defendants in bail proceedings. Given that a persons liberty is at stake, compounded by numerous ambiguities associated with law and justice systems, intelligible mechanisms like ML and VA could be the mitigator. Additionally, a rationale for substituting human objectivity and convention for AI decision-made mechanisms could be argued as being for economic reasons (cost-benefit) and efficacy – an interrogation here is of Kantian ethics and a contrast of Deontological and Teleological principles [29].

It would appear that there has not been any published empirical work directly examining analytical reasoning and decision-making by means of AI-based decisions (encompassing ML and VA) that: (i) can mitigate the ECs emanating from AI-generated bail decisions; (ii) can in-

crease understanding and decrease ambiguities made from complex AI-generated bail decisions; and, (iii) lead to policy and governance change for improved justice practices. Accordingly, this study is significant as to my knowledge: (iv) AI has not been applied to any criminal jurisdiction in Australia as a predictive decision-making mechanism; (v) and there has not been any published empirical research on AI to support predictive decision-making in the NSW criminal justice jurisdiction, with a specific focus on bail decisions.

1.5 Thesis Structure

This section provides a brief outline on the chapter structure and their contents, as well as, the research objectives aligned to that chapter.

A theme which will become apparent in this thesis, as similarly covered within aspects of the literature, is an argument on research being conducted and analysed in an exploratory way to ascertain its practical and empirical value of one or more concepts where the outcomes could be expanded upon. As it were, this thesis was conceived and completed in a manner whereby there is not one singular study; rather it is triadic of studies and designed as such that each complements the next, as illustrated in Figure 1.2.

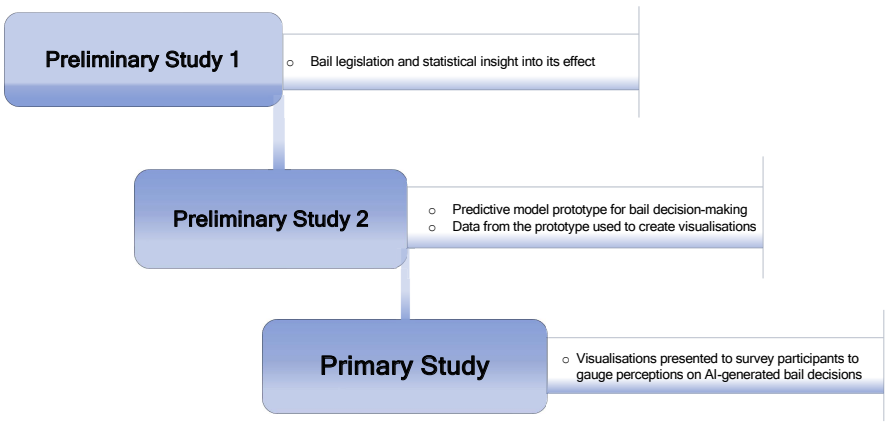

Figure 1.2: Triadic research formulate.

Chapter 1 contains the introductory matters. Chapter 2 details some of the literature reviewed as part of this thesis, aligned with RO1.

Chapter 3 contains Preliminary Study I. The first study is the groundwork on the bail amendments and the systemic consequences of conventional practice over a specified period. It provides the groundwork on the subject of bail and the relationship of the legislation having been enacted in 2013, then the amendments having been legislated and taking affect thereafter. This timeframe is important in the context of this thesis for several reasons. Firstly, the bail legislation and amendments introduced a schematic process by which the courts were required to decide bail for defendants in NSW. Secondly, most of the data in this thesis are conditioned upon the legislative commencement dates and the proceeding years that followed up until the time when this thesis was finalised. The statistical methods applied are linear regression and multiple linear regression. Preliminary Study I is aligned with RO1 and RO2.

Chapter 4 contains Preliminary Study II. This study is the exponent of the schema and systemic matters from which the bail amendments became the catalyst for predictive modelling being applied to bail decisions. Using two different mathematical decision instruments, namely, binary logistic regression and a tree-structured classifier, along with data obtained on bail decisions made in NSW criminal courts, predictive models are developed off of specified features within the existing bail schema. In order to balance consistency through comparison, it takes the observed systemic consequences to specify an alternative decision-making option as relief. Intelligibility of the predictive model outcomes then becomes a key motivation – apply visual analytics to the outcomes – although determining if this can be achieved ethically in accordance with the ethical principles of fairness and explainability. Preliminary Study II is aligned with RO2 and RO3.

Chapter 5 is the Primary Study. It is the culmination of the two preliminary studies of the systemic consequences of conventional decision-making for an alternative – that alternative is in predictive models being designed to make decisions and become intelligible through visual analytics. As a user-study, it is an examination of stakeholder perceptions on the alternative decision-making option by which to relieve the systemic consequences for the potential benefit of the greater population. The strength in this survey (and the research collectively) are the trademarks of real-world data and real-life bail decisions, that by intention, is wholly designed for authenticity and to sustain an ethical and moral sequence from data collation to model development to visualisations.Survey responses are calculated by evaluative metrics. Primary Study is aligned with RO2 and RO3.

Chapter 6 contains the conclusion and future work. This summarises all chapters and offers reflections and prospective developments from this project.

1.6 Contributions from this Thesis

In this section, the following items are those to which this thesis presents as contributions (in bold font).

As part of this thesis, public perceptions on AI-driven decision-making in the criminal justice domain are considered, building on other research conducted outside of Australia [30][31][32]. A key factor in the primary study will be to **gauge the public perceptions within the Australian legal jurisdictions**, and moreover, those who are likely to be affected by legal decisions – as a **user-study, it is designed to measure the perceptions of those participants who could be affected by such decisions, for instance, court-users and other stakeholders from two State criminal jurisdictions**. Participant perceptions were sought on court-based decisions on the premise that, if implemented in any Australian state or territory criminal jurisdictions, predictive decision-making would directly impact those participants who may be victims, employees, the disadvantaged, or those subject to the punitive interventions.

As AI appears to be building momentum, its complexities make it mostly intelligible to domain experts. Thus, gaining insight into those who may be unfamiliar with AI-generated decision-making more generally is salient in conjunction to whom consequential legal decisions were or will be made conventionally. As such, this thesis **analyses participants on their *confidence* and *agreement* for AI-generated bail decisions with the aid of visualisations individually created with real-world data.**

Apropos the ethicality of AI-generated decisions, the primary study **analyses selected visualisations as ethical mitigators to *fairness* and *explainability*; and, moral principles *deontology* and *teleology* as antecedents to AI-generated bail decisions.** As ethical and moral decisions permeate all facets of life; the evolution in AI, machine learning and prediction, facilitates the axiomatic dilemma on defining and regulating ethics and morality to minimise harm, yet advance understanding and acceptance.

This thesis has a specific focus on prediction. Consequently, predictive models were built based upon **nine well-defined predictors/variables, designed with specific intent of encompass the ethical parameters of *fairness* and *explainability* in unison with the variables for predicting bail outcomes**. The data from these models was then applied to create the visualisations used in the primary study.

Chapter 2

Literature Review: the amalgam of conventional philosophy and technological modernity to decide bail

2.1 Background and Motivation

The discourse on modern methods to support, supplement or supersede the conventional methods in decision-making – that is AI-driven rather than solely human-driven – forms the motivational premise detailed in the following.

A report on Australian prisons noted some significant issues regarding imprisonment rates and court administration processes [33]. For example, the rate of remandees awaiting court outcomes is almost twice the number now than it was in the year 2000 and the average time in custody has increased by 1.3 percent to 5.8 months over the same period to 2020 [33]. More specifically, in the state of NSW during 2020, the average remand time increased on previous years to 6.1 months - one factor being blamed for this increase was systematic issues with court processing [33]. Table 2.1 is extracted data on those defendants who proceeded to trial/hearing with a "bail refused" status from the initial arrest to finalisation; the data table also displays the court delays calculated by median days from arrest to finalisation, and the number of defendants relating to this criteria (the 2011 and 2012 data was published differently regarding the legal proceeding types). The data is based on three levels of the NSW criminal court structure; and extracted from the Bureau of Crime Statistics and Research (BOCSAR) [13].

Following the 2014 amendments, Table 2.1 shows defendant numbers fluctuating in all courts; yet the court delays/defendants remanded increased in the Local and District Courts, although the Supreme Court displays some sharp variations. Essentially, an implication from this data is that the defendants awaiting court outcomes are doing so while incarcerated at the peril of administrative and legal circumstances, only to find themselves released without any further penalisation.

Table 2.2 presents a different perspective when contemplating days on remand or court delays for defendants who are eventually discharged without penalty. During 2018–2020, defendants whose status was "bail refused" in all three court levels are counted equally in most

Table 2.1: Number of defendants bail refused (in custody) awaiting hearing/trial and court delays (median days denoted in brackets) by court level. Tabulated data extracted from BOCSAR [13]

Year	Local Court	District Court	Supreme Court
2011	6082 (35)	1179 (502)	49 (660)
2012	6318 (35)	1080 (517)	95 (814)
2013	5343 (80)	1009 (578.5)	54 (824)
2014	5175 (79)	1053 (595)	51 (849)
2015	7081 (81)	985 (658)	52 (774.5)
2016	7738 (82)	1176 (713)	59 (788)
2017	7979 (82)	1250 (716)	54 (839.5)

categories, although two categories of "not guilty of all charges" at 5.57 percent (n=579) and "all charges withdrawn by prosecution" at 4.74 percent (n=493) [13] are notable. These two categories, while proportionate to the number of matters finalised (n=10,393) may not seem overwhelming, implies defendants were remanded without any form of sentence being handed down by the courts despite lengthy periods of incarceration. Moreover, an implication that may be drawn from these two categories is a probability that decision-makers were erroneous by remanding some individuals, as opposed to granting conditional liberty. A further discussion could be had on other categories where defendants were remanded and whether bail would have been a more suitable option, for example, incarcerating mentally ill or cognitively impaired persons, although that topic is not of direct relevance here.

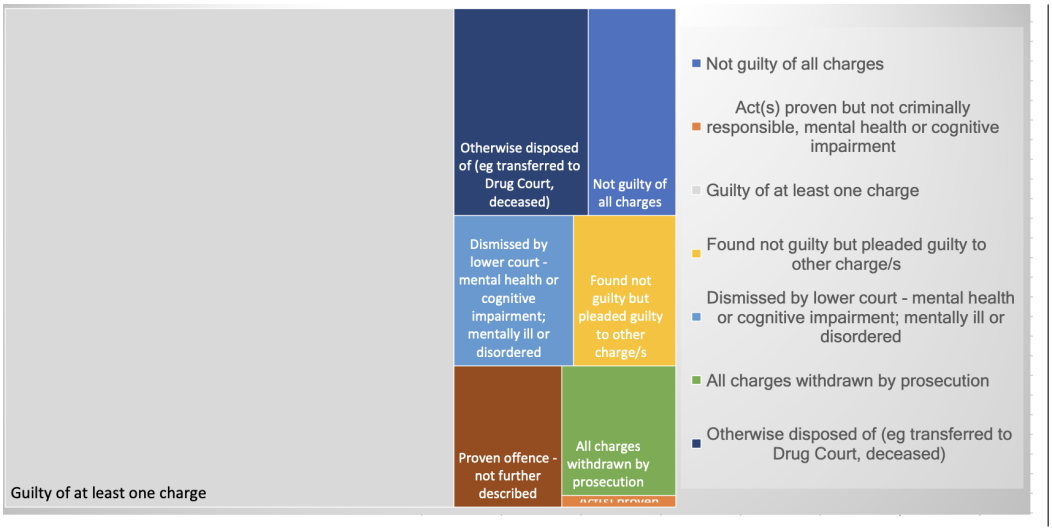

Table 2.2: Outcome by category of defendants refused bail in NSW criminal courts (2018–2020 (data for table extracted from BOCSAR [13]).

Apart from the legal and moral issues on probable erroneous decisions, another issue concerns the cost-effectiveness of potentially unnecessary incarcerations. In NSW, the budget concerning incarceration is increasing [34], and over the 2014-2015 financial year, the net cost to keep an individual in custody in Australia was calculated to be $61,179 annually as opposed to approximately $6,500 annually in the community on court-issued orders [35]. The assumptions made from these figures are on cost-benefit efficiencies: an accused on bail would cost significantly less than on remand, and again much less than community-based orders, as

generally there is limited intervention required for individuals on bail. Notwithstanding the cost-reduction argument, another issue is the time on remand that was not solely risk-based but attributed to laboured court processes and erroneous decisions.

Considering economic benefit and efficiency on sentencing, [15] commented that AI or algorithms designed to undertake such a task is fiscally responsible and would create efficiencies in court administration in addition to other benefits of consistency, transparency, and predictability. Subsequently, the objective is to improve the efficiency and accuracy of the criminal justice system, and it stands to reason that bail decisions conventionally made by the judiciary could be replaced with contemporary methods, inferring that human intelligence could be substituted by artificial intelligence.

Whitby [36] aptly reasoned that it may be the ideal that moral decisions are only made by humans; however, as technology becomes second nature to human life, it should be accepted that decisions will be made by machines. If the rationale by [36] is to be accepted in regard to a criminal justice system, it is therefore appropriate to examine the presuppositions that AI is to think and arbitrate like a human, its decisions should be moral and ethical, reasoned and justifiable; are to be good, right, fair and explainable.

Accordingly, the next section will review theories on ethics and morality, propose a conceptualised real-world model, and review the literature on VA as a means to mitigate the ECs assumed from AI as a decision-making mechanism.

2.2 Related Work

2.2.1 Moral and Ethical Considerations in criminal justice and AI

The appropriateness of algorithms determining someone's liberty – if it is good or right, and to what consequence – are moral and ethical considerations, and fundamental to AI-driven decision-making in criminal justice. This section will review these considerations underpinning AI and decision-making as they relate to the criminal justice domain for this research, namely, *Fairness* and *Explainability* and principles in *Deontology* and *Teleology* (or Consequentialism). Initially, considerations in Rawlsian theory on rights and justice are introduced.

Considerations on Rawlsian theory concerning rights and justice

It was opined by renowned western philosopher John Rawls' [*1971*] in *Principles of Justice* that for institutions to achieve the most equitable balance of rights and justice for all individuals within a society, those institutions must be arranged and ordered accordingly so that it maximises benefit for all its members, and that every individual should be granted the same rights to a complete system [37][38]. Rawls underscored one disadvantage of a justice system is within its incorrect denial or interference with one's freedom due to sanctions or punitive action posing a threat to an individual [38]. Interpreting the Rawlsian position on AI as a decision-making mechanism in an adversarial criminal justice system, it is reasonable that an equitable outcome would be achieved by machines rather than humans, for realistically, AI would make the same

decision each time. Contrastingly, it can be argued that justice systems relying on human-based decisions leaves an inequity, which is evident by lower courts having the remedy of appeal to higher courts. Then, the ideal needs consideration: trust in a machine to make decisions that are equitable. Rawls' first principle says that individuals are entitled to "equal basic liberties" [39], and if this is accepted as a fundamental tenet of AI, then there are several considerations that need to ensure what is right and what are its consequences.

Considerations on Deontology and Teleology

On morality, machine ethics and ethical evolution, standard ethical theories were premised as either *deontological*, where actions are good or bad, right, or wrong, or *teleological*, which asserts the outcome is of consequence [40]. *Deontology* asserts the intention is more important than the outcome and it "makes the right prior to the good" [29]. *Teleology* asserts the outcome is more important than the intention – outcomes for what is right or wrong are concerned with consequences - and it "makes the good prior to the right" [29]. In differentiating the 'right' from the 'good', the former asks of individualised efforts for others and themselves, and has one developed "moral code" and natural "moral instinct"; the latter asks more broadly on circumstances of subject matter and its appropriateness for a given society [40].

The literature on Deontology is extensive and cannot be expounded here, therefore, a consideration of it will be based on Rawlsian theory in terms of justice and fairness. The deontological position is that the outcome is inconsequential, inasmuch as all individuals are equal and there must not be an uneven distribution, or disparity of intention, to arrive at a particular outcome; justice would be principled upon fairness and equality, and those principles underlying it are developed and agreed upon equally by citizens, legislators and the judiciary alike [41]. If this is then advanced upon AI-generated decisions in criminal justice, all stakeholders would have shared input into its development and implementation. Then, in its practical sense, the deontological principle suggests that if the intention is to use AI in decision-making, then its embodiment must be persuaded by many stakeholders, and the decisions of consequences must be fair and equitable.

The literature on Teleology is expansive alike to that on Deontology and cannot be expounded here either, yet consideration if it can be made on its recognised branch of Utilitarianism. Teleology, which is also known as Consequentialism, is the principle of "right conduct" in that the conduct or actions may be considered right, based upon its outcomes [42] whereby the actions and outcomes benefit a majority [43]. Moreover, in the discourse on explainability (as will be discussed in this Chapter), the justification upon what is considered a right or wrong action may be done so premised upon the consequences or outcomes of the said actions; then also in quantifiable enumeration, what value do those actions and outcomes have for individuals [42]. A consequentialist argument was made on the implementation of technology in law with a focus on algorithmic decision-making: its application and use "...must be guided by clear goals, such as decreased disparities, increased efficiency, or fewer incarcerations"[44]. The implication by the author is that using advanced technology must have the ultimate objective of liberty, to significantly reduce an accused being detained prior to a hearing and to minimise racial

disparities.

A final point of relevance on Deontology and Teleology in its broader application examined in the literature was on their correlativity, that is, to resolve whether their is any relationship between them or are they mutually exclusive. The contention was made on theoretical and measurable qualifications, to determine any distinction between them [45], and in the context of moral judgment, whether harm to one or more individuals is acceptable or unacceptable [43].

Considerations on AI in criminal justice: Fairness

Not to be confused with algorithm fairness, with its own consideration more technically (discussed later in this Chapter), this differs as it presses the moral and ethical implications of decisions made in the law-justice domain materially.[1] Folger and Cropanzano [*1998*, *2001*] were cited by [46] as having developed fairness theory: an individuals displacement of blame from themselves onto the justice system for a perceived incorrect decision. These individuals consider the authority is at fault for not making the alternate decision in their favour [46]. Miller [47] remarked on fairness as it would apply to administrative law: substantive fairness asks if a decision or outcome was fair; and procedural fairness refers to how a decision is made. Miller [47] highlights issues on 'fairness in AI' as it relates to cultural and linguistic concepts, understood to mean that models and algorithms when designed for the Australian legal environment, should acknowledge that a fairness construct in a material sense, may not be universal. In another literary source, the fundamentals of fairness in decision-making processes was noted as: consistency for people and time; restrained bias; accurate information; means to vary poor decisions; support for people affected by poor decisions; and ethical standards [46]. The meaning of fairness in decision-making while nuanced, equivalently maintains the explication in ethics and morality.

The same contention is made on *algorithm fairness*. Castelnovo et al. [48] stated pragmatically that fairness as a construct contains many elements, and will have "different meanings and nuances," and conjecture is expected in moral and ethical matters. Kaas [49] held that there is not any broad agreement on the definition of ethical concepts such as fairness, and similarly, [50] suggested that in the attempt to find the meaning of fairness, conflicting positions will ensue. Nevertheless, it would be instructive to have an agreed definition, particularly given that 'fairness and justice' were second to 'transparency' as the most important ethical principle found in a collective assessment of AI guidelines [51]. A succinct definition of fairness was issued by the Australian Government in its AI Ethics Framework, stating that "AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups" [52]. Algorithms formulated for criminal justice decision-making are intended to represent fairness and lessen bias, but they are not infallible and predictive errors will occur and will be challenged [53].

Where the path remains opaque, [54] explained that decisions made by computer are more simplified, for it is automated and ordered, although the authors expressed an argument that

[1]For the purposes here, the meaning of 'materially' contrasted with 'technically' assumes the outcomes based on the fairness principle are corporeal as opposed to automaton.

automated decision-making is flawed due to transparency, fairness and bias. A focus in their paper was on the often-cited court case *Loomis v The State of Washington*, whereby the applicant Loomis had attempted to challenge a decision ultimately influenced by an algorithm in the program titled Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (see [55]). Loomis, as they stated, was the so-called test or landmark case in adversarial practising jurisdictions globally to argue whether an algorithm can be relied upon over a human-made decision. While the decision made by a U.S. state higher court did not favour the applicant, an outcome that had future implications was to place a disclaimer on any result made by a program such as COMPAS [54]. The authors noted however that a judge's decision would be persuaded upon a label of high-risk next to any offender's name based on a decision made by COMPAS. Nonetheless, the judge in the Loomis case made a telling remark that should also be considered in light of AI being used in any jurisdiction, which was to draw a clear distinction between "considering" and "relying" on a computer-based decision [54]. This distinction raises an important point: courts can mitigate an AI-driven decision by overriding it with a human-made decision.

A significant focus found in the literature was on public trust in machine-driven decisions. In a study by [30], it was reported that Americans were not in favour of algorithm decision-making. In one category tested on criminal risk assessment for offenders eligible for parole – which is a score based on data from offenders assessed against their own and others with criminal convictions – 56 percent of the public surveyed felt algorithmic assessments determining parole were unacceptable with serious consequences. Also, 49 percent favoured a risk assessment determining parole suitability; the sentiment of those surveyed indicated an even balance on fairness. As automated decision-making systems become more common, accountability increases, affecting matters concerned with intentionality, fairness and outcomes [56].

A discussion of this type regarding machine ethics in criminal matters was undertaken in a more recent study. Fine [31] examined the American public perception on the use of a machine learning approach for bail and sentencing risk assessments against conventional approaches. Even though participants had expressed ethical concerns regarding machine learning risk assessments in bail matters, they were not persuaded in their judgments on the sentencing aspect between conventional and non-conventional methods [31]. In the sentencing section of the study, the results demonstrated that participants favoured conventional risk assessment approaches to computerised ones; and, their perceptions of offenders were impacted by evidence from experts and the seriousness of the offence, but not from computerised assessments. Fine [31] determined that participants challenged the use of machine learning in bail considerations and its ethicality, finding it an unacceptable method of decision-making.

In light of the research on people perceptions, it is fitting to conclude on fairness with a theory that supports human cognitive capacity in decision-making. Stibel et al. [57] challenged an argument put forward by Johnson-Laird and Byrne [*2000*] that cognitive performance and thus rational decisions are negatively impacted when choices are increased [57]. These researchers undertook an examination of human cognition and choice variation, terming it Collapsing Choice Theory (CCT), and hypothesised "working memory limitations" prompted by increased choices benefited cognitive performance and decision-making [57]. The researchers found that

giving participants either a choice of three, or one-hundred alternatives, did not have any significant bearing on their judgment or choices, in fact, determining that there is a distinction between judgement and choice. Another outcome was that incorrect choices were lessened when participants were able to mitigate regrettable decisions they had made. The most critical outcome, but also a limitation, was the impact on participant decision-making from biases. An implication from the findings from [57], if assumed in a criminal justice context on bail determinations, is that a human would not be overwhelmed if multiple factors were at play in decision-making, which suggests that any challenge to fairness of human-driven decisions based on cognitive capacity may be unjustified.

Waltl and Vogl [1] indicated that a feature or variable in an algorithm that could influence an outcome, such as age, gender, or race, could in fact lead to a discriminatory outcome and consequently would require objective assessment. However, in the circumstance where a variable had been different, [56] commented that affected individuals can be provided with "contrastive explanations", which basically provides individuals with an alternative outcome to a decision. It would appear this would address the issues raised by fairness theory (see [46]).

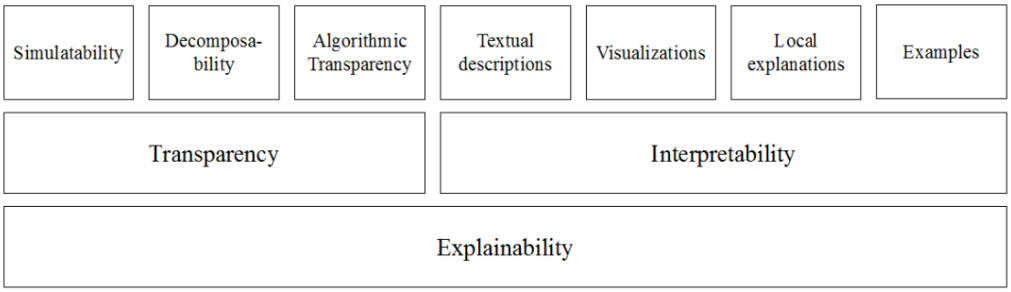

Figure 2.1: Taxonomy of Explainability in Algorithm Decision Making [1].

Hedden [58] puts forth a position that algorithms inherently have bias, and that "calibration" is the only feature in predictive algorithms not impacting fairness. As was observed by [54] and [59], bias in AI and justice can occur in numerous ways, such as, legal and technical. In a machine learning model for example, legal bias concerns how a model and data are applied; technical bias concerns the techniques upon that a model and data would operate [54][59]. Mannes [60] remarked that training data, erroneous data collection, and the inclusion and exclusion of variables, would impact the results and argued that public data sets are inherently biased, particularly in recidivism forecasting [60]. More optimistically, [15] contended that bias and other discriminatory matters can be overcome. On human judgement and bias in court decisions, [15] indicated that subconscious bias is an issue; more discerningly, "vividness heuristic" as a type of cognitive bias, implies an individual gauges a probability of something occurring on their recollection of other occurrences.

Berk et al. [53] analysed data over an eight-year period to determine if they could eliminate racial bias from decisions made in a U.S. criminal court on offenders who were denied bail following arrest. The single biographical indicator the researchers applied was 'age'; and, one of the approaches they varied was training the data for white offenders only, which had the effect to remove racial identification. A dissertation by [61] researched the use of algorithms

in various criminal justice settings, and one of those studies attempted to mitigate against racial discrimination on the part of the prosecution through the design and implementation of a Blinding Algorithm Model, which effectively attempts to remove the race feature from incident reports prior to it being read by a prosecutor. A criticism of [61] dissertation is made on the suggestion from one sub-study to it being the initial empirical examination of obscuring racial features in decision making; however, another sub-study uses the COMPAS model for risk assessment but does not appear to mention the critique of that model concerning racial bias. As such, these two sub-studies appear contradictory in that one tries to consciously exclude racial biases when the other included it without acknowledgement, perhaps inadvertently.

A recent study by [44] reviewed bias in the law and decision-making. A sub-study asked participants if they trusted a human or machine to make legally based decisions. The results showed that the participants had more trust in a human to decide over a machine. This was the case in another sub-study by [44] where participants were told that there was bias in both the human and machine outcomes, yet the participants still held more confidence in human decisions. The resulting confidence and trust for humans over machines as demonstrated in the research on decision-making must then be attributed to a fundamental characteristic of AI, and while these studies may not have directly identified it, the characteristic of explanation is pertinent.

Considerations on AI in criminal justice: Explainability

As was formulated in the taxonomy by [62], Explainability (xAI) encompassed several aspects in decision-making, including Transparency and Interpretability. Transparency regards the individual having a clear understanding upon how an AI-based decision is made; Interpretability, or "post-hoc interpretability", provides practical information that is intelligible [62]. Ng et al. [63] nominated two issues within the broader challenge of transparency: one of invisibility and the other of complexity. The authors are essentially claiming that individuals will be unaware and have limited or no knowledge that AI-based decisions are being made about them, and as such, an adherence to transparency and explainability, alike, are integral for institutions. xAI was characterised as the provision of a reason or cause from one entity's or agent's decision-making to another; however, AI has elements of complication, ambiguity, and obscurity, and reasonable to contend this would be exacerbated for the lay person and those with lesser technical proficiency's – any difficulties understanding AI may lead to mistrust by users and raise ethical concerns [47]. Trust in autonomous systems can be gained through two approaches: firstly, assess the level of understanding of any decisions made to ensure they are interpretable and explainable; secondly, provide unambiguous explanations [47]. Consistent with the other literary sources, [60] affirmed xAI was one approach to manage these concerns; however, argued that an explanation may not be sufficient, and trust in this mechanism will be needed from an individual or more broadly. Even though [60] expressed partial support for xAI, he failed to make mention of 'visual analytics' or 'analytical reasoning' as explainable options, contrary to other literature (as discussed later in this Chapter).

It is of relevance to examine interpretable algorithmic models as constructs. A framework from [62] details the necessary elements for an Algorithmic Decision-Making (ADM) model to comply with xAI principles (see Figure 2.1). Explainability as the foundation is divided into two types, Transparency and Interpretability. There are then three sub-types of Transparency: Simulatability (the entire model should be understood); Decomposability (each feature/parameter should be understandable); and Algorithmic Transparency (the algorithm should be decipherable). There are then four sub-types of Interpretability: Textual descriptions (although [62] termed this "Textual explanations" is descriptive wording being used to provide model justification); Visualisations (qualitative overview of data); Local explanations (a particular feature of a model is described); and Examples (although [62] termed this "Explanation by example", is understanding the model through what it has learned) [62].

Mittelstadt et al. [56] identified some of the ethical queries that an individual may consider when a decision-making algorithm is applied to them: was the process fair and what could change to improve favourability in the future? The authors concluded, if decision-making algorithms are to be trusted and have accountability, they need to be explainable and interpretable [56]. Therefore, the objective must be to build that trust based on measures outlined in Lipton's taxonomy, but also certain measures such as validity (how well does it perform against others) and accuracy (are the machines and results trustworthy). A *classification table* and a *Receiver Operating Characteristic (ROC) curve* are two indicators in classification modelling that provide a visual output on accuracy and performance [64][65]. Even though classification tables and ROC curves are based on and contain mathematical properties, they maintain a level of interpretability and explainability through visual representations that would not have otherwise be gained through a written equation.

Table 2.3: Classification table exemplar for Actual and Predicted bail decisions. Literal notations are True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) where the actual or true classifier values are represented vertically, and the predicted classifier values are represented horizontally

	Bail Granted (Actual Decision)	Bail Refused (Actual Decision)
Bail Granted (Predicted Decision)	**True Positive (TP)**	**False Positive (FP)**
Bail Refused (Predicted Decision)	**False Negative (FN)**	**True Negative (TN)**

Table 2.3 is an exemplar of a classification table. The letters are literal notations of True Positives (TP), True Negatives (TN), False Negative (FN) and False Positives (FP) [66]. The vertical columns represent the actual or true classifier, and these are put against the outcomes from a predicted classifier in the horizontal columns [66]. At this basic level, the classification

table provides one discernible option to support some of the principles outlined in the framework by [62], most notably, algorithmic transparency, explanations and interpretations.

Another means to support matters of explanation, interpretation and transparency is the ROC Curve. The ROC is purposed with measuring the performance of a ML model: for explanatory purposes, Figure 2.2 is an example of an ROC curve with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis [66]. The ROC curve in this example illustrates a dotted line on the 45-degree angle that represents 0.5 or a prediction of 50 percent that denotes an outcome of not greater than or less than chance, which is not the intended result for classification, as more certainty is generally sought [64]. The curved line with the black dot moving in a direction away from the dotted line toward the 1 on the y-axis, is a demonstration of greater accuracy [64] as demonstrated in Figure 2.2.

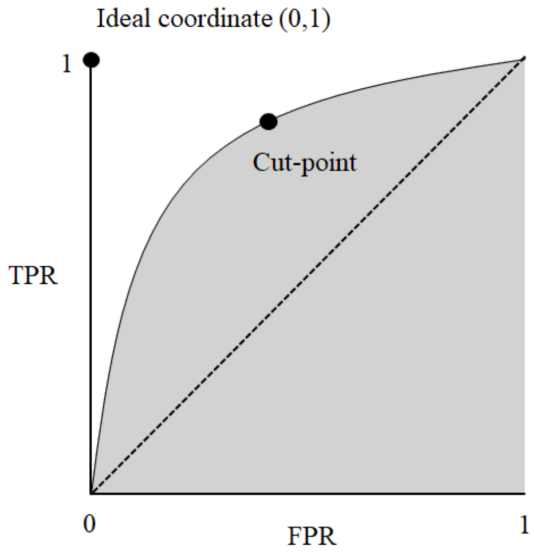

Figure 2.2: ROC curve example [2].

An example from the literature illustrated how the ROC curve can be applied in criminal justice research. Ghasemi et al. [67] initially measured recidivist rates on a large dataset and then assessed the accuracy of three ML instruments (decision tree, random forest, support vector machine) against a traditional yet respected predictive assessment tool (Level of Service/Case Management Inventory (LS/CMI). The evaluative methodology applied by the authors was through the ROC curve and classification table as observable references. The results demonstrated that all four statistical methods had a very similar level of accuracy with random forests (.734) and LS/CMI (.736) sharing the closer margins. Therefore, the results at around 73 percent is indicative of accurate classifications.

Miller [47] asserted that any resolution for xAI is not to add more layers of AI, rather it lies within the "human-agent interaction problem" characterised as "the intersection of artificial intelligence, social science, and human-computer interaction." However there needs to more effective means to achieve this, as such, several alternatives were put forward in the literature of "interactive methods", "visualisations", "verbal (natural language) explanations" and "interactive interfaces" all of which are elements within visual analytics [56].

Considerations on ethical and moral machines

The supposition as to whether machines can be made, or are, ethical and moral, is a dilemma expounded in the literature. Moor [68] wrote extensively on the existence of Machine Ethics and reasoned on this as a philosophical dilemma – given the technological world that now exists – that computers are making decisions for humans that have ethical implications. Moor [68] made two pertinent points on the importance of Machine Ethics: first, with machine evolution comes machine autonomy; and second, human ethics will need to be programmed within these machines. Moor [68] proposed having several ethical values, notably explicit ethical agents and full ethical agents. In defining these values, [68] stated that humans are full ethical agents, having "consciousness, intentionality and free will" – explicit ethical agents would be those machines with the capacity to assess the most appropriate ethical decision in the same that a human can. It appears the author was unconvinced that machines have attained the 'explicit' status, but stressed an importance on it, given machines will eventually gain more autonomy and responsibility. Even though it seems the author is unconvinced if computers can become full ethical agents with human ethical capacity, an emphasis made on ethical theory signals a design of this program type and what it would need to comprehend for real-world circumstances, must occur.

The morality of AI as a decision-making means has been questioned in the literature. Ryberg [69] expressed objections to AI technologies being applied as an alternative mechanism to conventional human-made decisions in criminal matters at sentencing. It was emphasised that the proportionality aspect in formulating the algorithms and building a database that provides the morally acceptable outcomes for each offence is thwart with difficulties. While [69] emphasis on 'morally desirable' and 'morally preferable' positions in sentencing is understandable, it does not contend that desirability and preference in moral discourse is subjective, and what one believes is morally acceptable punishment may not be agreed upon by another. This sentiment was similarly expressed by [70] where the morally correct action is debatable as those involved will have varied opinions. Additionally, [46] deliberated that "moral codes" and "moral content" may be acceptable or unacceptable between groups of people or individuals.

The opposition to morally operable machines was challenged in the literature. Whitby [36] offered the comparative proposition of "flint axes to computers" implying historically, humans have utilised some type of mechanism to support physical and intellectual ability, and AI is not any different. This proposition is understood to mean that there was not any reasonable opposition to an axe being used as a necessary tool supporting human advancement, correspondingly, there should not be any opposition to a computer being used as a necessary tool supporting human advancement. An articulation of this nature was made some two centuries ago by moral philosopher Adam Ferguson [*1792*]:

> ...the progress of the species itself will, without their intending it, keep pace with the ordinary pursuits, in which successive generations are engaged [71].

A question put forward in the literature asked what approach is most suited to implement machine ethics in AI systems, Bottom-up or Top-down? Battle and Heer [72] contended that neither strategy is a complete representation of suitability. Kaas [49] constructed a persuasive argument in favour of the Bottom-up approach to be best suited to implement machine ethics, and extrapolates his position on Reinforcement Learning as a Bottom-up approach over both supervised and unsupervised learning methods. It is useful however, to initially outline the reasons why the author is against the Top-down approach. Kaas [49] argues that Top-down is underpinned by deontological and teleological ethical principles that are restrictive and indefinable, rhetorically asking, what defines fair, right, good and harm? Moreover, multiple ethical rules or principles can be conflicting and domain dependent, that is, determining which action is the most appropriate and by what entity. Contrastingly, Kaas [49] favours Bottom-up as it learns as it is training through feedback and discovery, which allows it to be flexible where rule changes will not impede its direction and development, and it is independent of domain specificity. Reinforcement Learning as a Bottom-up approach for [49] is more favourable due to its organic ethical development, made through self-evaluation or feedback as it learns and is rewarded, but importantly, user feedback also shapes a machine to behave ethically. Even though [49] is arguing one approach over the other, he noted two pertinent observations: Bottom-up is not immune to the subjectivity of ethical principles and rules; and, a machine considered ethical does not designate it as having moral agency.

Regression algorithm models as predictive instruments to decide bail

Theory on machine ethics as previously mentioned is somewhat divided, although a consensus can be stated that the technological evolution is undeniable, and AI and ML are going to be integral. Subsequently, the issue becomes the preferred algorithmic approach or model to address the ECs. Regression is essentially the statistical measure to determine what effect a dependent variable has on one or more independent variables – it is a robust approach to predictive analyses [73]. In criminological and justice domains, a binary logistic regression (B-LogR) model is favoured, as fundamentally, the outcomes being sought are dichotomous [74], for example, bail granted versus bail refused, imprisonment order versus community order. It shows good predictive utility [75] and satisfies assessing or measuring its statistical significance on a dependent variable [74]. This however is not to discount other models that are effective in the criminological and justice domains. For example, the multinomial B-LogR model, as an extension, is effective for it can assess three or more categories [74] and was also said to exhibit low bias and "mean absolute error" [76]. Nonetheless, there are limitations to these models. Presumably, bias can occur due to methodology and data errors [77] and small sample sizes can negatively affect results [76]. It is arguable then to suggest these limitations are hardly at the severity of those in so-called black-box models. See Table 2.4 for a summary.

Table 2.4: Benefits and Limitations of Linear and Logistic Models – from the literature.

Benefits	**Limitations**
Less complicated than more advanced models [47].	Regression coefficients and weights can be erroneous (e.g., false positives to false negatives) [47].
Multinomial exhibited low bias and 'mean absolute error' [92].	Results negatively impacted by small sample sizes [92].
Ease of interpretation and use [95].	
Good predictive utility [91].	Binary Logistic Regression is restricted to binary outcomes [47].

Tree-structured classifier model as predictive instruments to decide bail

The next approach, a Tree-structured classifier (TsC) model, is expected to satisfy ECs in criminological and justice domains. TsC are supervised classification and predictive models [78], which categorise datasets into smaller subsets [79] ultimately providing a visualised figure that represents the data, structured in a tree-like format [78][79] (see Figure 2.3).

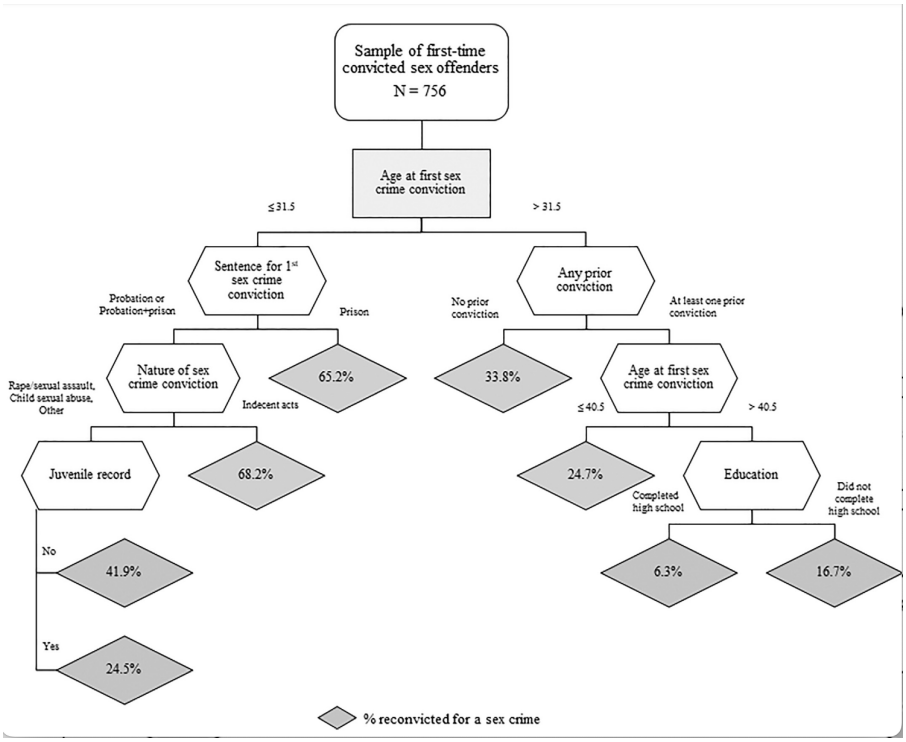

Figure 2.3: Screenshot of a tree-based model predicting recidivism in sex-offenders (CART method used to analyse risk factors for first-time sex-offenders)[3].

This approach was said to be favoured by social scientists in criminological domains for its adeptness in constructing the probability of recidivism on relevant factors [3]. Additionally, it resolves complexities in decision-making while simplifying interpretation and use [80], similarly characterised by [81] as being intelligible. TsC also shows good predictive utility and accuracy

[82][75]. TsC was also touted as a preferable approach as it corresponds well with flowcharts [80]. Importantly also, a TsC can be visualised [80]. Despite the many benefits of a TsC, there are limitations. The literature suggests that the models are not always interpretable, such as black-box models, and trade-offs are needed for interpretability and accuracy [82] and the more model complexities lead to negative outcomes [81]. See Table 2.5 for a summary.

Table 2.5: Benefits and Limitations of Tree-structured Classifier Models – from the literature.

Benefits	**Limitations**
Good predictive utility with accurate models [46][91].	Over-fitting can occur if training data is unbalanced [96].
Ease of interpretation and use; resolves complexities in decision-making [96].	Model complexity (bigger trees) can result in worsened outcomes; complexities lead to limited understanding in decisions [97].
Supports flowchart transition/interpretation [96].	Models are not always interpretable; trade-offs are needed for accuracy and interpretability [47].
Outcomes can be visualised [96].	'Black-box' models lack transparency, interpretability [46].

Conventional to Contemporary Predictive Instruments to decide Bail

A former chief justice of the High Court of Australia, in a speech on the current and future implementations of technology in Australian courts, opined:

> Technology is woven into our daily lives. It is the now, and the future. One does not need to look too far to see mistaken disregard of technology in the past [83].

More specifically on AI, technological influence was declared on a macro level. The OECD [84] noted the influence AI has had globally, bringing about transformations in economies and workplaces through increased productivity and innovation, and conceivably, enriching the lives of people and societies [84]. It is essential nonetheless to draw from the literature a meaning of AI, expressed in the guidelines from [84]:

> [When programmed with] human-defined objectives [AI is] a machine-based system that can make predictions [or] decisions influencing real or virtual environments [and is] designed to operate with varying levels of autonomy [84].

A definition cited in a criminal justice aspect was that AI is "the scientific understanding of the mechanisms underlying thought and intelligent behaviour and their embodiment in machines" [28]. Meaningfully, the objectives are human driven, yet programmed into a machine to make predictions that have real-world implications.

Model Prototypes

Earlier work on a model prototype for decision-making in the Victoria criminal justice jurisdiction was exhibited in the literature. The project conceptualised a decision support system

for sentencing in the Magistrates' Court of Victoria [85]. The authors decision-support system was said to apply a statistical-algorithm approach where data of previous sentencing outcomes or matching similar cases would inform decision-makers; the algorithm framework, said the authors, was an integration between decision trees and argument trees. Once satisfied with the theoretical model, the author's implemented it into a practical model using a web-based program, which generated responses or prompts from weighted variables to ultimately determine the sentence. Another literature source expressed their motivations for a decision supported machine was underpinned by consistency and efficiency – their argument on consistency was based on the probability that decision-makers examining the same facts can make different decisions [86]. The authors also claimed a benefit of their model for decision-making was "transparency", which will improve public understanding while lessening criticisms in legal decisions. The limitations expressed by the authors on their prototype were more legal than technical, which was likely due to the fact their research had not tested the technical elements on actual cases, consequently, its effectiveness could not be assessed [86]. Notwithstanding, the authors demonstrated that a schematic model can become an algorithm prototype model.

A recent survey considered predicting outcomes of criminal cases using various ML classifiers. Shaikh et al. [87] referred to their research undertaking as a "legal approach" which surveys the narratives of relevant legal cases in search of specific text or phrasing that create the features. The relevant information is entered into a database upon which the chosen ML classifier analyses and provides the outcome. The authors selected eight ML classifiers (see [87] for a full list) to analyse and predict the outcomes of 86 cases. It was determined that 64 cases were correctly predicted by the eight classifiers, yet the Classification and Regression Trees (CART) was the most accurate (91.86 percent) and had the best performance (91.76 percent). The claim made by these authors on CART is inconsistent with that made by [82], as noted in this Chapter; although, the authors reported that all eight classifiers provided respectable predictions in accuracy (between 85 percent and 92 percent). The research also showed 22 cases as having a minimum of one incorrect prediction and two cases wrongly predicted by all the eight classifiers. A limitation of this research was found to be in the high number of predictors and descriptors that were arbitrary, for as it was contended, predictors that are weak can be counterintuitive [88]. Despite the limitation and some predictive errors, the research by [82] shows statistical promise in that data analysed from legal case-narratives can provide reasonable outcomes.

A similar research question to that previously mentioned was taken by other scholars, although a discernible difference was apparent as to the most effective model. Zeng et al. [82] hypothesised that the ML classifier performance of the Supersparse Linear Integer Model (SLIM) would provide superior accuracy and interpretability over eight other classification models at predicting recidivism, including CART (see [82] for the eight models). These authors were of the opinion that SLIM is well-suited for quantitative research in criminology; it has similarities to conventional linear risk assessments for recidivism where a user is able to observe what input variables are influencing the result, satisfying transparency; yet SLIM as a ML technique is different to more notable ML approaches in its calculable properties (e.g. scalability). The methodology compared nine ML classifiers, including SLIM, on features that they said were

normally accessed by police and judges (e.g. prior arrests and imprisonment history) to predict six offence categories of arrest, drug possession, and four types of violence. A limitation from this data collection method is that bias can be contained in the data, given that a portion of it has come from policing records, for example, prior arrests. Despite this, the researchers reported their preference for SLIM over the other approaches for it uses a simplified numerical scoring, proven by its accuracy and interpretability (on metrics later discussed in this Chapter). Zeng et al.[82] emphasised the ethical consideration of transparency can be mitigated as a user of SLIM in predictive assessments has influence in variable input and can observe what influence the variable has in the assessment result.

Further discussion on variables was considered in the literature. Stobbs et al.[15] claimed that any type of algorithm coding for decision-making that has multiple variables can be produced. Stobbs et al.[15] stated that the NSW courts are presently guided by 30 factors in accordance with the sentencing legislation, although they reported over 200 factors to reference from; however two key factors, criminal record and offences while on bail, were identified. This discussion is complemented by a real-world model: Jung et al.[89] designed and tested a model to determine whether accused persons should be granted bail, assessed by two features, age and prior failures to appear [89]. After analysing over one-hundred thousand pre-trial detention cases, the authors claimed their model would have allowed judges to detain one-third the amount of people, and would not have increased the risk of any individuals failing to appear at the next court hearing if they were bailed [89]. A limitation in the research was found where a miscalculation may have occurred as to the reasons a bail authority released an individual due to inaccurate or missing information.

A different research study considered what variables were more prominent in failure to appear (FTA) matters. Zettler and Morris [90] applied a logistic regression model and analysed bail cases over a year to ascertain from nineteen predictors, such as criminal history and prior FTA, what were more likely for defendants failing to present at court. The researchers' design variation applied six models, five that were grouped by race and gender to limit bias, and a general model. The results demonstrated that the predictors of being male, impoverished and not having a prior felony charge[2], increased the chances of FTA. Despite the encouraging results, there were several limitations noted from this research. Firstly, the outcomes are only specific to that jurisdiction. Secondly, the monitoring of persons on bail by a government service is not applicable everywhere, and consequently, the government intervention in this instance could positively affect compliance. Thirdly, the information on the subjects may have been limited, particularly the data on homeless status.

Despite the persuasive arguments on the amount of predictor variables, the benefits in applying criminogenic factors in predictive models is evident. A meta-analysis conducted by [91] on predictors of adult offending and recidivism concluded some of the most significant criminogenic factors were age, gender, criminal history, pro-criminal associates, family circumstances, and substance-related issues. Additionally, a synthesis of actuarial risk measures in recidivism provided a higher correlation than personality measures due to heterogeneity of reoffending

[2]A felony is similarly equated to an indictable offence in Australia.

factors [91]. In another meta-analysis, [92] reviewed 136 studies regarding "clinical versus mechanical" or human versus algorithm prediction in the human health and behaviour field. The results demonstrated that the algorithm prediction was more accurate than human in predicting criminal behaviour. The meta-analysis reviewed studies that dated from 1936 (parole success or failure) to 1988 (criminal behaviour) and it is noteworthy that forecasting criminality in that 52-year period did not utilise sophisticated computer software and any comparison drawn on these outcomes on current forecasting should be made cautiously. Notwithstanding, it is appropriate to note that statistical prediction was equal to or more favourable than human-made prediction.

There were further promising results on machine versus human prediction reported in the literature. A study undertaken by [93] examined the effectiveness of ML on release decisions for inmates seeking parole from New York state. The researchers applied a Neural Network algorithm to analyse 18,688 cases: 70 percent used as training data, 20 percent as validation data, and 10 percent as testing data; 10 predictor variables were used including age, sex and race; and the results were visualised. The researchers claimed the accuracy of their model was calculated at 76.8 percent. The result while auspicious, had several limitations. For example, the model as a so-called 'black-box' program that used hidden layers, and the algorithm was undisclosed [1]. An additional limitation was found from at least one attribute of 'race', which was certain to create inherent bias. Berk et al.[94] compared a ML program to human-based assessment on released offenders and found that 10 percent of the offenders they deemed suitable to be released, had reoffended within two years, which was more accurate than the conventional assessment with a 20 percent reoffence rate. Despite the favourable results, a notable limitation in the methodology was observed regarding the high number of variables, whereby, the greater the number of variables or features applied can result in greater complexity and difficulty in interpretation [95].

A more conventional decision-making approach in prediction was found in the literature. Lum et al.[4] conducted a pilot-study study to determine the effectiveness of a predictive instrument, and associated issues in bias and fairness in police charge categories. The predictive instrument applied was the Arnold Public Safety Assessment (APSA) – it can be described as being a more traditional statistical measure using a six-point scale, using various features (e.g. prior failure to appear at court and prior convictions) resulting in a binary outcome. Its purpose is to determine if an accused person should be released or held after being charged with one or more offences[3] and the probability of the accused, if bailed, reoffending and/or failing to appear at court. The results from the scale assessment are then weighed against a decision matrix that was created for the pilot study, shown in Figure 2.4, which provides the recommendation. The researchers relied upon data from criminal cases over a 12-month period (2016-2017) in a U.S. criminal jurisdiction. Given there were several items analysed by the authors, for brevity, there were two notable outcomes: 27 percent of bail cases were incorrectly assessed by the APSA, which would have resulted in unnecessary intervention and restrictions

[3]This is similar to police bail decisions in NSW.

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	OR - NAS	OR - NAS				
FTA 2	OR - NAS	OR - NAS	OR - NAS	OR - Minimum	SFPDP - ACM	
FTA 3		OR - NAS	OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended
FTA 4		OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended	Release Not Recommended
FTA 5		SFPDP - ACM	SFPDP - ACM	SFPDP - ACM* / Release Not Recommended	Release Not Recommended	Release Not Recommended
FTA 6				Release Not Recommended	Release Not Recommended	Release Not Recommended

Figure 2.4: APSA decision matrix [4].

on those offenders; and, 10 to 20 percent of offender cases analysed were found to have had the charges unsubstantiated by the court (which corresponds with the data in NSW, detailed earlier and referenced at Table 2.2). By their own admission on limitations, the authors noted a small sample size, and they had not analysed biases in the data impacting judicial decisions and recidivism. Despite these limitations, the research by [4] underscores the principles of ethics and morals in criminal justice where human assessments present fallibility.

Schematising bail decisions

As was noted earlier, the NSW criminal justice system is burdened with economic stressors, administrative issues, and greater overall incarceration numbers, attributed in part to increases in remandees. It is legislated within the *Bail Act 2013* (NSW) that a decision whether an accused receives or is refused bail is made by a *bail authority* – defined as an authorised justice of a court or a police officer – and that an accused person would need to demonstrate to that bail authority the reasons why their detention is unjustified [96]. Introduced into the NSW Parliament as the *Bail Amendment Bill 2014*, the Second Reading on amended bail laws by the then Attorney-General stated that the accused charged with a "show cause" offence would need to "show cause...that detention is not justified and be subject to the unacceptable risk test before bail can be granted" and that a bail decision would be made on the basis of a "risk assessment" [97], which is termed "Unacceptable risk assessment". The Attorney-General explained further that the matter of "bail concern" is indicative of five items: failing to appear, committing a serious offence, endangering the safety of victims, the community, or interfering with witnesses of evidence [97]. Effectively, the task of a bail authority is predictive: it is a probability assessment based on identified factors to decide if an accused person will reappear before the court at a future time to answer the charges, but also not to commit any further

offences while subject to bail. This was similarly opined by a NSW Supreme Court judge who said of the Show Cause test and Unacceptable risk assessment:

> Both tests also involve, although to a lesser degree for the show cause test, an exercise of the prediction of human behaviour, to which no certainty can ever be attached. Reasonable minds may well differ on the result of a bail application [98].

At Section 16 of the Bail Act 2013, a bail authority must reason on factors concerning the nature of the offence (Show Cause test) and the risk an accused poses, if any, to others (Unacceptable risk assessment) [96]. Along with the written notations, this is depicted in the Bail Act 2013 by schematic flow charts that ultimately leads to a binary decision of either *yes* (bail granted) or *no* (bail refused) [96] (see Appendix D for Section 16 bail flow charts). This was succinctly expressed in [99]:

> [T]he approach of the Court falls into a dichotomy. If there is an unacceptable risk, the Court must refuse bail; if there is no unacceptable risk, the Court must grant bail ... The schema of the Act suggests a two-stage task in which the Court would first call upon the accused person to show cause why his or her detention is "not justified". Subsection 2 of [section] 16A provides that, if the accused person does show cause why his or her detention is not justified, the bail authority must make a bail decision in accordance with Division 2 of Part 3, which is the unacceptable risk test. That test applies to all offences.

This explains what Stage 2 and Flow Chart 2 assessments in the Bail Act 2013 are designed to achieve – a binary or dichotomous outcome – if there is an unacceptable risk apparent, a 'no' decision is given and bail is refused; if an unacceptable risk is not apparent, the decision is 'yes' then bail is granted.

As shown in Figure 2.5, Flow Charts 1 and 2 from the Bail Act 2013 are reconfigured into a single Sankey diagram, and herein referred to as the BAILgram (see Appendix E for the BAILgram pseudocode; template accessed from [100]). The BAILgram process moves from left to right; the different colours differentiate each stage in the bail assessment; the channel over the top half is indicative of bail granted, while the bottom half is indicative of bail refused. The letter representations of 'FC1' symbolises Flow Chart 1: Show cause requirement; and 'FC2' symbolises Flow Chart 2: Unacceptable risk test. The colour intervals and literal notations signify a point where a decision is to be made in the same as the bail legislation schema.

It is intended that the decision-making process for bail in NSW criminal courts, demonstrated schematically and materially, can be replicated into a computerised model represented as a numerical score or measurable result. This proposition is supported in the literature where an algorithm can be formulated to produce a decision, represented as a numerical score or probability [53], and moreover, contemporary algorithm-based risk assessments would need to apply specific measures or predictors, such as "criminal history" and an event such as a "rearrest" [53]. Although, caution should be taken when using predictor variables such as "past arrests" rather than "past convictions" as it can significantly affect prediction outcomes [101].

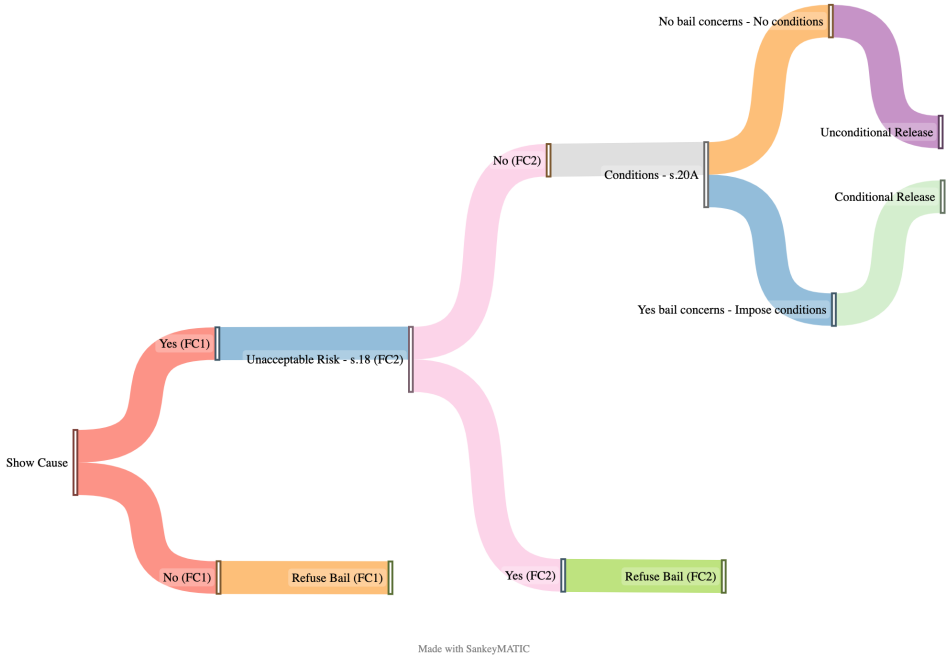

Figure 2.5: *BAILgram* – a reconfiguration of the Bail Act 2013 flow charts. The process moves from left to right; the different colours differentiate each stage in the bail assessment; the channel over the top half is indicative of bail granted, while the bottom half is indicative of bail refused. The letter representations of 'FC1' symbolises Flow Chart 1: Show cause requirement; and 'FC2' symbolises Flow Chart 2: Unacceptable risk test. The colour intervals and literal notations signify a point where a decision is to be made in the same as the bail legislation schema.

In adversarial criminal justice systems such as NSW, arrests and thus bail determinations are not required by law to meet the standard of proof that convictions require, namely, 'beyond a reasonable doubt'. As indicated in the Bail Act 2013 at Section 31 "Rules of evidence do not apply": a bail authority is only required to meet the standard of proof on a 'balance of probabilities' when considering circumstances of a case and accompanying evidence [96]. Accordingly, there is not a requirement to meet the reasonable doubt standard when making a decision, other than determining new and untested criminal matters. While the literature consistently calls for high accuracy in outcomes, a benefit presumably gained from Section 31 in probabilistic terms is that decisions would not have to meet the principle of absolutism, as perhaps might be expected in medical assessments, rather the principle of probabilism is more applicable.

2.2.2 Visual Analytics: a characterisation of it and its facilitation to AI

VA is the overarching discipline understood as the integration of visualised information analysis through computer processing [102]. It was expressed as an interplay between automatic and visual analysis with human input that provides an outcome based on the data [103]. According to [104] this interplay is the distinguishing characteristic of VA over other types of analyses due

to its dynamic and evaluative processes. Where more traditional methods of interpreting data can fail against large datasets, VA does not [103][104] as it merges computational tasks such as machine learning with interactive visual interfaces to assist the user in understanding and facilitating complex dataset analysis [105][106][107]. An example of a creative and interactive VA tool was demonstrated in an online property-news channel that enabled users to choose between suburbs and house prices over calendar years [5]. A benefit of this was it being user-friendly and easy to understand the outcomes.

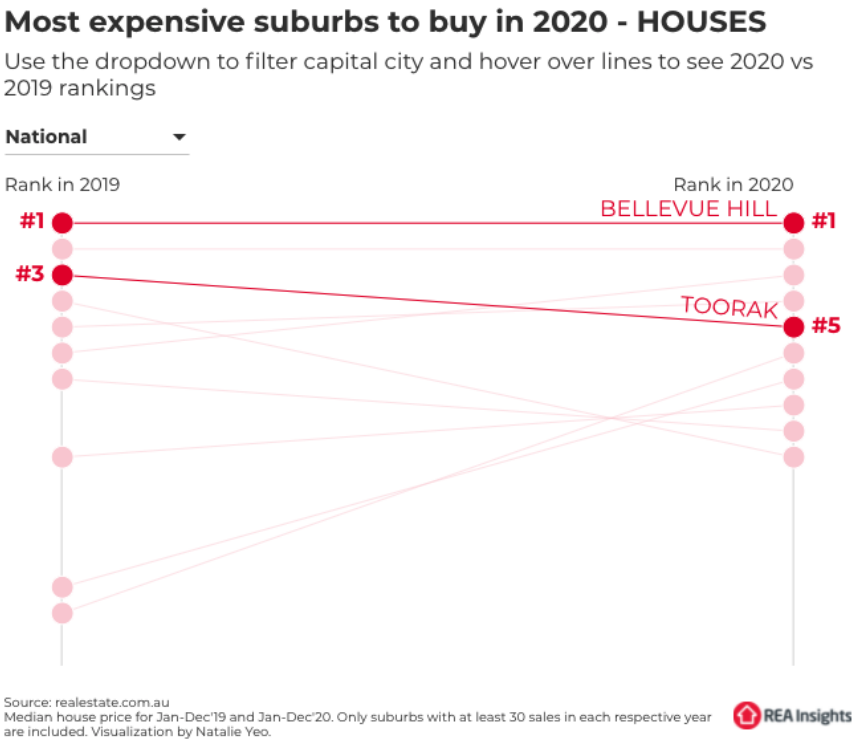

Figure 2.6: Screenshot of an interactive VA tool to compare property [5].

Even though it is not entirely evident from the screenshot in Figure 2.6 for its interactivity, there were several options available within it to obtain data on property and make comparisons to different suburbs, towns and cities throughout Australia by year and property value.

The conventional methods for statistical analysis, such as linear regression, are much improved by advancements in calculation supported by machines and automation, and what has benefited the interpretation of results or the calculated data, has been the emergence of *visualisation* [64], a sub-discipline of VA. It can be further specified in two branches, one *scientific* and the other *information* [103], the latter being of interest here. Information visualisation is focused on such things as business data and technical data, and due to the immense datasets produced from these data-types, they are not favourable to more standard visuals [103]. The benefit of visualisation through machine-based analysis is evident in more practical applications. For example, [6] applied a ML system with a linear regression model to predict future crime in South Africa, and they displayed their results utilising several visual analytical presentations through the Python Libraries application – one of those applied was WordCloud, which distinguishes the crime categories of four provinces in South Africa: (a) Gauteng; (b) KwaZulu-Natal; (c) Western Cape; (d) Eastern Cape (see Figure 2.7).

Figure 2.7: VA approach WordCloud used to display crime categories on four provinces in South Africa ((a) Gauteng; (b) KwaZulu-Natal; (c) Western Cape; (d) Eastern Cape) [6].

In their model, the researchers featured 27 crime categories across nine provinces from existing data that included population, area, and density. The authors reported that linear regression is an effective predictive model for crime occurrence; and ML was effective in drawing out valuable data patterns that led to effective visualisation of prediction that supported prevention. The study by these researchers demonstrated that different predictive models can be used effectively in data collection and visual analysis.

Visualisation approaches

As noted previously, an example of a visualisation approach was applied in a criminal justice study. There are many other visual analytics applications, and while there are too many to be covered here, what has been considered are those that the literature has demonstrated as being effective in the criminal justice discipline. At the outset, it is constructive to note four ML techniques from the literature that are implicit in VA applications. Endert et al. [10] listed *dimension reduction*, *clustering*, *classification* and *regression/correlation analysis* techniques, although only three are of consequence to this research, described summarily in the following.

The utility of dimension reduction is its ability to identify key attributes when using "conventional visualisation methods" to limit the complexities in the data, and enable the user to have process modulation [10]. Figure 2.8 provides an example of where dimension reduction is used. Classification supports the interactive nature between user and a working algorithm that anticipates and identifies incorrect input, where feedback and updates are provided while the machine is learning, which is effective in decision-making for the nature is dynamic [10]. The effectiveness of regression/correlation analysis is based upon VA approaches exhibiting promising outcomes for causation and observation while identifying multivariate associations in data variables [10].

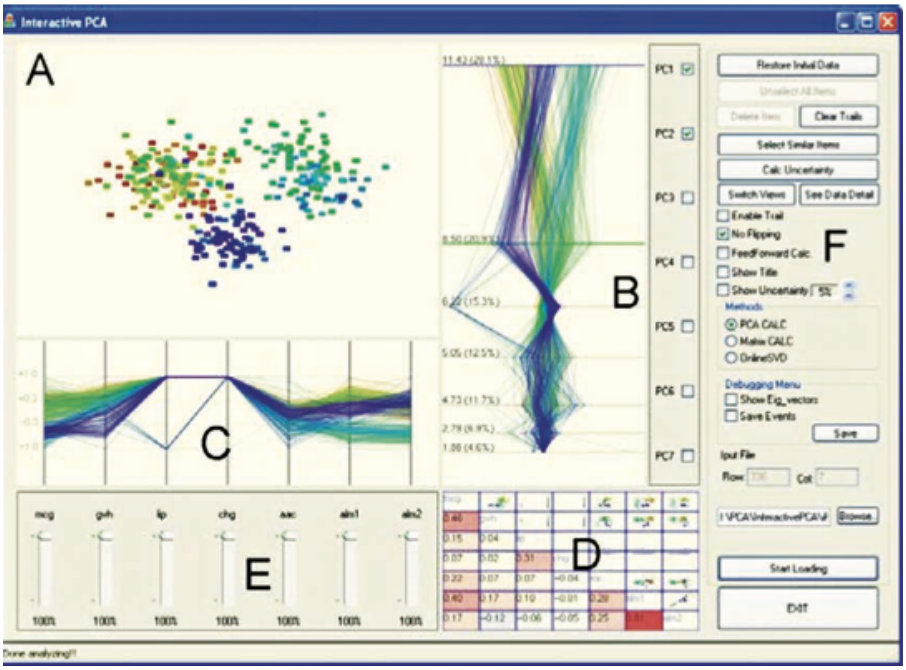

Figure 2.8: Screenshot of dimension reduction in an interactive visualisation tool where visual model parameters can be adjusted (E) and (F) to produce visualised analysis (A, B, C and D) [7].

Given this brief outline on techniques, it is necessary to consider the literature on VA approaches. Sun et al. [104] in their survey categorised five approaches: *space and time*; *multivariate*, *text*, *graph and network*, and *other applications.* The applications of relevance here are detailed in the following. Multivariate data methods involve the computational process of recorded data calculated by algorithmic equation to determine a given domains comparative value by way of analytical reasoning and visualisation, for example, scatterplot matrices [104] [108]. Text data includes documentation, which can be voluminous and unstructured [104] and essentially its qualitative nature when needing analysis requires an efficient process, so once the data is converted quantitatively, it can be interpreted visually. Text data was divided into two branches of Topic-based methods and Feature-based methods. As its name suggests, Topic-based is where information on selected topics are mined from text corpora and analysed through visualisation, for example, Natural Language Processing (NLP) [104]. Feature-based methods identify more frequently observed keywords within a text and visualises them in larger font as opposed to less frequently identified words in smaller font, for example, *word clouds* [104](see Figure 2.7). The authors referred to 'other' which is ambiguous although noted several methods, one being semantic interaction, seen as the synchronisation and adaptation between the user and a model, where a user's analytical reasoning forms a visualisation through their interactivity. As similarly iterated with dimension reduction, the user modulates the process through interactions [10]. This method is beneficial in the application of high-dimensional data being visualised in a two-dimensional view [104].

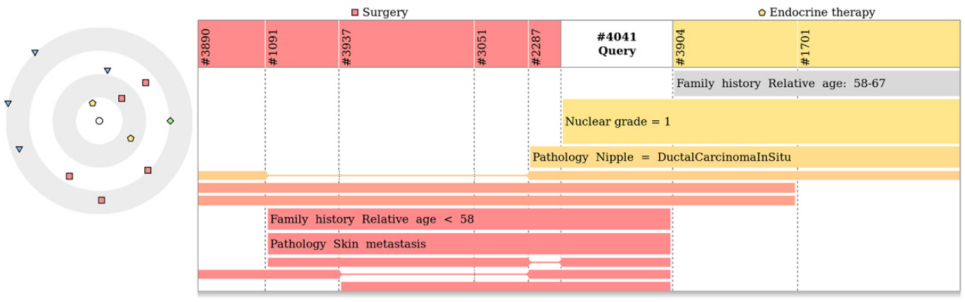

Figure 2.9: Screenshot of Case-Based Reasoning prototype using scatterplots and rainbow boxes [109].

In addition to the approaches mentioned, an xAI approach termed Case-Based Reasoning (CBR), was presented in research by [109]. CBR uses historical data from past cases that are similar; from those past cases that have an established outcome, a newer outcome can then be established [109]. The outcomes or decisions from that data can then be visualised for analysis [109][110]. In the paper by [109] two visualised formats of scatterplots and rainbow boxes were applied (see Figure 2.9). The authors indicated their preference for CBR was due to its explainability, and also, for rainbow boxes due to its capacity to present qualitative information [109].

Analytical reasoning through visualisation

Another sub-discipline of VA, *analytical reasoning*, facilitates the user to apply relevant visualisation approaches to support "interactive visual interfaces" [102]. The aim of analytical reasoning using visualisation is to provide the human-user with clarity and interpretability in an algorithmic model. This proposition is supported in the literature where it was asserted that machine-based decisions can rely on visualisations to contextualise features used in a given model [1][62]. As such, the benefit of utilising a machine is to fill the void of humans created in areas like bias and memory [111], although the position on memory was challenged by CCT (as mentioned earlier in this Chapter).

Conversely, there were misgivings in parts of the literature as to the benefits of visualisation when undertaking analytical tasks. Micallef et al.[112] tested if visualisations benefited participants on probabilistic reasoning using Bayesian psychologically based problems: the results reported that visualisations were not overly beneficial in improving accuracy for participants when presented with text, although participants estimated errors were reduced when presenting textual information with a diagram. Another misgiving implied more complex algorithmic models such as random forests, while accurate, lacked in the provision of explaining "how input variables are producing a predicted outcome" [82]. Given the differing perspectives, it is reasonable for a middle ground to be found, which balances human with machine interactions.

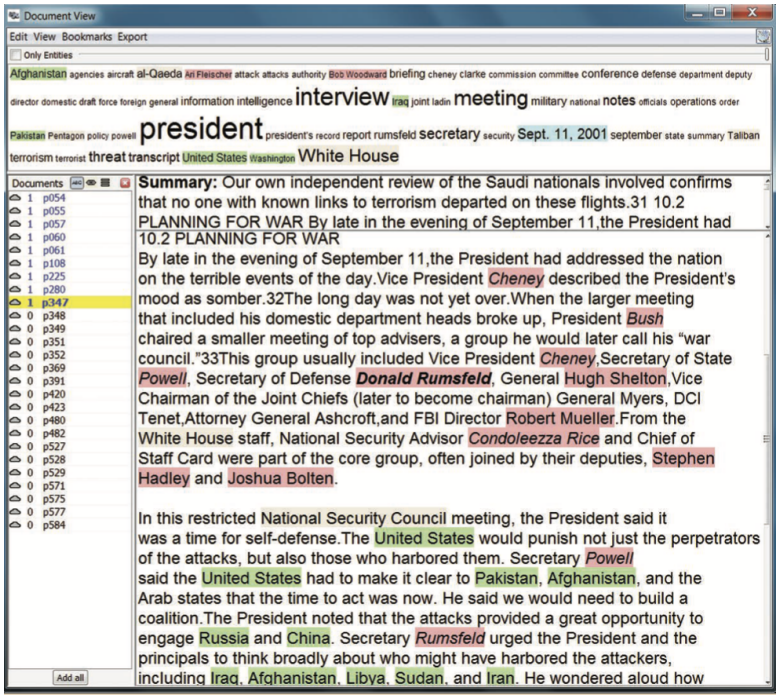

Figure 2.10: Screenshot of JIGSAW that shows all documents relating to the individual in question along with other related subjects [8].

Earlier in this chapter, the Bottom-up approach on machine ethics Reinforcement Learning, endorsed by [49], was said to allow a user the function to regulate a goal-focused process based on rules and principles, which ultimately completes a task. This approach was also preferred by [10] and [113], as the "human in the loop" approach instills greater trust in the results, and the training between human and machine creates a reciprocity and synergy. Where user trust had wavered in visual analytical systems, [107] suggested a "mixed-initiative interface" could counter these doubts as it facilitates direction in the system for a user. Similarly, [111] preferred this approach referring to it as "mixed initiative systems", which are underpinned by semantic interaction. A study on predictive visual analytical tasks reported research participants preferred easily interpretable visual analytical tools over the more complex ones, although it was indicated that easy to use did not necessarily mean that more difficult computer-interface tools were not used [114].

The literature provided a seemingly user-friendly analytical option for analytical tasks. In intelligence analysis in security and criminal contexts, information is likely to be in a text or language format, making analyses more difficult; however [8] claimed to have provided a prototype that is able to support simpler analysis of large textual documentation. The authors said their concept JIGSAW merges text analysis and interactive visualisations, deciphering documents identifying text regarding individuals, locations and entities of interest intended to provide intelligence analysts with a holistic overview of these connections to ascertain circumstances of consequence.

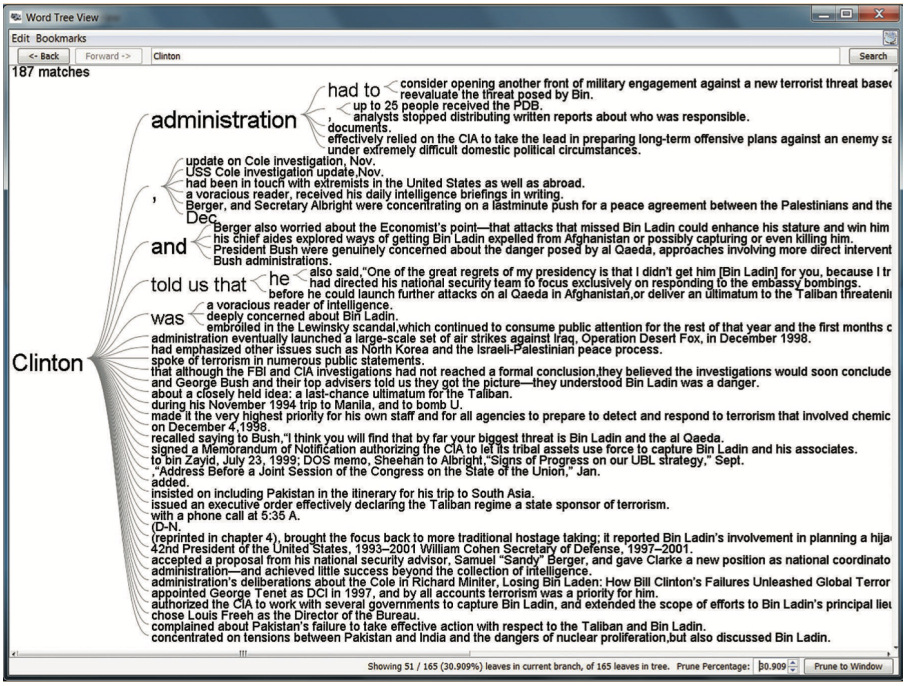

Figure 2.11: Screenshot of a word-tree from JIGSAW demonstrating the selected word and the most common phrases that follow [8].

Figures 2.10 and 2.11 are screenshots of JIGSAW displaying content specific to the subject matter being searched – in this example the participants were attempting to ascertain threats associated with a terrorism event. The authors tested the prototype using a 585-page intelligence report with tertiary students as participants. The participants were divided into four groups, three being control groups and one having complete access to the JIGSAW program. The authors claimed their results showed the group with complete access produced more favourable outcomes than the three control groups with limited access. A self-disclosed limitation of their research was that it needed a large participant size.

Notwithstanding, the JIGSAW concept provides a positive example on document analysis in a crime/justice domain, but despite being algorithmically structured, it provides visualised elements that accommodate ethical matters.

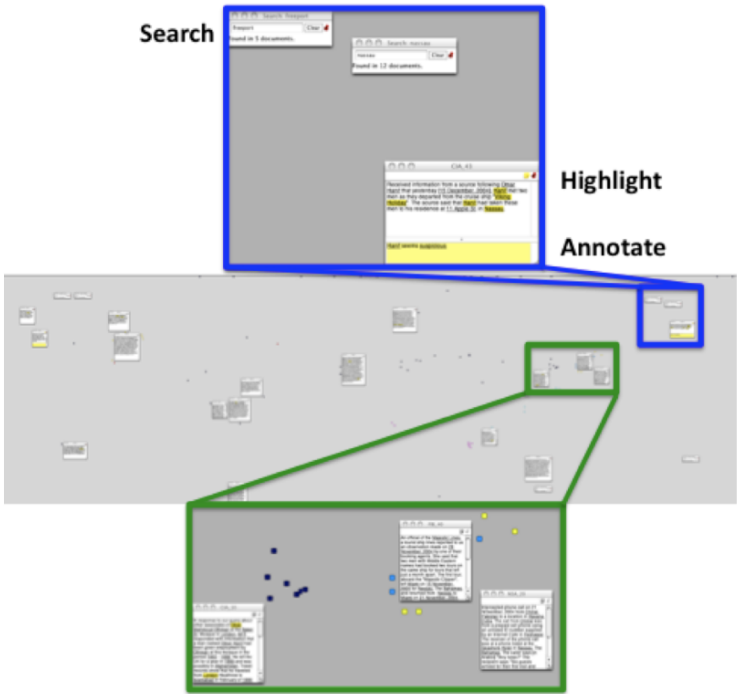

Figure 2.12: Broadened interface screenshot of ForceSPIRE demonstrating spatialisation and retrieval of documents using search, highlight and annotation [7].

Within the literature, two contrasting views on mitigating ethical considerations were made. One suggested it is achievable through a human-machine collaboration; the other claimed it is unreasonable to assume that mitigation against ethical considerations is absolute, rather it is a proposition of balance and making concessions [56][105][115].

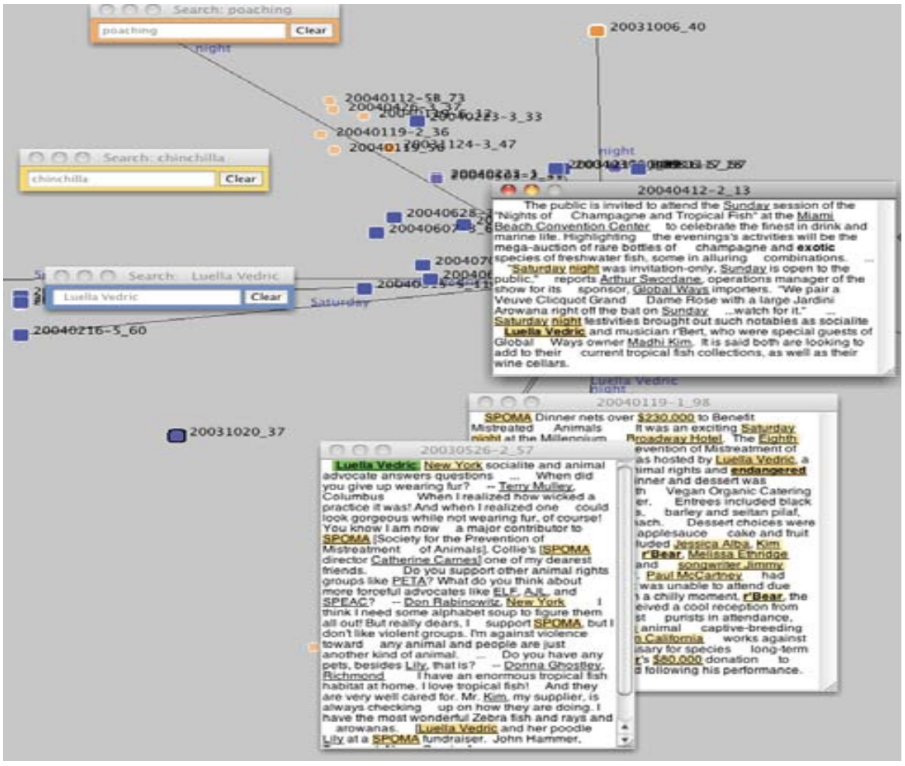

Figure 2.13: Screenshot of the StarSPIRE prototype interface visual encoding of document relevance and term importance[9].

Several iterations of the human/machine interactive models appeared to placate ethical development through VA. *ForceSPIRE* [10] and updated versions *StarSPIRE* [9] and *IN-SPIRE* [10] are machine learning prototypes that operate by synchronising data and visualisation with human interaction (see Figures 2.12, 2.13 and 2.14).

The approaches of direct manipulation, semantic interaction and foraging are applied, permitting a user to control the direction of the program manually instead of total automation [7] yet is guarded against algorithm aversion [9][116]. The authors claimed their prototypes could provide the user or analyst with greater clarity on complex data sets, particularly distinguishing similar and differing items [7]. The authors tested both models in their respective release years to demonstrate its effectiveness in analysing intelligence reports, and reported that these prototypes were effective in the said approaches where processing and reasoning of information was synthesised. The negatives of both studies were the participant sizes and the study generalisation; however, a hallmark of this study is framed in the utility of these prototypes for real-world application, as in security and intelligence analyses referred to by the authors.

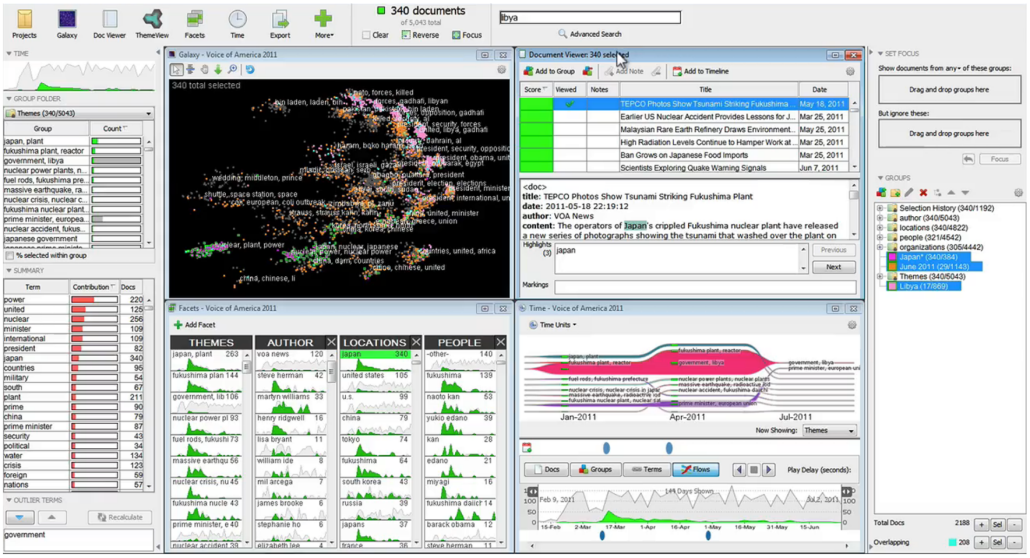

Figure 2.14: Screenshot of the IN-SPIRE prototype interface showing multiple insets [10].

A second prototype was identified in the literature that bears some similarities to those just described, although the authors had more specific intentions for their prototype to support social science researchers using machine learning approaches. Soroko et al.[11] stated that "computational social science" is relying more so on algorithmic based research and placed importance on it being transparent, accurate and explainable. The prototype *isift* is said to simplify locating relevant phrasing in legal texts using "low-dimensional real-valued vectors that represent the syntactic and semantic meaning of words" using two-dimensional visualisation [11]. The software design is purposed on being user-friendly for those who do not have expertise in machine learning and algorithms.

The screenshot at Figure 2.15 displays various visualisations of sequences within the *isift* program: (1) Document Map (sentences from legal documents in order); (2) reference sentence chosen by the user; (3) selected sentence chosen by the user; (4) similarity between 2 and 3 determined by 'similar' or 'not similar'; (5) scatterplot of sentence vector similarity; (6)

Graph demonstrates sentence and wording similarities and editing; (7) reclassifies sentences without word changes; (8) key displaying contrasts of 'not similar' to 'similar' and symbols [11]. While the prototype does not allow for model adjustments, it is interactive allowing for user control. Given there are more technical elements to this prototype not mentioned here, it does represent promise in mitigating ECs when applied to the criminal justice domain, particularly in identifying key phrases in voluminous documents.

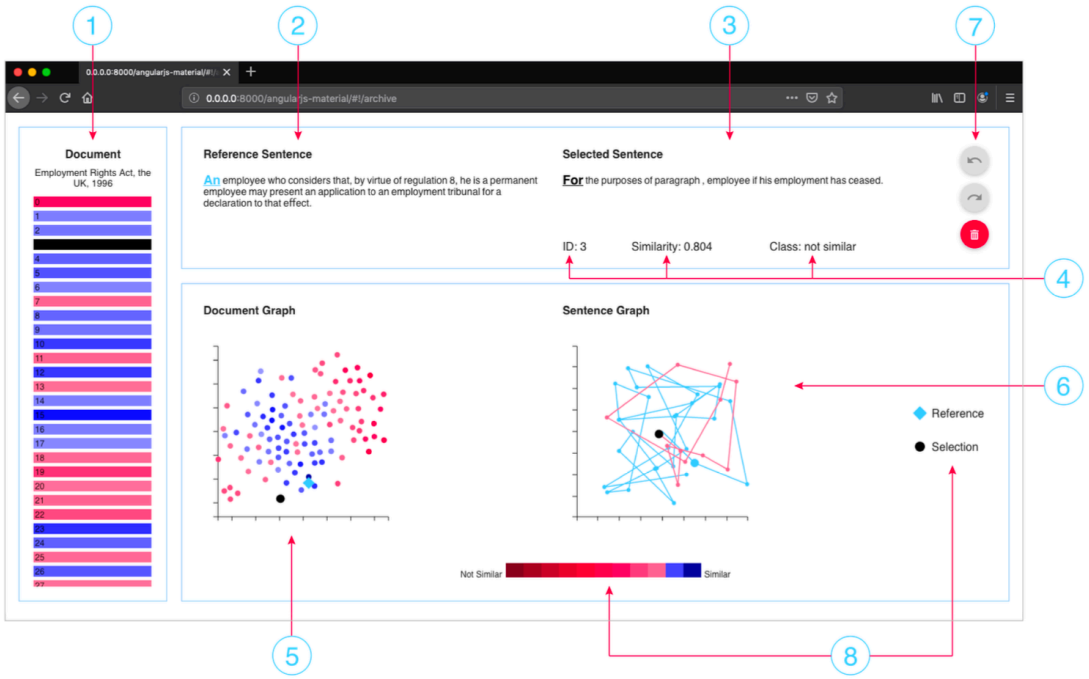

Figure 2.15: Screenshot of the *isift* prototype interface. (1) Document Map; (2) Reference Sentence; (3) Selected Sentence; (4) Sentence information for ID, similarity score, Class (similar/not similar); (5) Document Graph/Scatterplot; (6) Sentence Graph; (7) Editing buttons; (8) Key/Legend [11].

Another model found in the literature also appeals to ethical mitigation using VA. Steed et al. Steed et al. [12] designed a visual data mining software program called the Multidimensional Data eXplorer (MDX), which they described as applying statistical analysis and data classification techniques on complex datasets. While not a traditional research survey, the authors and developers of MDX demonstrated how it works on a large dataset and what visualisations can result from their framework. The significance of MDX was how it demonstrates and explains some of the ethical challenges seen in data analysis. This was demonstrated in their use of a multicollinearity filter in a regression model algorithm, which is firstly interpretable, and secondly permitted predictors to be added or removed depending on relevance (see Figure 2.16). The authors proclaimed that "[r]educing the number of predictors is helpful to exploratory analysis because it simplifies interpretation" [12].

Another model from the literature was creative in its approach to analytical reasoning and visualisation. The VA approach is CBR, and as discussed earlier, it is a form of xAI. The authors favoured this approach as it can explain the results that are both qualitative and

Input: Significant correlation threshold, $r_{threshold}$
Input: Array of Axis objects, $axes$
Input: Single dependent axis, $axis_{dependent}$
Output: Truly independent set of axes in display

```
    // Descending sort by r of each axis with axis_dependent
 1: axes_sorted ← SORT(axes)
 2: for Axis object axis ∈ axes_sorted do
 3:   for Axis object axis_compare ∈ axes do
 4:     if axis_compare = axis_dependent then
 5:       continue
 6:     else if axis_compare = axis then
 7:       continue
 8:     else
 9:       r ← CORRELATION(axis, axis_compare)
10:       if r > r_threshold then
11:         remove axis_compare from display
12:       end if
13:     end if
14:   end for
15: end for
```

Figure 2.16: Screenshot of the MDX prototype pseudocode [12].

quantitative, despite the use of algorithms. The algorithmic model by [109] was devised then tested for medical purposes, classifying patient therapy for breast cancer (see Figure 2.9).

According to the authors, the clarity demonstrated by this visualisation was equated to the visual interface that consists of three elements: (1) the scatterplot indicates the similarly procured cases; (2) the scatterplot and rainbow header boxes indicate the similar case amounts and levels; (3) the boxes show the alignment of similar cases to new cases, and delineates and orientates characteristics according to value classification. The authors measured the effectiveness of their model in three ways. Firstly, based on three publicly available datasets and four treatment options for breast cancer intervention – surgery, chemotherapy, radiotherapy, and endocrine therapy – the most suitable options were demonstrated through a visual analysis approach (see Figure 2.9). The second measure the authors assessed was model accuracy: applying three algorithms, they declared had provided good classification (although a limitation that the authors self-identified related to a small subset). The third measure was to determine the opinion of medical experts who might use the model. The feedback was suggestive of interface complexities, and subsequently, it was revised and re-tested using another medical cohort. The feedback was more positive on the CBR using visualisation, although the inconsistencies between the scatterplots and rainbow boxes was unfavourable. Notwithstanding, the VA approach by these authors provides an effective template that could be applied in other domains, most relevantly, criminal justice.

2.3 Summary

> The active pursuits of man result also from the exigencies of human society, or its need of establishments, to restrain disorders, and to procure the benefits of which it is susceptible [71].

This statement by a western philosopher in the eighteenth century is far removed from the current period, yet it has relevance in the context of ethics and morals for criminal justice – the procurement of a mechanism to benefit a susceptibility in human fallibility – an endorsement of AI supporting conventional decision-making.

The judge's statement in the Loomis case concerning fellow judiciary members having a choice to 'consider' or 'rely' on a decision being machine-driven was pertinent to the discourse upon whether AI should be used to support or supersede human-driven decision-makers in some legal matters. The coexistence of human and machine – the human-in-the-loop as it were – appears to provide some semblance of an organic and natural integration as AI permeates into real-world environments. That being so, assume momentarily the real-world environments of criminal courts assessing bail applications to decide whether an individual should be detained or released: generally, in larger judicial precincts, there are numerous court rooms, many accused, copious documents, and limited time to determine decisions of consequence. However, with the benefit of AI-generated decisions, a magistrate or judge (or a court) could rely on a program with the function to predict the suitability of multiple bail applications and filter the relevant information. The functionality of these concepts was evident in the literature, and some of these prototypes' simplicity balanced machine and human interactions with visualisations, and it was interpretable and transparent, among other ECs, thus establishing a viable mitigation approach to decision-making using AI.

While the literature portrayed the viability of ethical decision-making, there is content challenging its viability based on validity (effectiveness) and reliability (trust). A significant concern is founded on human ethics and decisions on an individual's liberty – empowering computers to be decision-makers to grant or refuse bail. Another concern is satisfying ideological thinking, such as conservative or non-progressives, that machines can be effective and trustworthy, and the benefits of AI outweigh any negatives. A pivotal challenge drawn from the literature is the facilitation or balancing of the ECs equally – one or more ECs might be affected or sacrificed for others, also referred to in the literature as 'trade-offs' – which consequently may result in a disparity.

The NSW criminal justice system's deficiencies in bail administration were identified from the literature, and how it could be reshaped and alleviated through the application of computerised decision-making. The implementation of an AI-driven program would be the overarching responsibility of the NSW government, and based on their current AI policy, it is a safe assertion any implementation would inherently contain the ECs described in this Chapter. The research on the U.S. criminal justice system using AI-driven decision-making was indicative of apprehensiveness from the public, particularly concerning ethics, yet there was a strong contention in the literature that the ethical matters can be mitigated, which invariably could be achieved

through mixed-initiative user interfaces using VA, and applying simple methodology such as limiting the amount of variables. Advocates on algorithm decision-making in the Australian criminal justice system asserted that if such a revolutionary concept was to be considered, a six-month pilot study should be undertaken in two or more jurisdictions, conducted concurrently with active matters, which also engaged both the legal profession and the public. If the supporting literature is to be relied upon, it is possible for the schematic bail model, or the BAILgram to be replicated into a realistic decision-making prototype – the proposal here is that a pilot study can be conducted on bail decisions using ML and VA. Without overreaching, it is reasonable to contemplate any success in a pilot study and prototype development could result in later iterations being adopted by bail authorities in NSW.

In light of this prototype proposal or any other, it is apt to conclude by reiterating a position expressed by [15] on algorithm decision-making: there is an inevitability for decisions to be contested by defendants, and such challenges will require human engagement from the judiciary, consequently, the function of the criminal courts would not be invalidated by technological developments. That being said, the coexistence of the two systems working to support or supplement one another, is both pragmatic and realistic.

Alluded to in this Chapter, regarding the implementation of AI-generated decision-making in a criminal justice framework, is for its objectives to contain improvements in efficiencies and accuracy. As important as that is, consideration must be given to those who would be impacted by the technological developments: at the macro-level, the public, in the provision of economic benefits; and at a micro-level, the stakeholders, by way of fair and impartial decisions. Furthermore, that with fairness, decisions are explainable, and are factored into such technological developments and their implementation.

It is at this juncture to underscore what this thesis intends to contribute by bridging a gap, namely: the amalgam of predictive modelling to assess actual bail decisions with a revised categorisation of nine predictors based on the Bail Act 2013; and, to survey what the stakeholders perceive of analytical reasoning aided by VA to decide bail for an accused.

2.4 Collection

This ancillary section details the search processes undertaken for the literature collections relied upon in this thesis. A synopsis of the research-based material can be found in Appendix C.

Mapping the process for search parameters or key terms initially began with various wording combinations with the view of developing a more concentrated and defined representation of AI in the criminal justice system. The intention was to expand from the fundamentals to a broad scope – the parameters and terms included, but are not limited to: “artificial intelligence”, “algorithms”, “data mining”, “risk assessment”, “prediction”, “visual analytics”, “analytical reasoning”, “fairness”, “bias”, “ethics”, “crime” and “law”. The generalisation of these terms then became more concentrated, for example, “artificial intelligence and law”, and “artificial intelligence and ethics” as the relevance of these subject areas developed. For the purposes of surveying and structuring the literature to allow for relevance and synchronicity, the following

categories prompted sections and subsections for this Chapter, but have been varied during the mapping process: Artificial Intelligence, Algorithms and Machine Learning; Prediction and Risk Assessment in Crime and Law; Philosophy, Ethics and Theory; and, Visual Analytics and Analytical Reasoning.

The sources were predominantly scholarly books and their accompanying chapters, peer-reviewed journals, and academic dissertations. The more common databases where these sources were published under are ProQuest, SpringerLink, HeinOnline, Arxiv, ScienceDirect, ACM Digital Library, and IEEE Xplore. These databases provide linking prompts to similar literary subjects, and this option was utilised based on title relevance and/or familiarity of cited authors. While searching through these academic journal articles, and similarly for books and chapters, the abstracts, introductions and summary of chapters were initially read to determine relevance. A significant benefit was gained from the use of bibliography/reference lists from the selected sources after identifying the relevant material. Online reports from tertiary think-tanks and government departments were accessed, as well as legislation and caselaw from NSW government websites. Several survey papers were made available from the principal supervisor.

The material selected was dependent upon the published date: any literature before 2000 was not reviewed mainly on the basis of wanting the most recently published material, and also, being mindful to the nascency of AI in the criminal justice discipline. Google Scholar was relied upon for some VA based sources in the attempt for the most current literature and published research.

Chapter 3

Preliminary Study I: Groundwork on the Bail Amendments Effect

3.1 Background and Motivation

The *Bail Act 2013* (NSW) saw two critical amendments made in in 2014 and in 2015 regarding the detention or remand of defendants under the "Show Cause" test and terrorism related powers, respectively [97][117]. The perceived intention of these amendments was to temper bail laws, yet contrary to this perceived intention, a higher threshold measure was introduced compelling defendants to provide justification for conditional release. It was asking bail authorities to risk-assess and forecast recidivism under stricter guidelines, and presumably, make it more difficult for a defendant to be granted bail therefore adding to the burden of an already strained justice system.

Remarks from one of three appellate court judges in a 2015 bail judgment provided an extensive commentary on the relatively new bail legislation and its impact on the justice system, as quoted in the following:

> By allowing de novo reviews the Bail Act 2013 facilitates the making of more interlocutory applications which serves to fragment the criminal process, potentially delaying trials. It does so in a manner that diverts this court from its task of hearing appeals from convictions and sentences. It is also a process that generally advantages the Director [of Public Prosecutions] in that in most cases he has superior resources than that available to an accused person. Thus the Director [of Public Prosecutions] has a greater capacity to litigate and re-litigate the issue of whether bail should be granted [118].

Intimated in this judgment was the bail legislation permitted a nuanced legal recourse for those challenging bail decisions made in lower courts through the appeals process to higher courts, consequently, impacting on the functions of the courts in a broader sense, and potentially more so when made by the prosecution. Extrapolating this one case out to an unbounded amount, it is reasonable to argue for bail decisions to made through artificially intelligible means to at minimum, increase efficiency. This will be explored in the proceeding chapters on decisions made

from the new amended bail laws, propositioning that automated decision-making has greater benefits for the justice system and for the population collectively. In preparation however, this preliminary study first seeks to draw on the intimation in the judgment [118], and explore if the bail legislation had fragmented the NSW justice system, based on the underlying presumption that court delays (equated with increased remand times) are a by-product of the Bail Act 2013.

Briefly, to draw on the economic perspective as to why remanding defendants is of consideration, in 2015 the cost ratio of one offender in the community was 10 percent to that of an offender in custody [35]. While not calculated in the report by [35], it does mention a remand drawback where unsentenced offenders become more costly than offenders who are sentenced. A juxtaposition of court delays to remanded offenders is a presumption of a greater burden on the justice system.

3.2 Statistical Methods

Data and Evaluation Metrics

This preliminary study will infer that the 2013 bail legislation and its subsequent amendments affected court delays, consequently increasing remand numbers. The data to test this hypothesis was gained from open-source publicly available information, which is a combination of the state's police and justice department figures [13]. Data subsets were labelled under "Criminal Court Statistics" and "Open Data: Incident by NSW" [13]. Time range captured is over a ten-year period from 2011 to 2021, allowing for two years before the new legislation commenced. Listed below are the selected categories (variables) denoted by upper-case letter, and tabulated in Table 3.1:

- (Variable A) number of incidences of Robbery with a firearm (yearly average);
- (Variable B) number of incidences of Sexual assault (yearly average)[1];
- (Variable C) number of defendants who Breached bail condition(s) (yearly average).
- (Variable D) number of defendants who Failed to appear (yearly average);
- (Variable E) number of defendants Refused bail in all offence categories at all court levels;
- (Variable F) number and Delays in all courts (except Children's Court) measured in median days for trial matters to be completed from arrest to finalisation where the accused was on remand (refused bail).

[1] "Sexual assault" data (variable "B"): child sex-offences and other sexual assaults (adults) were not differentiated by [13] as they are not individually classified by the Australian and New Zealand Standard Offence Classification (ANZSOC) [13]. As will be relevant for this research, child-sex offences are Show Cause items.

Table 3.1: Categories labelled A to F represent captured data between 2011–2021 (see [13] for complete datasets)

Year	Variable A	Variable B	Variable C	Variable D	Variable E	Variable F
2011	33	432	3096	93	7310	502
2012	32	437	3529	63	7493	517
2013	27	461	3487	68	8001	578.5
2014	25	480	2704	57	8240	595
2015	14	491	2096	60	10497	658
2016	14	508	3529	62	11368	713
2017	11	569	3487	48	9283	716
2018	13	577	3549	45	9909	716
2019	13	611	3998	48	10036	722
2020	9	625	4313	64	9230	725
2021	7	620	4262	68	9106	764

Variables C and D were selected as these offences feature prominently in the literature as being instrumental in predictive analysis, and also, are directly related to bail status; Variable E was selected for it is a total of all defendants charged in all offence categories having the status of bail refused. Variable F is accused persons with a bail refused status – defendants who proceeded to hearing/trial, yet found not guilty of all charges, or all charges were withdrawn by prosecution, or otherwise disposed of (i.e., transferred to Drug Court or persons deceased) – therefore charges had not resulted in any penalty but defendants were remanded (equivalent to median court administrative days). As will be demonstrated in this section, Variable F becomes the outcome variable.

Regression algorithm models

As was detailed in Section 2.3 and summarised in Table 2.4, regression algorithm models are robust evaluative methods to measure and analyse what effect the explanatory variables have on the outcome variable [73], provide good predictive utility [89] whilst being easy to interpret [93]. Ordinary Least Squares (OLS) method is selected for its statistical properties are highly effective, versatile and a strong measure of R^2 [119][120], which is a measure by percentage on the variation of the dependent to the independent variable in a model, and the closer to '1.0' suggests very little variability and goodness of fit. [121]. OLS assumes the "independent variables are linearly independent...and the error is normally distributed and uncorrelated with the independent variables" [119].

As seen in linear equations 3.1 and 3.2, the parameters denoted as β_0 and β_1 are unknown and the basis for OLS is to estimate the outcome variable denoted as y [120][122]. Linear regression (3.1) and multiple linear regression (3.2), noted as,

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

where y is the outcome variable, β_0 is the y-intercept, x is the explanatory variable, β_1 is the

regression line slope, ϵ is the error [123], and the multiple linear regression equation,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n + \varepsilon \tag{3.2}$$

where y is the outcome variable, x is the explanatory variable, β_0 is the y intercept, β_n are the coefficients, and ε is the error [123][124]. Given there are five unpaired explanatory variables or unpaired groups, ANOVA methods were selected [125]. ANOVA should provide an outcome on the overall fit of the regression models to the data [119], and to test the hypothesis.

Correlation analysis method

Correlation is an effective analysis method to identify multivariate associations in data variables [10] and can also determine what influence an explanatory variable might have on an outcome variable [126][127]. This method will test the existence of a linear relationship between two variables that are independent of one another. The strength of the association between coefficients is determined numerically, as such, a stronger relationship would see values closer to +1.0 or –1.0, whereas values closer to 0.0 mean a weaker relationship [126]. A one-tailed test is applied to ascertain the relationship between an explanatory variable (x) and the outcome variable (y) [127]. It is calculated using *Pearson r* and this result stated as statistical significance is measured against $\alpha <= 0.05$, suggesting there is a five percent or less probability of the result being obtained by chance [126].

Statistical Analysis Resources used to facilitate data evaluation

Evaluation is facilitated by the Data Analysis Toolpak [119] and Real Statistics Resource Pack [128].

3.3 Results

Statistical trend by category

Figure 3.1 illustrates the yearly averages of each respective category (variable). As listed in Table 3.1, the two most notable observations are in Variable C (grey trend-line) and Variable E (blue trend-line), showing an inverse proportion approximate to the year the first bail amendment was legislated (marked by a vertical pink dotted line) until 2017 where it stabilises, although it did not return to its previous level in 2011.

Figure 3.2 indicates the change as a percentage in all categories between 2011 and 2021. Incidences or offences under Variables A and B decreased by 77% and increased by 44%, respectively. Defendants who breached bail increased by 38%; although defendants failing to appear decreased by 27%. An increase by 25% is observed on defendants refused bail in all offence categories. The number of median days that court matters were delayed increased by 52%.

Correlation

Displayed in Table 3.2 are the results from a one-tailed correlation from 2014 to 2021 revealing all explanatory variables are positively correlated with outcome variable F, with the exception

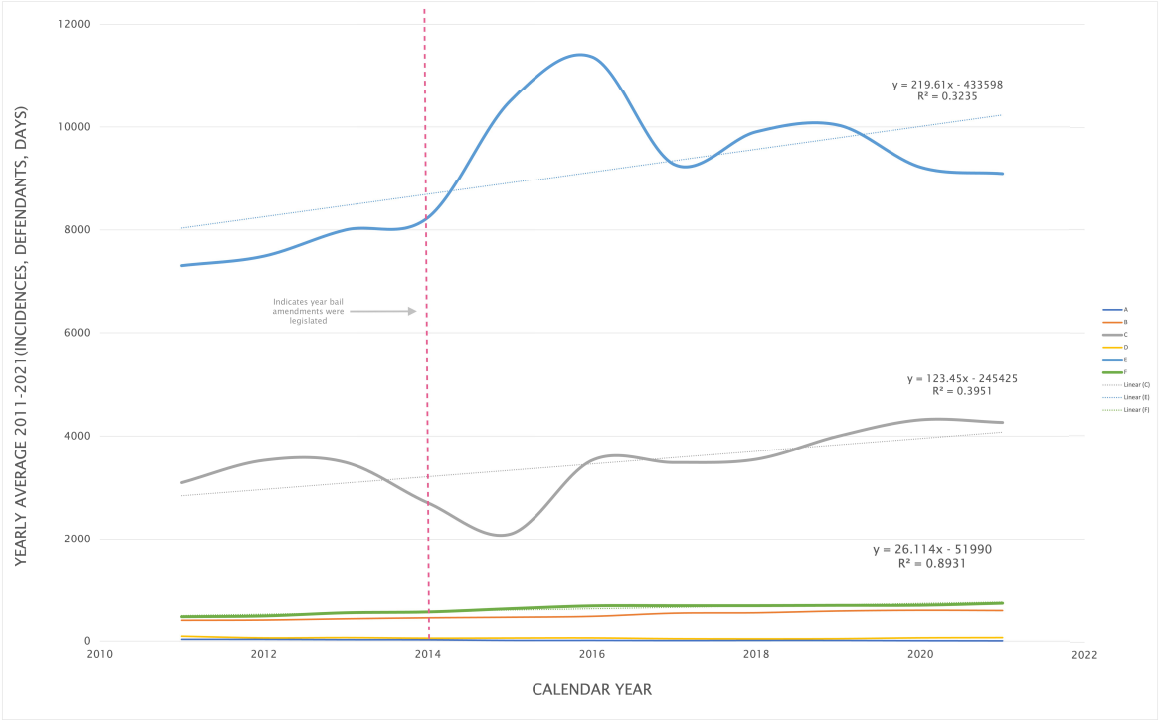

Figure 3.1: Trends calculated by yearly average of each Variable, noting the grey trend-line (Variable C) and blue trend-line (Variable E). Pink dotted line denotes the year the first enactment of two bail amendments.

of Variable A. Variable B (r = .821) and Variable C (r = .809) are indicative of a strong relationship. A significant relationship is also apparent for Variable A (p=.000), Variable B (p=.006), and Variable C (p=.007) to the outcome variable.

Table 3.2: One-tailed correlation (μ) Explanatory Variables A to E and Outcome Variable F based on data captured between 2014 to 2021.

	r	p
Variable A	–.932	.000
Variable B	.821	.006
Variable C	.809	.007
Variable D	.094	.413
Variable E	.276	.254

Revealed in Table 3.3 are similar results with respect to positive and negative correlation, with the exception of Variable D (r = –.585), having shown a negative relationship to the outcome variable between 2011 to 2021. All variables, apart from Variable C, have shown statistical significance to the outcome variable.

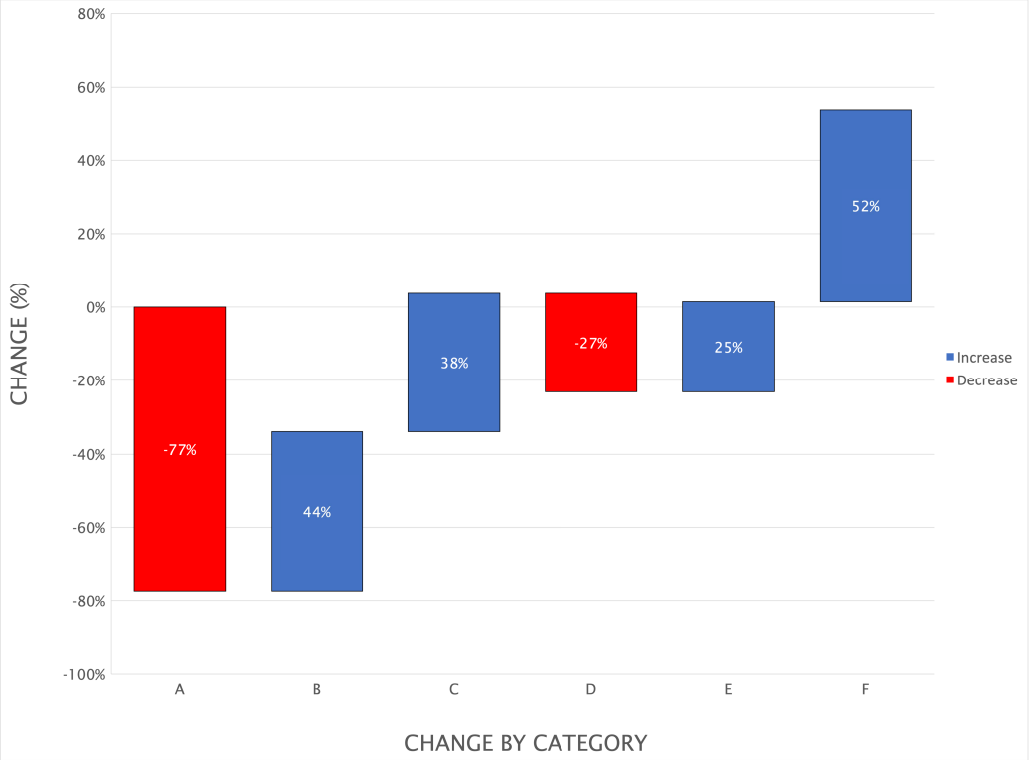

Figure 3.2: Change in percentage of all variables A-F. Data extracted from [13].

Table 3.3: One-tailed correlation (r) of Explanatory Variables A to E and Outcome Variable F based on data captured between 2011 to 2021.

	r	p
Variable A	–.984	.000
Variable B	.910	.000
Variable C	.469	.073
Variable D	–.585	.029
Variable E	.767	.002

Regression

A one-way ANOVA was applied to ascertain if any explanatory variables were observed to have statistical significance to the outcome variable. As can be seen in Table 3.4, none of the variables were statistically significant; and in Table 3.5, significance was only established in Variable A (p=.025, SE 2.246).

Table 3.4: One-way ANOVA of Explanatory Variables A to E and Outcome Variable F based on data captured between 2014 to 2021 ($\alpha <= 0.05$).

	p	SE
Variable A	.117	2.814
Variable B	.607	.406
Variable C	.195	.020
Variable D	.445	.895
Variable E	.637	.009

Regarding the bail legislation and its subsequent amendments, two regressions were conducted. Table 3.6 shows the results from a multiple regression analysis conducted on eight observations from 2014 to 2021. Similarly, Table 3.7 displays the results of a regression analysis conducted on eleven observations from 2011 to 2021 for the purposes of identifying any variability in the models [129].

Table 3.5: One-way ANOVA (p) ($\alpha <= 0.05$) of Explanatory Variables A to E and Outcome Variable F based on data captured between 2011 to 2021.

	p	SE
Variable A	.025	2.246
Variable B	.789	.280
Variable C	.214	.013
Variable D	.489	.489
Variable E	.240	.008

It is indicative from the R^2 values that there is very little variability in the models, the data is a good fit to the model, and there is some relationship between the explanatory variables and the outcome variable.

Table 3.6: Multiple regression values based on eight observations from 2014 to 2021.

Multiple R	.987
R^2	.975
Adjusted R^2	.913
SE	15.256
Observations	8

Figures 3.3 and 3.4 are line fit plots on eight observations between 2014 and 2021 representing the predicted outcomes, and it is apparent that the actual outcomes do not follow the predicted projections. A normal distribution of the data is plotted in Figure 3.5 (2011 to 2021) and Figure 3.6 (2014 to 2021), indicative of a positive correlation.

Table 3.7: Multiple regression values based on eleven observation from 2011 to 2021.

Multiple R	.993
R^2	.986
Adjusted R^2	.973
SE	15.125
Observations	11

3.4 Discussion

When the bail laws were amended in 2014 and 2015, a stricter model was implemented for a defendant to prove why their detention is not justified. Consequently, decision-makers had

the unenviable task to conduct risk assessments and future predictions on whether a defendant would comply with bail. Those refused bail were remanded in custody for an indeterminate amount of time, which as hypothesised, led to court delays. An exploratory analysis of this showed an upward trend in the bail refused category (Variable E) and a downward trend for breach of bail (Variable C) directly after the new bail laws came into effect. The trend was apparent for several years after and did not recede, although it did not follow the predicted projection, as shown in Figure 3.1.

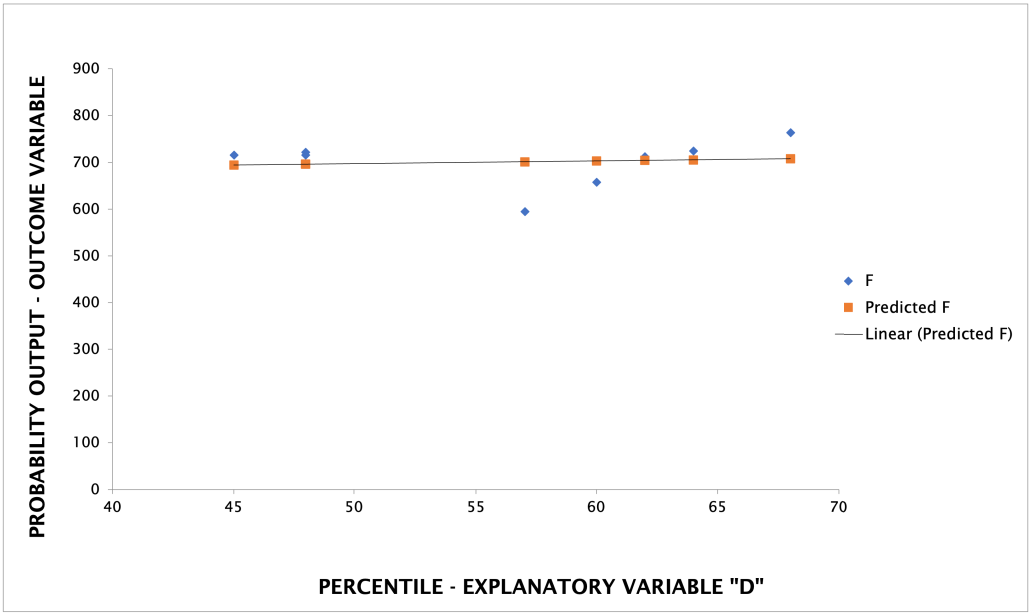

Figure 3.3: Line Fit Plot – Explanatory Variable D.

Variable E was the category expected to be highly correlated with the outcome variable (Variable F) as it was reasonably assumed that defendants being refused bail is analogous to court delays. Statistically, a correlation was evident from the 11 observations ($r = .767$) however this did not occur when the observations were decreased to eight ($r = .276$). It was expected that robbery with a firearm (Variable A) would be negatively correlated with court delays (Outcome Variable F) as a decline in the incidences was observed over the ten-year period to 2021. It was also expected for the sexual assault category (Variable B) to be positively correlated with court delays ($r = .821, .910$) given the offence seriousness combined with the threshold set by the new amendments, such as the Show Cause test.

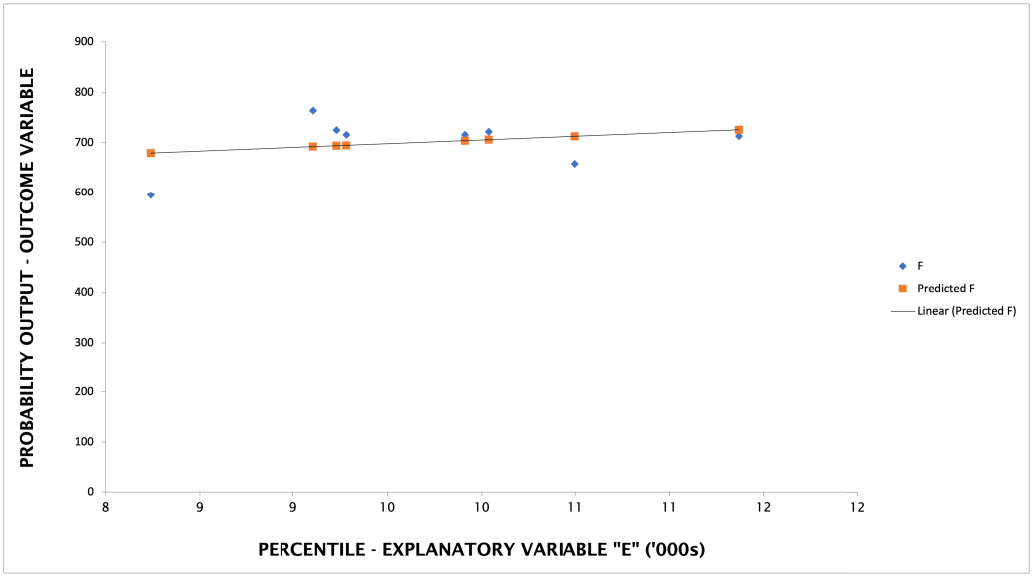

Figure 3.4: Line Fit Plot – Explanatory Variable E.

The one-way ANOVA on the two sets of observations did not reveal any results of significance, with the exception of Variable A ($p = .025$). It is noted however that using p-values as indicators can be misleading [129], therefore, other metrics were utilised as validation. Figures 3.3 and 3.4 display line plots or predictions of two explanatory variables in "D" and "E", and unexpectedly, neither follow the linear predictions. R-squared outcomes of both sets of observations reflect good model performance, and also, attributes any variations between the explanatory variables and the outcome variable within those values (R^2 .975 (2014–2021), .986 (2011–2021).

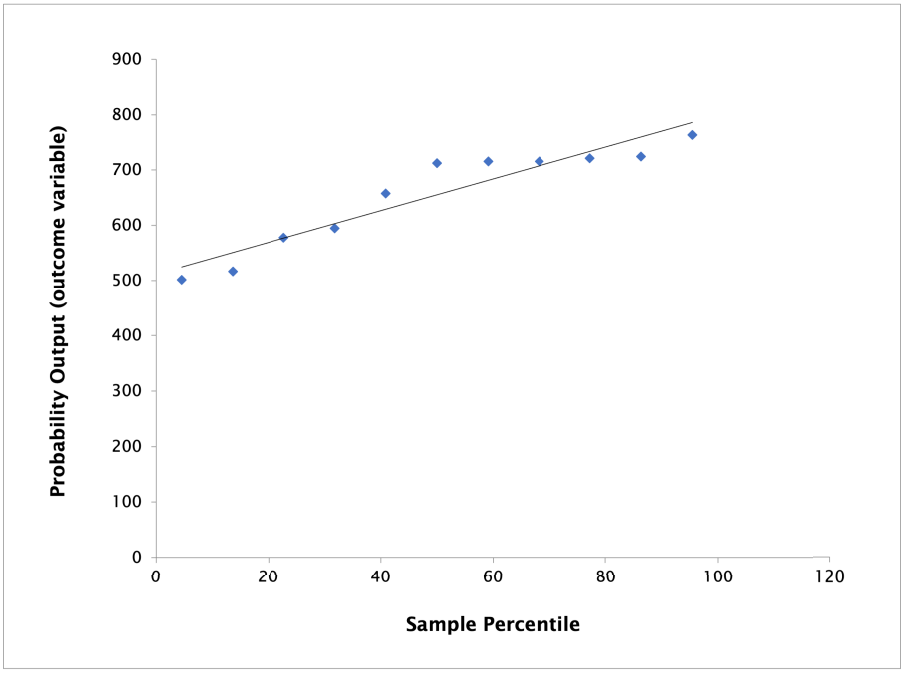

Figure 3.5: Probability output of eleven observations, 2011–2021 ($\mu = 655.14$, $\sigma = 91.65$).

Figures 3.5 and 3.6 represent the normal distribution of the year to year observations for the periods 2011 to 2021 and 2014 to 2021, respectively, and as demonstrated by the intersecting

trend line, the data distribution is lineal. Setting aside more complex statistical calculations, an observation of Variable F is that there is an increase of 52 percent regarding the median days in court delays over a 10-year period (shown in Table 3.1 and Figure 3.2). Consequently, further examination of the bail amendments' impact may be required to understand other factors contributing to court delays. Nonetheless, an identified need to lighten the burden from backlogged courts to assess bail applications was expressed by an appeals court judge – to remedy the delays, lessen the remandees and assist the lower courts in first instance applications – suggesting alternative means must be considered.

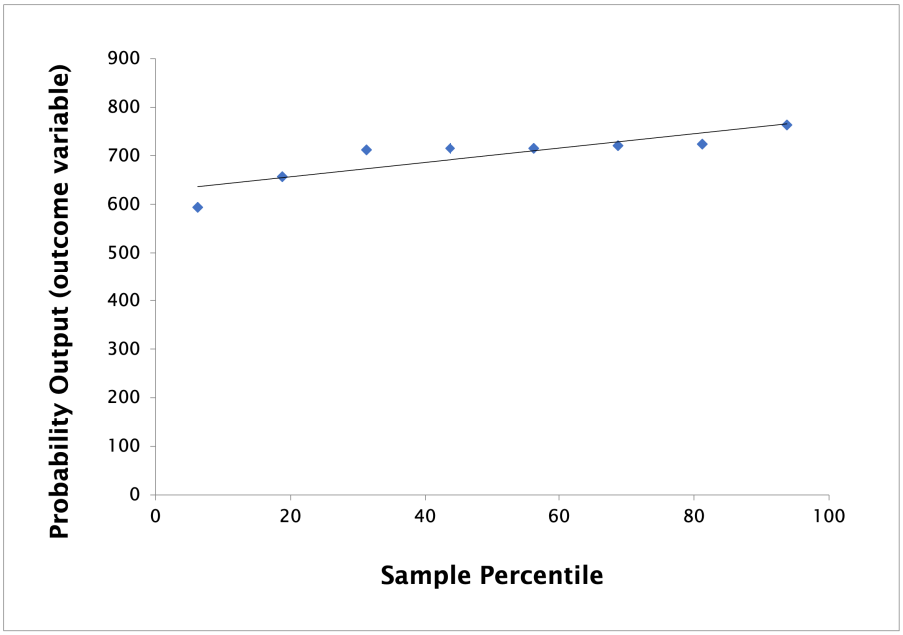

Figure 3.6: Probability output of eight observations, 2014–2021 ($\mu = 701.13$, $\sigma = 51.6$).

Reflecting on elements within Research Objective 1 (RO1) – to identify and explore ML techniques or algorithm-based models – methods in regression and correlation were applied for data analysis. While having been applied to assess the impact of bail legislation on certain categories rather than bail decisions *per se*, this did allow for the groundwork on theoretical and empirical ML and VA techniques in an undertaking on predictive and visual analyses.

Chapter 4

Preliminary Study II: Groundwork on Predictive Modelling to Decide Bail

4.1 Background and Motivation

Decision-making is a fundamental undertaking of criminal courts and it is expected that decisions are consistent with all other corresponding cases. An emphasis on "consistency of decision-making" was remarked upon in the Second Reading of the *Bail Amendment Bill 2014* [117]. Despite this, the very nature of bail decisions made by humans are presumably going to yield inconsistencies. As a justification, it may be argued that every case has its point of difference, and so there is an inevitability that the decisions will vary. Opposing this argument is that each decision should be consistent, for the criminal laws categorise a relevant offence and is guided by specific elements or proofs. Yet more importantly, the *Bail Act 2013* (NSW) relies on 24 determined factors of which the bail authority refers to when determining bail outcomes. This was remarked on in [130]:

> ... [section] 18 limits the matters to be taken into account in assessing bail concerns under the Act. Each of the matters is given equal priority. No one matter assumes dominant significance.

Even though a systematic process is legislated and adhered to, it must be accepted that inconsistencies will occur, allowing for a reasonable expectation in improved consistency. Moreover, where there are categories that remain constant on all bail cases, hypothetically, such determinations could be made that ensures improved consistency. What of the Court itself when it considers consistency, balance and efficacy when making bail decisions? The following extracts have been taken from judgments on bail cases that lead to this question:

> It is possible, if not on one view inevitable, that minds may differ in any particular case about how these authorities should play out in the difficult discretionary exercise with which I am presently concerned [131] ... Bail decisions involve a discretionary evaluative judgment on a variety of factors about which, and within limits, reasonable minds may differ ... But it must be recognised that most of these judgments are very specifically directed to the facts and circumstances of the case

> at hand. It is useful for "bail authorities" to have examples of how particular factual circumstances have been considered by Supreme Court judges. But every bail application presents its own unique factual matrix [132].

An opposition to what could be conceivably be a future of bail decision-making using "base rates" and historical data was expressed:

> It would not be an exaggeration to say that upon each occasion the applicant has been granted conditional bail, generally in respect of relatively minor charges, conditional bail has been imposed by a court, and regularly the applicant has disregarded those conditions. This is not a base from which confident predictions of future compliance with bail conditions can be made.[T]he approach of the Court falls into a dichotomy. If there is an unacceptable risk, the Court must refuse bail; if there is no unacceptable risk, the Court must grant bail ... That test applies to all offences [130].

In the last extracts, the NSW Supreme Court was referring to the Stage 2 / Flow Chart 2 assessments in the Bail Act 2013, conclusively: if an unacceptable risk is apparent, a decision of 'no' and bail is refused; if an unacceptable risk is not apparent, a decision of 'yes' and bail is granted. As demonstrated in the BAILgram (see Figure 2.1), the outcome sought from this predictive assessment is to decide if bail is to be granted or refused – a binary or dichotomous outcome – and therefore logical for the most suitable measure to be in a binary format and the chosen statistical methods are binary logistic regression (B-LogR) and a tree-structured classifier (TsC).

Machine Learning Models

This preliminary study intends on applying two algorithm models, and as stated in the literature, it is acceptable practice for data to be applied to more than one model to ascertain the most suitable approach [95]. B-LogR is one model that will be measured against the data for it is said to correspond well with ECs, as discussed in Chapter 2. Additionally, B-LogR is an effective statistical model when applying observational data [133] and in analysing binary outcomes [90]. A TsC is said to be a robust performer in predictive analysis, it benefits interpretability, is understandable and less complicated [80][81][134], and accordingly, it is one model selected and will be more suitable than other ML models in its relationship to the bail schema.

Uniformly, the literature on ML being applied in the criminal justice domain endorses logistic regression and tree-structured classifiers for predictive modelling and decision-making instruments. The justification for this given, for example, was that logistic regression is ideal for binary classification [82] [88] [89], and, mollifies ethical concerns beyond that of other techniques, such as Neural Networks [64] and Random Forests [88]. Similarly, more modern developments in ML such as Neural Networks, meant its application as a tool was met with greater complexities than that of relatively less complex ones like tree classifiers [135]. The literature also exhibited research having measured various ML methods where the predictive rigour of those selected

was comparable, two of those being logistic regression and tree classifiers [3][75]. Given these scholarly examples having sustained ethical AI considerations and applied suitable ML methods for predictive modelling specifically in the criminal justice domain, the B-LogR and TsC are appropriate in the undertaking on AI-generated decision-making for bail in this preliminary study.

4.2 Statistical Methods

There is argument bound on historical practice and expectation for statistical research to be tested against theories and hypotheses, and the causal elements and statistical results to be determined by deduction. This preliminary study does not intend to form hypotheses and develop causation based on the analyses underpinned by theoretical premise and model explanation, rather, it is centred on *predictive modelling*. It is imperative to provide the rationale for this inclination, however, constructive to differentiate 'explanatory' from 'predictive' modelling.

Predictive modelling is the application of algorithmic modelled numerical data to forecast a future event or make a predictive observation [136]. Alternatively, explanatory modelling is a resultative undertaking in examining causal outcomes based on theories and hypotheses [136]. While explanatory modelling contains measurable data, it becomes misrepresented by the theory underpinning it, leading to an inefficacy for explainable phenomena [136]. Notwithstanding, a requirement in predictive accuracy in the criminal justice domain is that the future has representations of the past and the future prediction is dependent upon mathematical rigour obtained through superior data collection and analysis [88].

Subsequent to or working diachronically with the predictive model approach is Exploratory Data Analysis (EDA). EDA is not wholly defined but is characteristic of analyses methods that precede conventional methods, upon which analysis can be drawn from, namely, visualisation, where hypotheses are not a foundational requirement in a study [136]. Another example of non-conventional methods is concerning the probability value. The disputation in the use of the explainable type of metric extended to significance testing. The customary p-value as it is commonly referred, is usually accompanied by arbitrary indicators of 0.05 or 0.01 [137]. The preference in measurement criterion by [137] is to instead apply and rely on 'confidence intervals' for its superiority to hypothesis testing, for example, in study replication. Taking into consideration the contentious yet compelling scholarly rationale, this preliminary study will break convention and be devoid of hypothesis, although a reasonable statistical overture will be presented in relation to the predictive model.

The groundwork in this preliminary study is to develop a working prototype to predict bail decisions using actual data – to forecast whether a defendant should be granted or refused bail based on nine distinct predictor variables – the data drawn from 101 written bail judgments. Once the outcomes were in an analysable format, it was evident there would be a disconnect with previously formulated hypotheses on causation for the bail decisions. This was supported by [136] having stated that sizeable datasets present correlational and pattern complexities, which become difficult to hypothesise.

A contemporary approach to measure the risk of reoffending is by ascertaining the relationship between independent/predictor variables and how it impacts on a dependent/target variable [53]. Similarly, this research will ascertain if any relationship is evident between the categorical variables to predict the probability that if released on bail, a defendant would reoffend. As mentioned, B-LogR and TsC are the applied techniques – for the purposes of analytical reasoning, those decisions will then be formulated into several VA outputs, to then be presented to participants who will be surveyed as the main approach in the Primary Study (see Chapter 5).

The data having been applied to Bail-14 produces a dichotomous outcome: bail is either *granted* or *refused*. Purposed as reliable measures on the performance of a ML model will be the ROC Curve and classification table [64] for their predictive vigour, reasonably justified by [53] who stated "...the more similar the classification tables, the less evidence there is for unfairness." On the point of fairness, yet lends itself more toward independent testing and replication of model accuracy expected in research methodology, is the technique of *cross-validation*. This technique is guided by the model to apportion or split cases in a randomised manner that ultimately supports model preference and accuracy, whilst counteracting overfitting [64][138].

Data Evaluation and Metrics

A collation of 101 bail cases across various levels of the NSW criminal courts formed the dataset, accessed from the NSW CaseLaw website. CaseLaw is an open-source publicly accessible website with written decisions on nominated cases in most jurisdictions of the NSW law courts. The search parameters contained "bail", the chosen four court levels of "Local", "District", "Supreme" and "Court of Criminal Appeal (CCA)" and the time period parameter from "2015" to "2023", as the year 2015 was when the last legislative amendment was enacted. The search parameters made available more than 600 cases however this eventually was filtered down to 101 cases based on detail and relevance, from the District, Supreme and Appeal Courts (see Table 4.1). It is to be noted that some of the cases contained two or more multiple parties heard at the same time (co-defendants), although each defendant was counted as one case and assessed individually, as the decisions and factors contributing to those cases were dealt with separately by the respective court.

Table 4.1: Number of defendants per court corresponding to data collected from case narratives (n=101)

Court type	Number of cases
Criminal Appeal	34
Supreme	60
District	7

The data was manually recorded corresponding to each variable. For instance, where a written judgement stated that a defendant had one prior failure to attend court, this was subsequently recorded under the predictor Failure to attend/flight-risk as "1". If a judgment did not contain the information to satisfy each of the predictors, it was disregarded.

Each predictor variable is recorded commensurate to the information in the Model Predictor Information Table (see Appendix B). All variables were binary coded either "0 = no" or "1 = yes" except for Criminal history and Seriousness of offence(s), as they required more specific coding. Under the class Seriousness of offence(s), if viewed on a scale can vary, given factors associated with harm to victims. Therefore, the seriousness was rated on the maximum penalty that could be imposed under the sentencing law for the index offence. Guided by domain experience in sentencing procedures and administration, seriousness was gauged by the penalty and the hierarchy of the court. For example, the three year marker is significant as a penalty of this degree can only be determined by superior courts and three years or more custodial sentence requires the NSW State Parole Authority to determine the schedule and conditions of an offender's release. *Show cause* under the legislation is a "yes" or "no" outcome and therefore coded the same in this instance, "1" and "0" respectively. Show cause data input determines if the defendant has shown cause why their detention is not justified, which resembles the NSW bail legislation, but was varied, giving greater weight to certain offences in the categories of sex crimes and domestic/family violence. It is important to note that the interpretation of this class is transposed for the two models.

One of the determining factors for bail in the NSW legislation refers to *if* a defendant was convicted, whereby the emphasis on 'if' is a subjective assessment made by a magistrate or judge on the probability for a conviction. In order to minimise this subjectivity (where essentially a human is tasked to refer to 24 factors and predict the probability that a defendant would or would not comply with conditional release) the 24 factors were condensed down to nine predictor variables that make up the Bail-14 prototype model, making the predictors less ambiguous. Attributes such as gender, age, and race have intentionally been omitted, which create biases in the results, and from a logical and moral standpoint, the weight of those attributes should not give cause to determine a defendant's liberty.

Classification Metrics

Table 4.2 lists the information-based and error-based measures along with each respective equation. Outcomes are determined numerically between 0 and 1 [66].

Table 4.2: Error-based and Information-based measures (equations in bold).

TPR: the probability of a correct classification of *Granted* **TP/(TP+FN)**.
TNR: the probability of a correct classification of *Refused* **TN/(FP+TN)**.
FPR: the probability of incorrect classifications – *Granted* being *Refused* **FP/(FP+TN)**.
FNR: the probability of incorrect classifications – *Refused* being *Granted* **FN/(FN+TP)**.
PPV: the probability of compliance given an outcome of *Granted* **TP/(TP+FP)**.
NPV: the probability of non-compliance given an outcome of *Granted* **TN/(TN+FN)**.

Nine variables in the Bail-14 prototype are derivatives of the twenty-four factors listed under Section 18 of the Bail Act 2013. It was deemed that some of those listed factors overlapped or are similar in context, and therefore, for the formulation of Bail-14, those factors were either merged or omitted. Also, there are items that appear subjective and their evaluative cogency is consequently limited. Given these factors, the nine predictors applied to the Bail-14 model are listed in Table 4.3.

Table 4.3: Nine predictor variables: *Bail-14* predictive model. Note: Appendix B contains more specific information on the nine predictors in *Bail-14*.

(1) Show Cause
(2) Criminal history
(3) Seriousness of the offences(s)
(4) History of violence
(5) Bail non-compliance (whether under the Bail Act 2013 (NSW) or other jurisdiction)
(6) History of non-compliance (with court-issued orders)
(7) Pro-criminal associations
(8) Threat/danger to the victim(s), public or others
(9) Failure to appear/Flight risk

Demographic identifiers, while relevant in certain assessments, have proven to create issues in fairness (among other ethical principles) in the justice and law domains – it is for these reasons that demographic identifiers will not be used. In saying this, the Bail Act 2013 does not contain demographic criteria.

Binary Logistic Regression

This is a parametric statistical approach to ultimately produce a dichotomous outcome. It is a cogent predictive measure in binary classification for values of 0 or 1, true or false, or yes and no [139], suitable for predictive modelling where variables of consequence need to be ascertained [140]. When considering this approach from an ethical standpoint in comparison to other ML approaches, B-LogR is understandable [139], interpretable, explainable and transparent. Notwithstanding the ethical dilemma that may arise from prediction as an undertaking in itself, an ML-driven regression measure is applied to determine and explain the relationship between the predictors and the dependent variable, and from this, evaluate the predictive power of the relationship [129]. While not being an objective of this preliminary study to measure the effectiveness of the model, it is relevant in the space of ethical principles, such as, explainability.

Data was split in several ratios to account for overfitting and to determine model accuracy [88][64][141]. A multivariate multiple regression statistical method was applied. The reason for this method is having two dependent variables, or that the response variable (Y) has two values that are qualitative [142] – in this instance, *granted* and *refused* – and as such, the actual court decision from each bail case is the target variable.

The model applied here is an archetype of a logistic regression algorithm [75][143][144], and inspired by a model from a scholarly source [89]:

$$p(y) = \frac{1}{1 + exp(a + b^1c^1 + b^2c^2 + b^3c^3...b^9c^9)} \tag{4.1}$$

where p is the probability of success (bail granted = y), a is the intercept, b are the coefficients corresponding to a recorded instance of the predictor c and refers to the binary instance of 0,"no"; 1,"yes".

As listed in Table 4.3, each predictor variable is accompanied by its coefficient, as demonstrated in Equation 4.1: Show Cause (*c1*); Criminal history (*c2*), Seriousness of the Offence(s) (*c3*), History of violence (*c4*), Bail non-compliance (whether under the Bail Act 2013 (NSW) or other jurisdiction) (*c5*), History of non-compliance (with court-issued orders) (*c6*), Pro-criminal associations (*c7*), Threat/danger to the victim(s), public, or others (*c8*), Failure to appear/Flight-risk (*c9*).

The designation of the codes and values were assigned by a code-based program *Colectica* [145], which is an 'add-in' on Microsoft Excel that supports coding variables. Another Excel add-in *Real Statistics Using Excel Resource Pack* [128] was applied, which enables logistic regression functions while facilitating categorical and variable coding. These packages analysed the data from the 101 bail cases on the probability of success or failure for a defendant being granted or refused bail based on the nine predictors listed in Table 4.3.

Applying domain knowledge in risk assessment and jurisprudence [146], three or more convictions was a satisfactory threshold of an individuals historical offending. *Seriousness of offence(s)* classification was based on the current penalty range for the index offence in accordance with the NSW sentencing legislation: equal to or less than 3 years imprisonment (1, low); greater than 3 years-not more than 10 years imprisonment (2, moderate); and greater than 10 years imprisonment (3, high). More specific details can be found in Appendix B.

The classification cut-off or threshold was finalised at 0.4. While the threshold is gauged or presumed by the intentions or purpose of the model, it was determined that 0.4 was a suitable marker, given the trade-off that the accuracy was at its foremost. When determining to grant or refuse bail, the chance bet is not a suitable option, rather, a definitive metric is needed: a defendant who scores equal to or below 0.4, would be refused bail; equal to or greater than 0.41 is indicative of a defendant being granted bail. Moreover, the literature suggested a threshold at 0.5 is an arbitrary marker, as it is seemingly not any more definitive than chance [64][138][147][148]. Accuracy is calculated by the addition of true-positives and true-negatives divided by the total number of bail cases tested [66]. The ROC Curve charts the effectiveness of the model in predicting a target, discernible by the curved line that follows the y-axis to the top left corner then turns to move parallel to the x-axis [64][138].

Several models were tested, aptly named after the split ratios: *Model 61-40* and *Model 51-50*. Additionally, the models were also tested on their performance by removing two predictor variables, selected on domain knowledge in addition to literature on recidivism [89][149] and these sub-models are labelled *-CRIM50* and *-SoO50*.

Tree-structured Classifier

Otherwise referred to as Decision Tree, the TsC is a non-parametric approach, propagated out of the Bail-14 dataset and is herein referred to as "Bail-Tree". Produced through the software *RapidMiner-Studio*[14], which is an open-source product, the data applied was imported through Excel, much the same as B-LogR, although with two distinct exceptions. First exception: the target variable does not require coding as the polynomials of 'granted' and 'refused' have a deliberate function as set parameters. Second exception: the Show Cause outcome is transposed, for in this process, the decision is determined by the defendant having to show cause why detention is not justified, which is the same as asking what legal argument can the defendant submit to justify their liberty, for example, not presenting any risk to the victim or community (whereas the B-Log-R interrogates the specific offence for which the defendant is being charged similar to the legislative framework, e.g., was the alleged offence committed while on bail or parole). As noted earlier, adjustments to this classification were made to give greater weight to categories under sex crimes and domestic/family violence. The target variable was labelled "Actual court decisions".

RapidMiner-Studio does not exclusively identify what algorithm is specifically applied under the "Decision Tree" operator although it closely replicates the perennial algorithms ID3 and C4.5 (see [150][151]). Bearing a closer resemblance to the C4.5 algorithm, Bail-Tree initially applies a top-down approach [151], meaning the first or root node is determined to be the most relevant predictor, and the second and so on, until the last two leaves display the outcomes. More specifically, it systematises the strength or value of all predictors selectively through data computation until it reaches a juncture, where the return path is again calculated by a bottom-up process, that ultimately selects the most relevant predictor at the root to then forms the branches and leaves by "splitting" according to their individual relevance or strength [150] [151]. In its more practical utility, the schematic flow charts in the Bail Act 2013 (see Appendix D) is also a top-down assessment upon which a bail authority is to reference in that decision-making process. Furthermore, the utility of a TsC was by simulating analytical decisions [78].

A split of the training (0.6/60%) and test data (0.4/40%), and to build the model, a "stratified sampling" option was applied as it was considered the most suitable for binomial calculation, "gain ratio" was the criterion deemed most suitable for splitting [14]. The confidence level was set at 0.1; the minimal gain 0.01; minimal leaf size set at 2 (size for split set at 4). The integers ranged between 0 and 3 and polynomials were Granted/Refused and Yes/No. Figure 4.1 provides the 'pseudocode' on this process.

Table 4.4: Bail-14 pseudocode exemplar to demonstrate simplified commands or syntax to build the T-sC.

1. Select bail cases.
2. For each predictor P, find the normalised information gain ratio from splitting P.
3. Let P best be the predictor with the highest normalised information gain.
4. Create a decision node that splits on P best.
5. Recur on the sublists obtained by splitting on P best, and add those nodes as children of the node $(P_i, (P_{ii}, P_{iii}...P_{iix})$.
6.Those will be children of node P which has the highest information gain ratio.

4.3 Results

Binary Logistic Regression

***Model 61-40*:** Training was conducted on 61 cases randomly selected, subsequently, the remaining 40 cases was the test data. Classification performance in Table 4.5 shows an overall accuracy of .775 (78%) – Successful Observations: 19 (.90); Failed Observations: 12 (.63). Figure 4.2 displays the ROC Curve (AUC .845) and is indicative of moderate to strong model accuracy.

Table 4.5: Classification Table – *Model 61-40*

		Actual Class	
		Granted	Refused
Predicted Class	Granted	19	7
	Refused	2	12

TPR suggests the probability of a correct classification of Granted is 90% (.90), while the TNR suggests the probability of a correct classification of Refused is 63% (.63). FPR at 36% (.36) and FNR at 9% (.09) refers to probability of incorrect classifications. PPV (.73) and NPV (.86), means the probability of compliance given an outcome of Granted is 73% and the probability of non-compliance given an outcome of Granted is 86%.

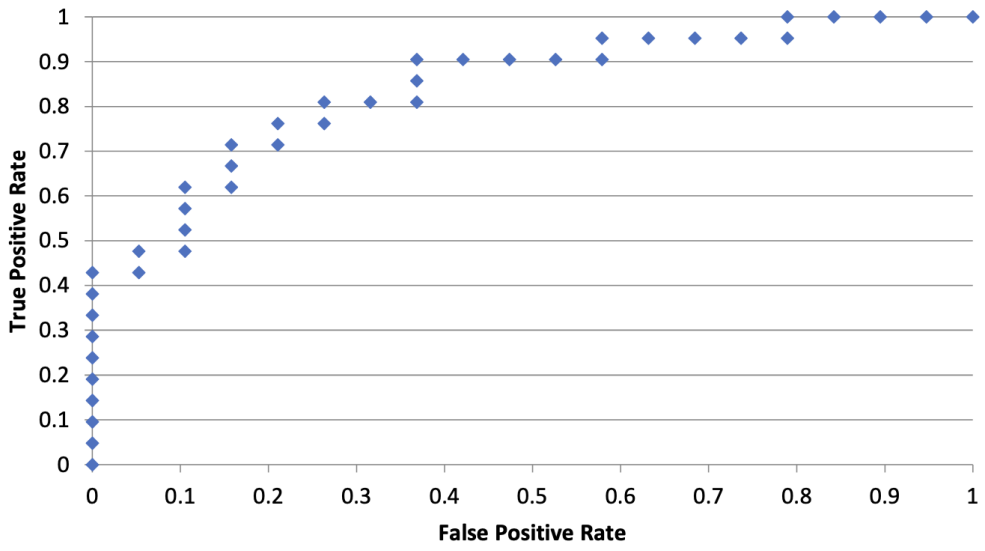

Figure 4.1: ROC Curve - *Model 61-40* (AUC .845, 95% CI).

Statistically assessed by means of predictive value (p<.05), the model was rendered significant (p=.02). The variance-covariance matrix presented in Table 4.6 shows Criminal history and Failure to appear/flight risk resulted in a negative correlation (1.8E+08). Bail non-compliance featured prominently: a negative correlation with History of violence (-.09) and similarly with Pro-criminal associations with a positive correlation (.09). Show cause was negatively correlated with Bail non-compliance (-.08) and positively correlated with Pro-criminal associations (.02).

Table 4.6: Variance-Covariance Matrix – *Model 61-40.*

	Criminal history	Seriousness of Offence(s)	History of violence	Bail non-compliance	Non-compliance with other orders	Pro-criminal associations	Threat towards victim/ witness/ public safety	Failure to appear/ Flight risk	Show Cause Offence	Actual Decision
Criminal history	1.80E+08	-1.52	-3.26	0.34	2.26	-0.98	-1.72	-1.80E+08	1.19	0.96
Seriousness of Offence(s)	-1.52	0.33	0.31	-0.16	-0.24	-0.06	0.18	0.54	-0.24	-0.05
History of violence	-3.26	0.31	1.25	-0.09	-0.76	0.85	0.27	0.28	-0.34	-0.89
Bail non-compliance	0.34	-0.16	-0.09	0.56	-0.22	0.09	-0.14	-0.11	-0.08	0.18
Non-compliance with other orders	2.26	-0.24	-0.76	-0.22	2.64	-1.85	-0.06	-0.51	-0.24	0.68
Pro-criminal associations	-0.98	-0.06	0.85	0.09	-1.85	2.70	-0.04	-0.63	0.03	-1.25
Threat towards victim/others	-1.72	0.18	0.27	-0.14	-0.06	-0.04	1.50	0.28	-0.66	-0.21
Failure to appear/Flight risk	-1.80E+08	0.54	0.28	-0.11	-0.51	-0.63	0.28	1.80E+08	-0.34	0.48
Show Cause	1.19	-0.24	-0.34	-0.08	-0.24	0.03	-0.66	-0.34	1.52	0.06
Actual Decision	0.96	-0.05	-0.89	0.18	0.68	-1.25	-0.21	0.48	0.06	1.69

Key
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00

Model 51-50: Training was conducted on 51 cases randomly selected. Subsequently, test data was taken from the remaining 50 cases, the performance displayed in Table 4.7. Overall accuracy was 76% (.76) – Successful observations at 20 (.83) while Failed observations at 18 (.69). TPR and TNR values are .83 (83%) and .69 (69%) respectively; FPR .31 (31%) and FNR .16 (16%); PPV .71 (71%) while the NPV .82 (82%).

Table 4.7: Classification Table – *Model 51-50.*

		Actual Class	
		Granted	Refused
Predicted Class	Granted	20	8
	Refused	4	18

Figure 4.2 displays the ROC Curve (AUC .845) determining accuracy. At .845 the model represents moderate to strong accuracy. The predictive value was calculated at .01, which determined model significance ($p < .05$).

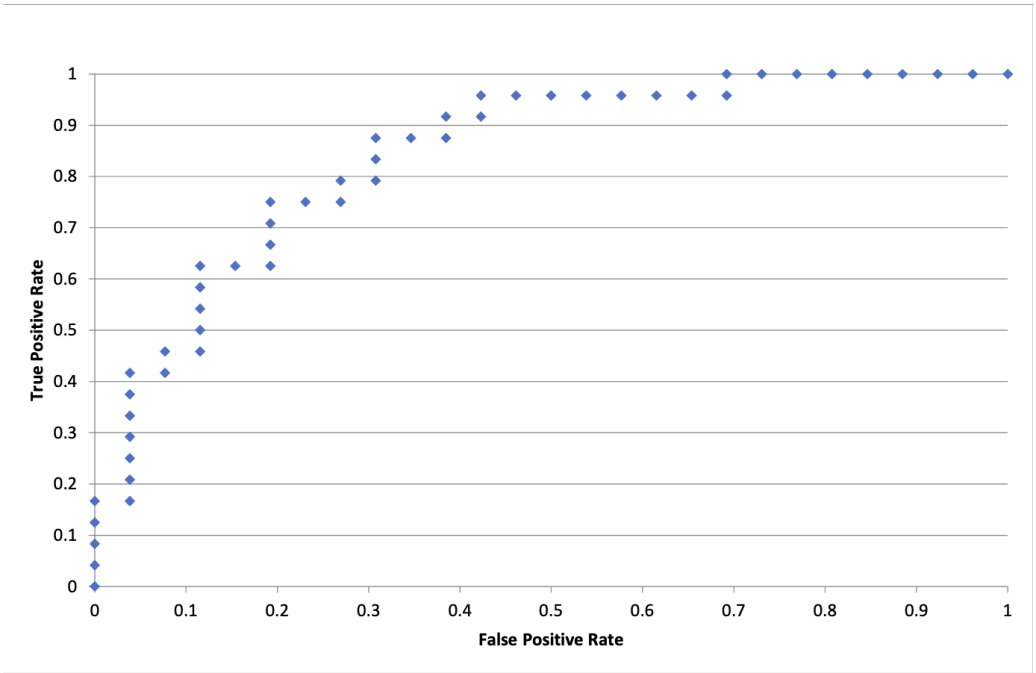

Figure 4.2: ROC Curve – *Model 51-50* (AUC .845, 95% CI).

Taking into consideration the literature on evaluating predictive modelling performance, [89][149], it was warranted to measure what relevance the prominent two predictors had on the performance of Model 51-50, namely, Criminal history and Seriousness of offence(s). To measure this, a datapoint comparison was undertaken using the ROC values of the TPR and FPR from Model 51-50, with the ROC values of the two sub-models, following the deletion of the relevant classification.

Table 4.8: Classification Table – sub-model -*CRIM50*.

		Actual Class	
		Granted	Refused
Predicted Class	Granted	22	11
	Refused	2	15

In Figure 4.4, the two sub-models are denoted as "-CRIM50" and "-SoO50". Datapoints and the models performance are noted after the class was removed.

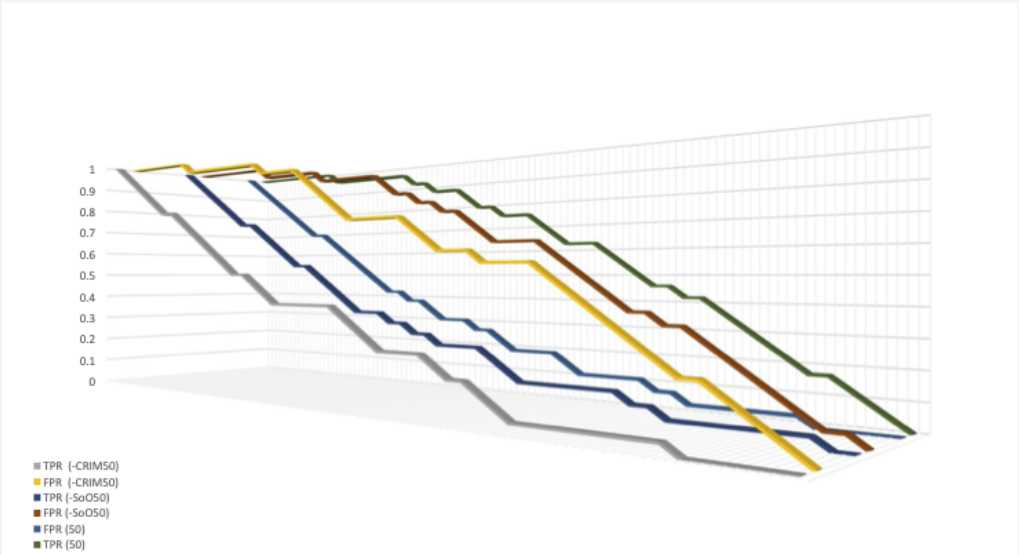

Figure 4.3: TPR v FPR – *Model 51-50* – comparison to sub-models where one predictor variable was removed.

As referred to previously, the full model determined the TPR as 83% (.83) and FPR as 31% (.31). After Criminal history was removed, the sub-model determined the TPR as 91% (.91) and FPR as 41% (.42). Then, Seriousness of offence(s) was removed, the sub-model determined the TPR 83% (.83) and FPR 35% (.35).

Table 4.9: Classification Table – sub-model -*SoO50*.

		Actual Class	
		Granted	Refused
Predicted Class	Granted	20	9
	Refused	4	17

Table 4.10 shows the variance-covariance matrix, where a positive correlation is observed between Seriousness of offence(s) with History of violence (.036) and Failure to appear/flight Risk (.049), while Seriousness of offence(s) has a negative relationship with Bail non-compliance (-.097), Non-compliance with other orders (-.040), and Pro-criminal associations (-.009). A positive correlation was observed between History of violence and Failure to appear/flight Risk (.030).

Table 4.10: Variance-Covariance Matrix – *Model 51-50*.

	Criminal history	Seriousness of Offence(s)	History of violence	Bail non-compliance	Non-compliance with other orders	Pro-criminal associations	Threat towards victim/others	Failure to appear/ Flight risk	Show Cause	Actual Decision
Criminal history	3.16	-0.25	-0.83	0.34	-0.21	-0.60	-0.30	-0.30	0.10	-0.24
Seriousness of Offence(s)	-0.25	0.12	0.04	-0.10	-0.04	-0.01	0.06	0.05	-0.06	0.02
History of violence	-0.83	0.04	0.40	-0.07	0.15	0.14	-0.05	0.03	-0.20	-0.14
Bail non-compliance	0.34	-0.10	-0.07	0.39	0.06	-0.28	-0.14	-0.13	0.09	-0.06
Non-compliance w/ other orders	-0.21	-0.04	0.15	0.06	1.25	-0.86	-0.25	-0.08	0.17	-0.29
Pro-criminal associations	-0.60	-0.01	0.14	-0.28	-0.86	1.62	0.51	-0.05	-0.32	0.16
Threat towards victim/others	-0.30	0.06	-0.05	-0.14	-0.25	0.51	1.25	-0.13	-0.50	-0.19
Failure to appear/Flight risk	-0.30	0.05	0.03	-0.13	-0.08	-0.05	-0.13	0.81	-0.15	-0.13
Show Cause	0.10	-0.06	-0.20	0.09	0.17	-0.32	-0.50	-0.15	1.03	0.25
Actual Decision	-0.24	0.02	-0.14	-0.06	-0.29	0.16	-0.19	-0.13	0.25	0.94

Key: 3.5, 3, 2.5, 2, 1.5, 1, 0.5, 0, -0.5, -1

Tree-structured Classifier

Overall accuracy was 72.5% and classification error 27.5%. As shown in Table 4.11, the TPR value is .66, suggesting the probability of a correct classification of Granted is 67%; TNR value .77, suggesting the probability of a correct classification of Refused is 77%.

Table 4.11: Classification Table – TsC (Accuracy)

	Actual Class		
Predicted Class	Granted	Refused	Class Precision
Granted	12	5	70.59%
Refused	6	17	73.91%
Class recall	66.67%	77.27%	

The PPV and NPV were .71 and .74, respectively, conveying the probability of complying given an outcome of Granted is 71% and the probability of non-compliance given an outcome of Granted is 74%. Probability of incorrect classifications is measured by the FPR 22% (.22) and FNR 33% (.33); the probability of granted being refused was 22% (.22), and conversely, refused being granted was 33% (.33).

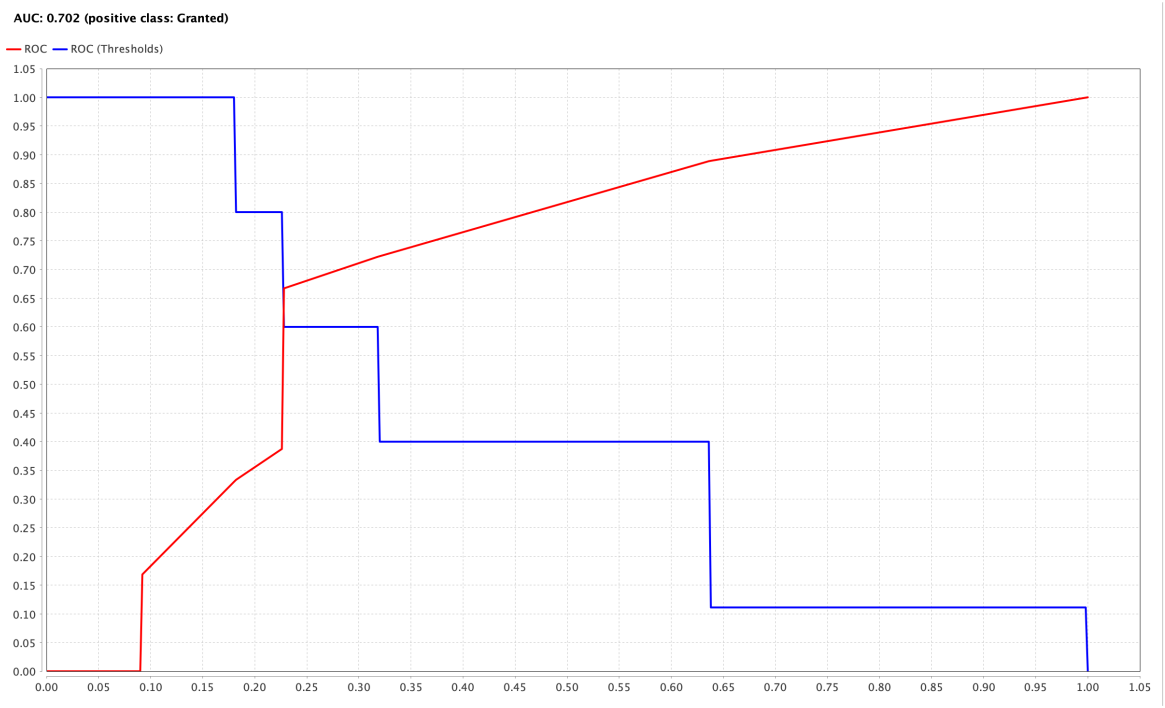

Figure 4.4: ROC Curve for TsC model (AUC .702). The red line denotes the standard plots on x-axis and y-axis and the blue line denotes the ROC Threshold (values on y-axis are reversed). Graph output is a feature of the "Performance" classification parameters by [14].

Figure 4.4 displays the ROC Curve as a measure for the TsC models' performance at a tree-depth of seven. Denoted by the red line (AUC .702) the model is moderately accurate. When selected at tree-depths of "eight" and "nine", the overall accuracy was 72.5%. When there was not any specified tree depth (indicated by -1) the accuracy was maintained at 72.5%. Shown in Figure 4.5 is the TsC model descriptor for the unspecified depth and Figure 4.6 displays the TsC visualisation for the tree-depth at "eight".

```
Seriousness of Offence(s) > 2.500
|   Show Cause (Y/N) = No: Refused {Granted=0, Refused=5}
|   Show Cause (Y/N) = Yes
|   |   Criminal history > 2.500: Refused {Granted=1, Refused=8}
|   |   Criminal history ≤ 2.500
|   |   |   Criminal history > 1.500: Granted {Granted=4, Refused=2}
|   |   |   Criminal history ≤ 1.500
|   |   |   |   Bail non-compliance > 0.500: Refused {Granted=0, Refused=3}
|   |   |   |   Bail non-compliance ≤ 0.500
|   |   |   |   |   History of violence > 0.500: Refused {Granted=0,
Refused=2}
|   |   |   |   |   History of violence ≤ 0.500
|   |   |   |   |   |   Pro-criminal associations > 0.500: Granted
{Granted=3, Refused=2}
|   |   |   |   |   |   Pro-criminal associations ≤ 0.500: Refused
{Granted=2, Refused=3}
Seriousness of Offence(s) ≤ 2.500
|   Pro-criminal associations > 0.500
|   |   Seriousness of Offence(s) > 1.500
|   |   |   History of violence > 1.500: Granted {Granted=4, Refused=2}
|   |   |   History of violence ≤ 1.500
|   |   |   |   Failure to appear/Flight Risk > 0.500: Refused {Granted=0,
Refused=3}
|   |   |   |   Failure to appear/Flight Risk ≤ 0.500
|   |   |   |   |   Show Cause (Y/N) = No: Granted {Granted=2, Refused=0}
|   |   |   |   |   Show Cause (Y/N) = Yes: Refused {Granted=1, Refused=2}
|   |   Seriousness of Offence(s) ≤ 1.500: Granted {Granted=2, Refused=0}
|   Pro-criminal associations ≤ 0.500: Granted {Granted=8, Refused=2}
```

Figure 4.5: TsC model descriptor results at a tree-depth of "eight" based on *Bail-14* data.

The trade-off in the attempt to increase overall accuracy impacted the true-positive and true-negative values, For example, when the minimal gain parameter was adjusted to .05, it resulted in increased TPR .77/TNR .54, although the overall accuracy was reduced to 65%.

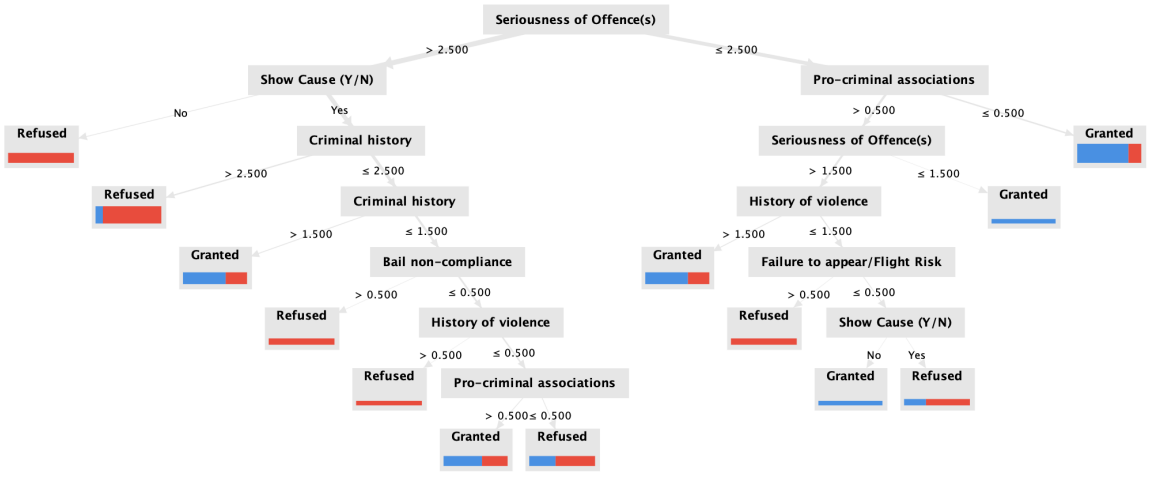

Figure 4.6: Screenshot of the TsC at a tree-depth of "eight" of *Bail-14* data.

Statistical data comparison of BOCSAR [13] (2015–2023) and *Bail-14* predictive model output

Upon examination of the Bail-14 predictive models error-based and information-based results, it is meaningful to undertake a comparison of the statistical data on bail related matters over the period 2015 to 2023. The data comparisons relied on have been extracted from [13].

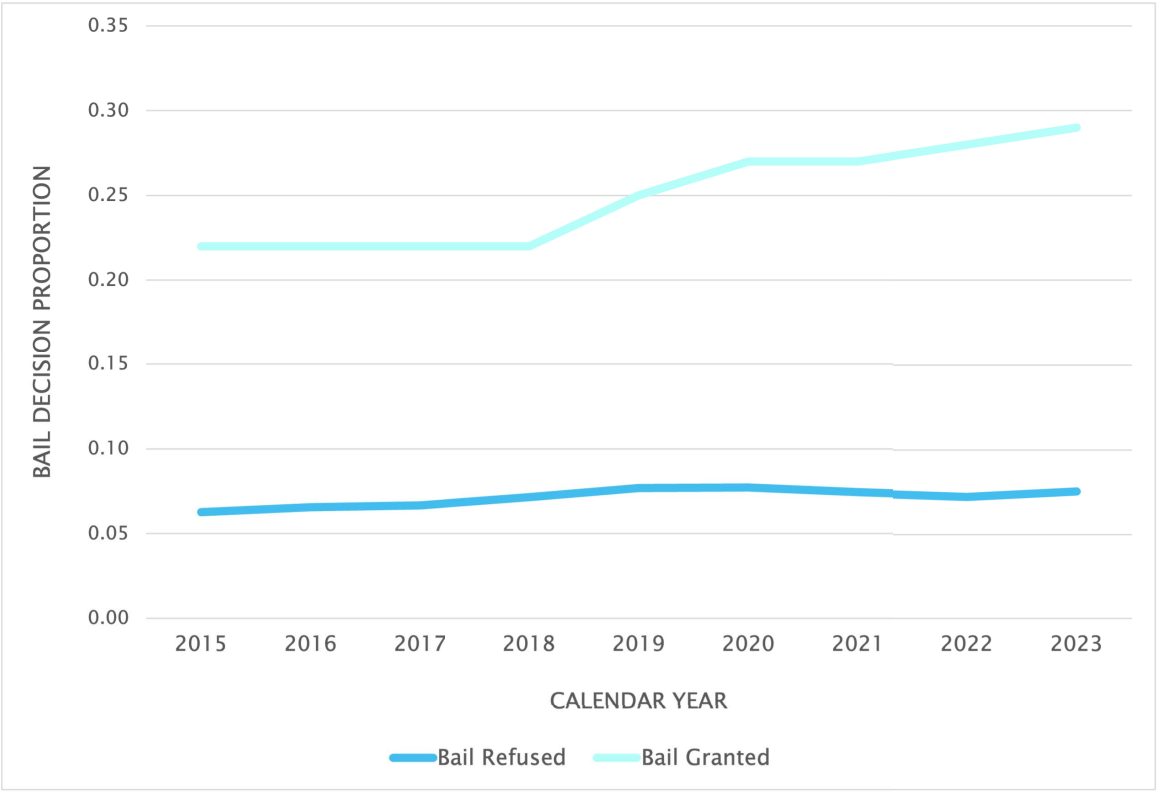

Figure 4.7: Bail decisions proportionate to the total number of bail matters at finalisation. Raw numbers were extracted from [13] and calculated as a proportion to the total number of defendants who had bail matters before all adult courts in NSW over the period 2015–2023. Note:"*finalisation*" refers to a defendants bail status at their final court appearance.

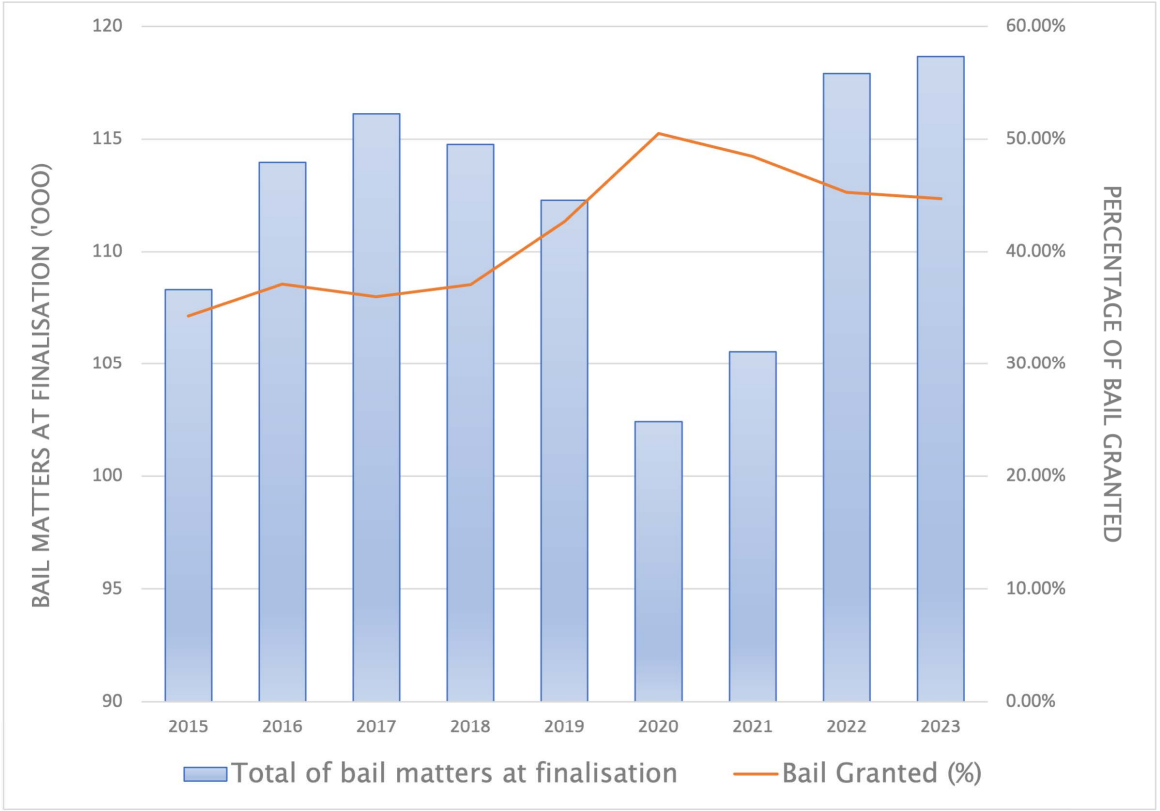

Figure 4.8: Bail status at finalisation – all defendants compared to percentage of defendants granted bail. Data extracted from [13].

Figure 4.7 graphs the decisions of the criminal jurisdictions in NSW (excluding Children's Court) proportionate to defendants who were refused bail to defendants granted bail between 2015 and 2023. Notably is the gradual increase in granted decisions from 2018 onwards, although the refused decisions remained stable.

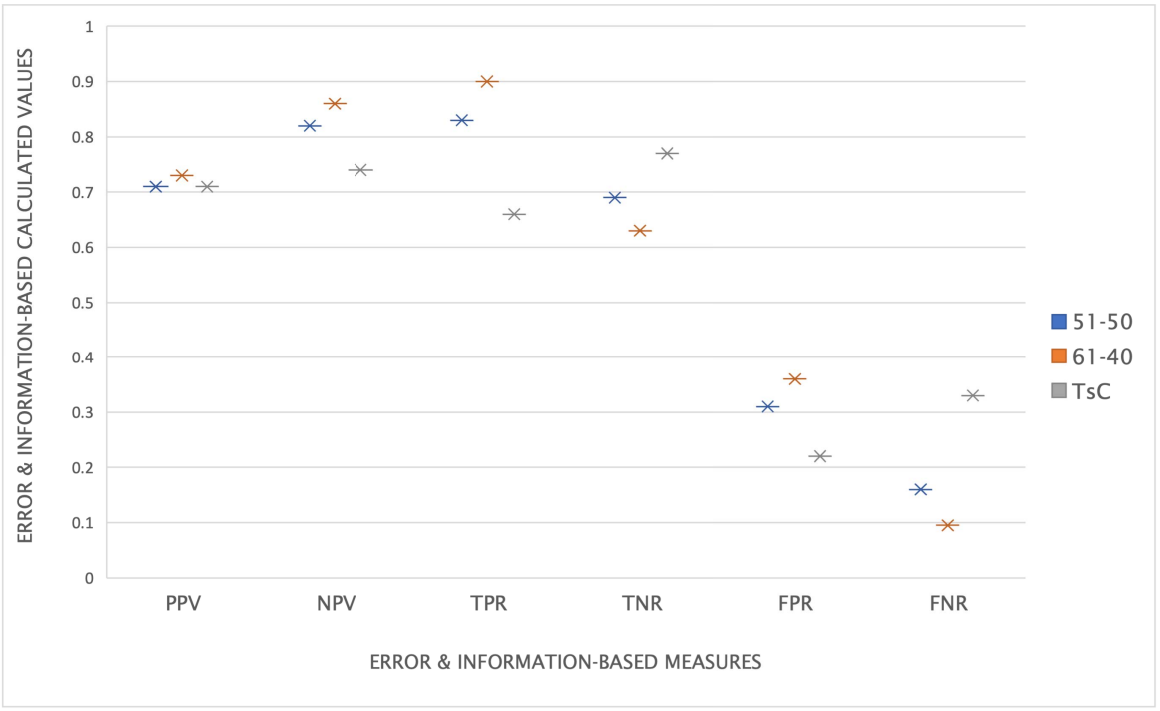

Figure 4.9: Error measures of the two regression models and tree-structured classifier.

Figure 4.8 charts bail granted numbers compared to the total of matters with a bail status at finalisation. Notably, there appears to be an inverse relationship: when bail matters increase, the number granted decreases; and when bail matters decrease the number proportionate to that for granted increases. Anecdotally, the inconsistency in the inverse relationship in the 3-year period from 2019 to 2022 could be accounted for during the pandemic when courts had to modify their decisions due to unprecedented circumstances [13], but this cannot explain the other inverse inconsistency between 2015 and 2019. Figure 4.9 displays the Error-based and Information-based measures from the two predictive models. The box plots are mostly concentrated in the same areas for each measure, although the TsC does not demonstrate the same consistency as the two regression sub-models.

In order to draw a reasonable comparison of the two models to recent bail statistics, in particular, categories of "granted", "refused" and "breach of bail", probability distributions were calculated and then compared to the error and information based measures. Recalling that earlier, an inverse relationship was observed between two variables, it was therefore pertinent to base the predictive measure outcomes within the same period. The following equation on probability distributions was applied:

$$f(x) = P(X = x) \tag{4.2}$$

where P is the probability, X is the random variable, and x is the mean [152]. Table 4.11 lists

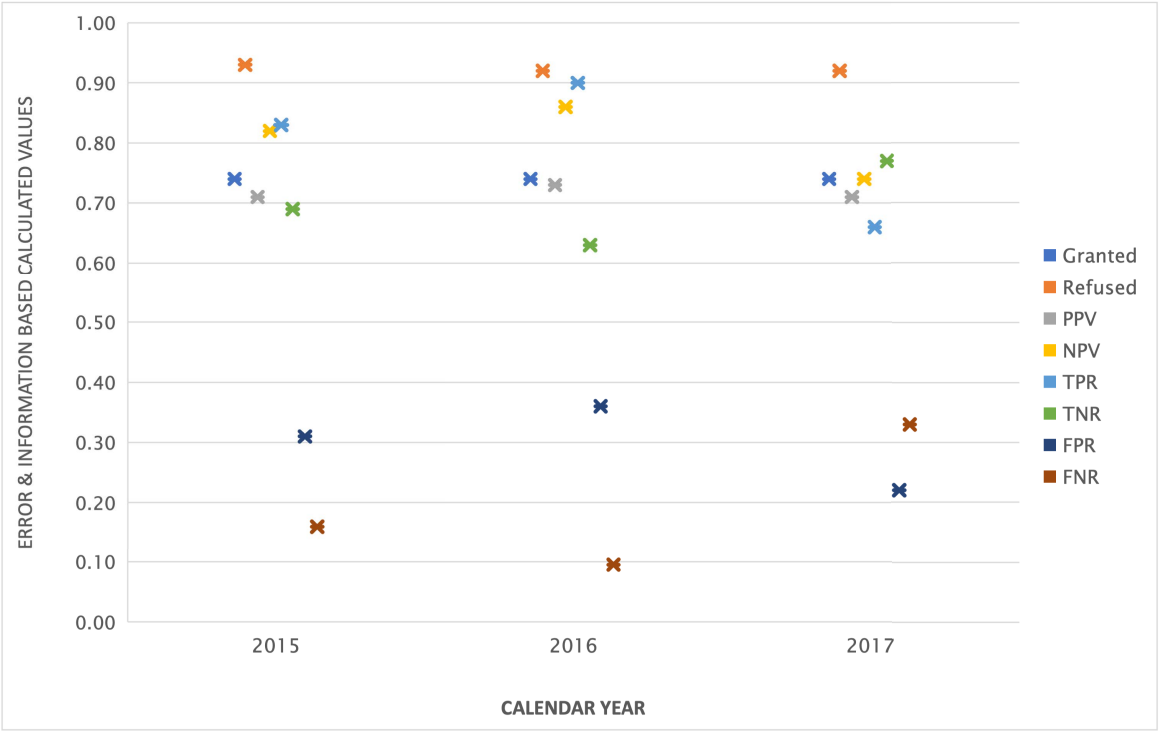

Figure 4.10: Comparison of probability distribution to the error-based and information-based values from *Bail-14*.

the results from this equation where 'Success' equates to bail granted and 'Failure' equates to bail refused.

Table 4.12: Success-Failure values comparison with Error-based and Information-based measures by year.

	2015	2016	2017
Success	0.74	0.74	0.74
Failure	0.93	0.92	0.92
PPV	0.71	0.73	0.71
NPV	0.82	0.86	0.74
TPR	0.83	0.90	0.66
TNR	0.69	0.63	0.77
FPR	0.31	0.36	0.22
FNR	0.16	0.10	0.33

Figure 4.10 details the values relative to the calendar years from Table 4.12, noting the categories of 'Success' refers to the probability of Granted and 'Failure' refers to the probability of Refused. It is apparent that the majority of the calculated values within each year share a uniformity with Success (Granted) and Failure (Refused).

Lastly, it is of interest to revisit the earlier anecdote regarding the inverse relationship between bail granted and the total numbers of bail matters finalised, although the focus being on a 3-year comparison on bail granted, breach of bail, and the PPV and NPV. As shown in Figure 4.11, the rates of defendants being granted and breaching their bail is relatively balanced. PPV and NPV data from *Bail-14* when measured against those categories – the probability

of compliance given at outcome of Granted, and the probability of non-compliance given an outcome of Granted – a corresponding pattern emerges.

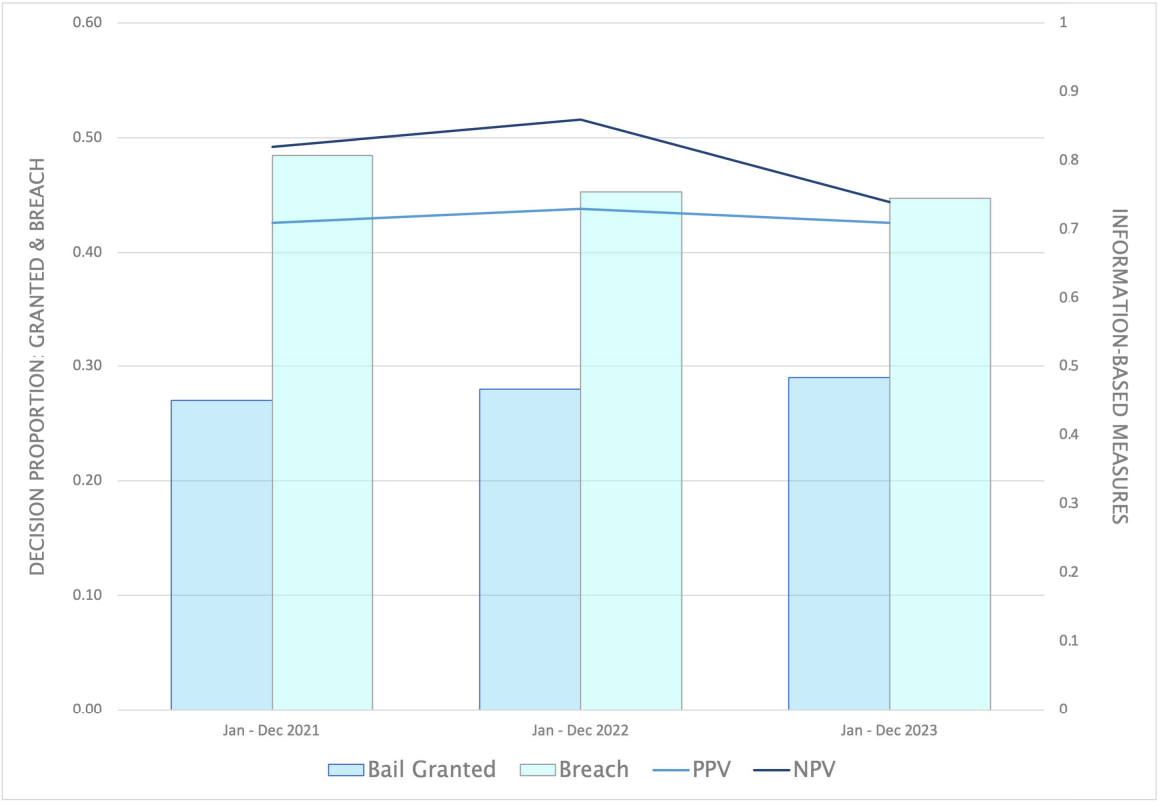

Figure 4.11: A 3-year comparison of information-based measures PPV & NPV to Bail Granted and Breach of Bail.

Limitations

This preliminary study was retrospective, meaning, the data was obtained from past cases and then considered against the outcome of the conviction or appeal. Along with the limits of retrospective nature of this study, was using open-sourced data. This meant, the quantity and quality of published judgments for bail hearings was a factor in data numbers. Firstly, judges (or other administrative authorities) are selective on what cases they want published, and bail decisions appear to feature less than other matters. Secondly, some cases are restricted and were not available at the collation time. Thirdly, while select bail decisions are published, the key variables are not contained within all written judgments and could not be used. Therefore, quality and quantity negatively affect the total numbers and consequently meant a low number for training and testing. More favourably would be for a greater amount of published judgments/decisions to accommodate greater data numbers for training, testing and validation.

Data collected from the written court judgments was a laborious exercise – searching for the keywords and phrasing required each judgment to be read in its entirety, then re-read at different junctures to ensure the key variables were registered accurately.

A limitation may be drawn on the data being specific to the State of NSW and not any other, given Victoria and Queensland both apply the Show Cause test in their bail legislation.

Lastly, a small yet valid criticism can be made on the RapidMiner-Studio software. The tree classifier option did not definitively nominating an algorithm, which draws on principles

such as transparency and explainability, as a criticism of predictive modelling more generally.

4.4 Discussion

Preliminary Study II applied two predictive models, a regression and tree-structured classifier. These models titled under the umbrella of the prototype *Bail-14*, were purposed for the Primary Study in Chapter 5, and through a syntheses of data and visualisations, real-world data will be cast to demonstrate real-world outcomes. Yet, equally or more fundamentally to this preliminary study was to examine the ethicality and morality of this undertaking in its entirety – from data collation to the analyses – as a principled measure of integrity before embarking on the primary study underpinned by *fairness* and *explainability*, of which is aligned with RO2.

Regarding FP and FN, these decisions could result in either a defendants' liberty being wrongly denied (false positive), or liberty being given incorrectly (false negative): values lower to zero are the ideal, yet in reality the values from these models were still dubious, and likely not to be in the range for acceptable policy [88].

Table 4.13: Predictor relevance order tabulated from Figure 4.7. Note: left-side tree (L) Criminal history node repeats although provides two different binary outcomes as expected; right-side tree (R) stopped at the sixth node.

Relevance (ascending order)	Predictors (L)	Predictors (R)
1	Seriousness of Offence(s)	Seriousness of Offence(s)
2	Show Cause (Y/N)	Pro-criminal associations
3	Criminal history	Seriousness of Offences(s)
4	Criminal history	History of violence
5	Bail non-compliance	Failure to appear/Flight risk
6	History of violence	Show Cause (Y/N)
7	Pro-criminal associations	-

Statistical prominence was shown in TsC on both these manipulated predictors, which is consistent with domain insight where a defendant's criminal history and offence seriousness level is expectantly taken into consideration when determining bail. Research conducted on decision-making in criminal justice also supported this contention [153]. Literature on predictive modelling in crime-related domains assessing risk and recidivism, for example, reflected the results of this preliminary study (although noting that wording variations within the literature do not detract from the premise on meaning). A meta-analysis on predictor domains measuring recidivism found criminal history was a prominent feature [91]. Scholarly work on AI and criminal justice assessing recidivism and risk reported predictor factors – of the types the same to – offence seriousness, current and prior violence, prior convictions, and failure to appear [28]. Notwithstanding, there could be a material reason why Seriousness of the Offence(s) in the TsC was a prominent attribute, that being, the 'gain criterion' having a preference for those attributes with a higher amount of values [150]; however, it supports the principle on *algorithmic*

fairness as to whether those predictors are rightly used in predictive modelling. In support of *explainability*, Figures 4.5 and 4.6 relating to the TsC identify one path to higher probability of bail being refused, the main predictor being Seriousness of the Offence(s). Similarly, Table 4.13 displays the relevance of each predictor in order as selected by the TsC in Figures 4.5 and 4.6, with the exception of two predictors (History of non-compliance (with court-issued orders); Threat/danger to the victim(s), public or others).

Analysing the data output from matrices in Tables 4.6 and 4.10 and Figure 4.3 on B-LogR, is resultative of similar conclusions in the literature and as demonstrated by the TsC: Seriousness of Offence(s) and similarly thereafter Criminal history, are consequential to predictive modelling on bail and other domains. While not to directly intended to compare one model against the other, it is nonetheless inferred by the EDA approach, to consider the performance or strength of the models on their predictive capability. B-LogR and TsC yielded some persuasive outcomes, notably the overall accuracy was in the 70th percentile. This level of accuracy is moderately good, given other research with the sole objective to devise a predictive model for court decisions resulted in 79 percent accuracy [154]. TPR in all B-LogR models produced a strong result, indicative of a good predictive model; and, the ROC values from Model 51-50 and its sub-models is some indication to its robustness. Comparatively the TsC was lower, concluding its predictive strength was secondary to B-LogR. This outcome was not reflected in research comparing logistic regression to tree models in predicting recidivism, having determined their decision tree performed better than the logistic regression approach [78]. There was not any preconceived objective to draw comparisons on the regression and tree classifier models to current bail statistics; however, as another manifestation of EDA, to compare and contrast results to results, became apparent.

Chapter 5

Primary Study: Survey of Participant Perceptions on AI and Visual Analytics

5.1 Background and Motivation

When handing down a decision in the NSW Supreme Court, the judge said no fewer than seven times the phrase "plain English" [155]. Contextualised, the judge wanted legal orders to be drafted in comprehensible language for the defendant, rather than legalese. While this judgment was not a bail matter, it exemplified the difficulty courts have when explaining orders to defendants, particularly those who are disadvantaged in some capacity. In another example, a former chief justice of Australia's highest court reflected on the utilisation of technology in legal matters, and the intellectual capacity of court-users (i.e., victims and defendants) being able to access and read electronic documents – he expressed concern that this would disadvantage those individuals who may already be at a disadvantage [83].

It is circumstances such as the two just mentioned which upon this study is fundamentally seeking to explore, with the prospect for policy change in the future. The contention here is around conventional communication for complex legal decisions, by simplifying the content by using imagery, charts and graphs, many of which have become familiar through weather forecasts and financial market analyses. Similarly, in the AI space, where analytical reasoning of complex data is required, visual analytics provides a means to simplification, and when extended to decision-making, then intelligibility of those decisions should be made easier. While in theory this appears acceptable, several relevant issues within this process emerge, namely ethicality and morality of AI-generated decisions. More specifically, how such decisions were generated, how such decisions are to be made intelligible, and who would be impacted by such decisions. On the intelligibility premise, a less explored yet fundamental construct within analytical reasoning is *visual literacy* [156], that is, being able to make intelligible the elements contained within a visual output.

In the institution of justice, matters on confidence in, and agreeing to, AI-generated decision-making, are imperatives for all relevant stakeholders and the public collectively. The former high court chief judge emphasised that the implementation of technology in courts must coincide with "public trust and confidence" [83]. Further on his AI discussion, the former chief judge

spoke about algorithmic fairness, in particular, on treating "subgroups the same regardless of their distinctions, like race or gender" and ensuring "statistical parity in the outcomes" which suggests the decisions would need to represent balance for all individuals and ensure any biases are omitted in the data [83].

Foundational to decision-making in this study are the ethical and moral principles of *fairness* and *explainability* and *deontology* and *teleology*, which were detailed in Chapter 2.

5.2 Statistical Methods

Alluded to in previous chapters was an expectation for statistical research to observe historical practice mandating theories and hypotheses to precede causal explanations through hypothesis testing or confirmatory research and analyses [157][158]. Articulated in the literature was a dissuasion away from historical practice of "statistical significance", encouraging an alternative approach, which is more about conceptualisation on a subject matter through inquiry and building on those outcomes without having preconceived notions [159]. The framing of this approach is Exploratory Data Analysis (EDA): non-experimental and descriptive, underpinned by mixed-methods entailing both quantitative and qualitative methods [160][161].

EDA can be understood as a nuanced approach to statistical analyses. Coherently expressed in the literature as a decades old movement away from conventional methods, to one of discernment and inquiry in data exploration and visual analysis [157], both numerically and visually [136]. EDA is not constrained by a specific definition, which is beneficial in itself allowing for flexibility in statistical analysis and findings. A study does not necessarily need to begin with hypotheses, and aim for statistical significance, rather it is dependent upon the researcher's understanding in the subject area and the potential outcomes [158][162]. Applying EDA is not antithetical to confirmatory data analysis, it however precedes the confirmatory process and assures potential discoveries that may not have been realised deductively [158][162]. Subsequent to EDA, this study will apply *descriptive statistics and analysis*. This approach engages constructs and variables, operationalised to connect the said constructs and data in a meaningful way [123]. Data analysis is undertaken through univariate and bivariate methods [123][126]. The survey was produced with the benefit of Research Electronic Data Capture or "REDCap" [163][164].

Composed of eight questions and two vignettes, the Bail-14 survey has its own defined constructs. A construct is described as an abstract or preconceived phenomena of the researcher pertaining to research matters of interest [165]; and where there is greater complexity within the phenomena, seen as a multi-dimensional construct, comprehensible descriptions are required on operationalisation, such as the measures and analysis methods used [123]. The overarching construct of the survey is *Intelligibility of AI in decision-making*, which denotes the *perceptions or opinions of participants on AI-generated decisions, conceived by analytical reasoning through visualisation, explainability and visual literacy.* As will be detailed, each construct is implicitly connected to a variable, which supports their operationalisation, and this will be extrapolated for each question and vignette in the survey. Scales for each question or vignette are noted.

Constructs are operationalised by the following scales. Visual Analogue Scale (VAS) was selected for its functionality, performance and representation of values [166], and increases the options for statistical tests [167]. Nominal Scale (NS) was appropriate for categorical responses [123] (i.e., check boxes), and to be noted was combined with theoretical underpinnings of Semantic Differential Scale (SDS) – that these responses were designed in essence of SDS as it effectively measures perceptions of participants [168]. A Binary Scale (BS), simply applied for yes/no responses [123].

Constructs

Contrastive or Counterfactual Explanations in AI-generated decisions or predictive modelling is one approach to provide intelligibility [169][170]. Contrastive explanations aid the "explainee" by not having to necessarily interpret a decision, rather it provides an alternative decision as a way to explain or justify a given decision [48][56]. The first section in the survey "H-Bar" (labelled from its namesake Horizontal Bar), asked participants to what extent did they agree or disagree to a statistical decision cut-off point at "0.40" – is X justified given the outcome of Y – where one defendant scoring 0.41 would be granted bail while the alternative option of refused bail was imposed on another defendant with a score 0.39. A 5-point VAS was applied to measure 'Agreement': strongly disagree (1) to strongly agree (5) with a neutral value (3). A secondary question was set (using the "Branching Logic" feature in REDCap) – it asked if they would reconsider their response on a decision cut-off point after learning of the offence upon which the defendant was bailed. The cut-off point was denoted by a vertical dotted line, as it was beneficial to ensure participant's were drawn to this particular value [171]. A 2-point BS was applied: Yes (1), No (0).

Individual fairness refers to one defendant being treated equally despite any individual characteristics that are perceived to categorise or differentiate from other defendants. It imparts the premise that similar people are treated similarly and receive similar decisions [48][172] – as such "similar people" refers to defendants. Guiding this construct in the survey was a Sankey visual (see Appendix A.3 and A.4) and a question on individual fairness, suggested by age, gender and impairment, as possible determinants. A 5-point VAS was applied to measure 'Confidence': not confident at all (1) to extremely confident (5) with a neutral value (3).

Group fairness refers to all defendants being treated equally despite any group characteristics that are perceived to categorise or differentiate from other defendants. It is the equal and proportional treatment of groups with varying characteristics to be treated equally [172]. Guiding this construct in the survey was the TreeMap visual (see Appendix A.5) and a question group characteristics with possible determinations on race, culture and religion. A 5-point VAS was applied to measure 'Confidence': not confident at all (1) to extremely confident (5) with a neutral value (3).

Visual literacy in this context encompasses a persons intelligibility to conceptualise and interpret images borne from AI-generated data. It was said to be an essential cognitive attrib-

ute in the development of an informed society [156]. Bail-Tree was the title of fourth section, named as such for the tree classifier was the visual reference, and developed on the data from Preliminary Study II. A 5-point VAS was applied to measure 'Difficulty': difficulty (1) to easy (5) and a neutral value (3). Secondary questions were set (using the "Branching Logic" feature in REDCap) following selected VAS responses, and measured using a NS.

Moral Principles made up the last section of the survey, labelled under "Situational example" using a vignette and two postulates corresponding with Teleology and Deontology, and to be reasoned analytically using the Circle Pack visualisation. Additionally, this construct within the literature contends a juxtaposition on Deontology and Teleology of an inverse relationship or their not being any correlativity – therefore to remedy this contention, any relationship or lack thereof will be tested mathematically. Variables for both vignette's were 'Agreement', measured by a 5-point VAS: strongly disagree (1) to strongly agree (5) with a neutral value (3). Secondary questions were set asking participants for the reasons to their initial response (using the "Branching Logic" feature in REDCap), measurable by NS.

Visualisations

Selected visualisations were created through two open-sourced programs: *inetSoft Visualize* (H-Bar, TreeMap, Circle Pack); and *RapidMiner-Studio* (Sankey and Bail-Tree). The labels or titles of the visualisations were selected due to their relevance with the software's own title of each respective visual, but also to make it easier for participants to identify each visual in accordance with the questions or vignettes. Purposely, visualisations were created to be aligned with the relevant construct.

The following are descriptions of each visualisation as it appears in the survey. All visualisations have interactivity available, although for the purposes here, screenshots are presented as exemplars (refer to Appendix A). The interactivity is important to note for it would be necessary where there are many decisions grouped together and one or more users required specific detail.

Visualisation 1 is termed "H-Bar", which is an interactive bar graph with the standard two axes of 'x' and 'y' where each horizontal bar represents the defendant and the value in iterative predictive output (see Figure A.1); its interactivity allowed for each individual bar to be highlighted with the pertinent information specific to that defendant. Figure A.2 shows the first interactive step to highlighting an individual case, in this instance defendant with the code-identifier 'CGF': illuminated by a red bar, a predictive value of 0.41 and literal decision over the bar indicating bail "Granted" . Note the dotted line in Figure A.2 is not part of the original visualisation and inserted for illustrative purposes only to demonstrate the cut-off or threshold between the decisions of Granted and Refused.

Visualisation 2 is called a Sankey diagram: its multicoloured 'channel' like display moves from left to right; each channel represents a defendant and is linked to the predicted decision on the right side of the graph (Figure A.3). Its interactivity allows for a selected defendant to be highlighted detailing the decision iteratively, and as shown in Figure A.4, defendant with code-identifier 'KJC' was selected and illuminated by a light blue channel, leads to the bail

decision of Refused.

Visualisation 3 is identifiable by it's multicoloured brick blocks and its namesake (yet not to be confused with a tree classifier). The parameters are conditioned in such a way that the predicted decision is assigned a size and colour, and the defendant's code identifier is labelled on the brick and the numerical output in the legend on the right screen side. This can be viewed both in a static or interactive format (refer to Figure A.5 for static screenshot).

Visualisation 4 is the Circle Pack and as its name suggests, it is a group of circles bound to one another; each individual circle represents a defendant and referencing the legend, indicates a decision of granted as correct (dark blue) and a decision of refused as fail (light blue). Displayed as a static screenshot in Figure A.6, it can also be interactive.

Visualisation 5 is a tree-structured classifier and as a different algorithmic model and output to the previous visualisations, requires a more detailed description."Bail-Tree" displays the relevant predictor classifier in sequential order, and in between each classifier is a numerical value, with the exception of Show Cause (Yes or No) – "Seriousness of Offence(s)" is the most relevant predictor in Bail-Tree. The numerical value is assigned through examination of the defendant's case, which ultimately leads to the decision. As a process being applied to any defendant, the categories and numerical values are fixed, for instance, the branch between Seriousness of Offence(s) and Show Cause indicates a numerical value of >2.5 – this value meant that any defendant charged with an offence rated above 2.5 would follow that branch to the next most relevant predictor, which in this example is Show Cause, and if the response was "Yes", the next most relevant predictor, being Criminal History, would ultimately result in the outcome of "Refused". The same process would be undertaken for each defendant. Even though Bail-Tree is interactive, for descriptive purposes, Figure A.7 is a screenshot of the pathway to determine the decision for defendant with the code-identifier 'KJC' that resulted in a decision of bail "Refused" (this decision is consistent with the logistic regression model, displayed as a Sankey diagram in Figure A.4). Therefore, the seriousness of the offence KJC was charged had a value greater than 2.5; the offence was classified as Show Cause; and KJC's criminal history had a numerical value greater than 2.5. Piecing this together, KJC was refused bail.

Participants and Apparatus

Participants

Recruitment occurred at the following locations: Lismore and Byron Bay (NSW); and Brisbane, Beenleigh and Southport (Qld). Each location is commensurate with one day. A total of 11 participants consented to undertake the survey: 2 surveys partially completed, 1 participant consented to the survey by scanning a designated QR code with a mobile phone and although it was activated it remains incomplete. In total, 9 observations formulated the data. A preamble on anonymity and confidentiality was provided – identifying information of any kind was neither asked nor required – a part of the research design to provide more reassurances to participants on anonymity. Participants were only provided with basic written information or prompts accompanied by the visualisations. The selection of these visualisations and their configurations were intentionally modest by design, to minimise complexities, as was the approach in another user-study [109].

The reasons for the survey being directly targeted at a specific population were: firstly, as decision-making in the justice domain generally has had or would have some direct impact on this population, it was more relevant to seek their perceptions; secondly, there is also a statistical advantage in sampling probability of this population. While the study design was initially purposed for direct engagement with each participant while they completed the survey for interactive reasons, it was later determined *ad hoc* that other recruitment methods should be explored. For example, a hyperlink and QR code were activated, where for example, the survey could be conducted on a handheld device at a later stage at the participant's choosing.

Apparatuses

Figure 6.1 is an image of the apparatuses used for in-person (direct) survey: iPad, with a 25.9cm/10.2 inches touchscreen display – 2160-by-1620-pixel resolution at 264 pixels per inch – aided by the touchscreen for improved navigation and interactivity. Supplementing this was a MacBook Air, with a 34.54cm / 13.6 inches LED display – 2560-by-1664 native resolution at 224 pixels per inch – utilised for greater image clarity of visualisations in support of the iPad.

Figure 5.1: Survey apparatuses – MacBook and iPad

Survey software

Study data were collected and managed using REDCap electronic data capture tools [163][164]. REDCap is a self-operated program, which allows for the user to develop their own survey design. It is a secure, web-based software platform designed to support data capture for research studies, providing: (1) an intuitive interface for validated data capture; (2) audit trails for tracking data manipulation and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for data integration and interoperability with external sources [163][164]. Participant consent was requested prior to the commencement and at the conclusion of the survey, the conclusion is supported by the e-Consent framework [173].

5.3 Results

This survey was designed to complement the Bail-14 predictive model seen in the previous chapter, to provide real-world data to participants, with the intention of ascertaining their perceptions on AI-generated decisions through visual analytics.

Table 5.1: Participants perceptions on Contrastive Explanations in response to H-bar.

Response variable	Response (%)
Agree	44.45
Disagree	22.22
Unsure	33.33

Participants were informed in the introductory stages of the survey, prior to responding, that the defendants and cases are real, consequently the data contributing to the visual analytics, is also real. Even though this was not a measurable asset, anecdotally it gave a sense of authenticity to the survey, that may not have otherwise been gained from fabricated data.

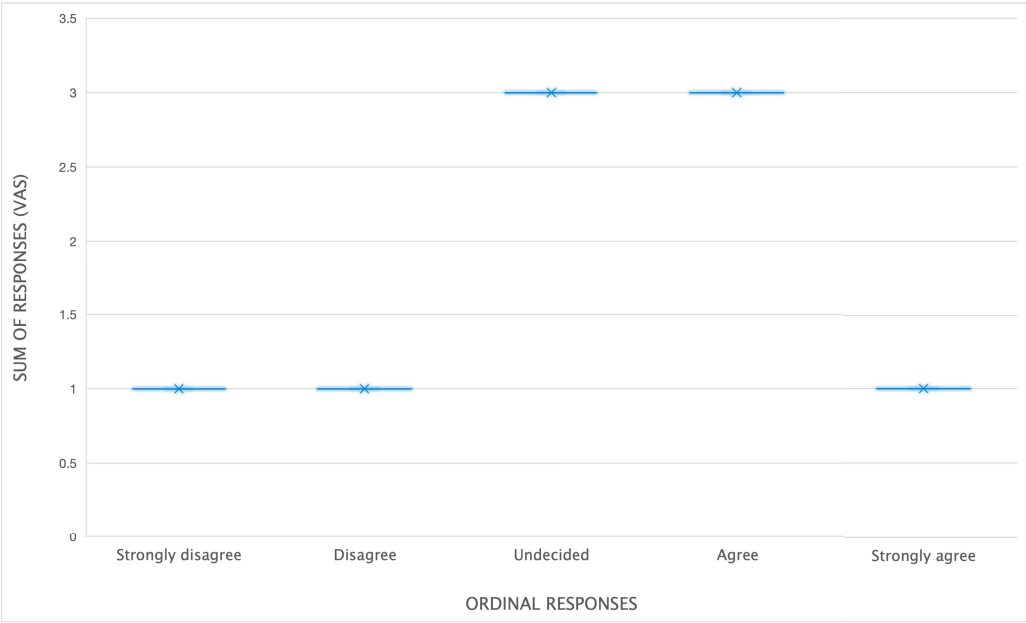

Figure 5.2: Participants perceptions of H-bar in response to Contrastive Explanations.

The structure of the results presented in this section are done so in the same order as the questions and statements in the survey. Visualisations used in the survey are available in Appendix A.

Table 5.2: Participants perceptions – Change (%) on Contrastive Explanations to H-bar visualisation.

	Would reconsider response	Would not reconsider response
Participant response	77.28	22.22
Change from Agree	75	25
Change from Unsure	100	0
Change from Disagree	50	50

Beginning with the Contrastive Explanations construct and 'Agreement' variable, participants were shown two versions of the visualisation H-Bar: the first displayed the collection of decisions from the predictive model; the second displayed the collection of decisions in the same, with the difference being two horizontal bars were illuminated to identify the two contrastive decisions. Additionally, a dotted vertical line was displayed on the second H-Bar visual, at the 0.40 threshold. Participants were asked to consider it fair or unfair to have a decision cut-off point at 0.40.

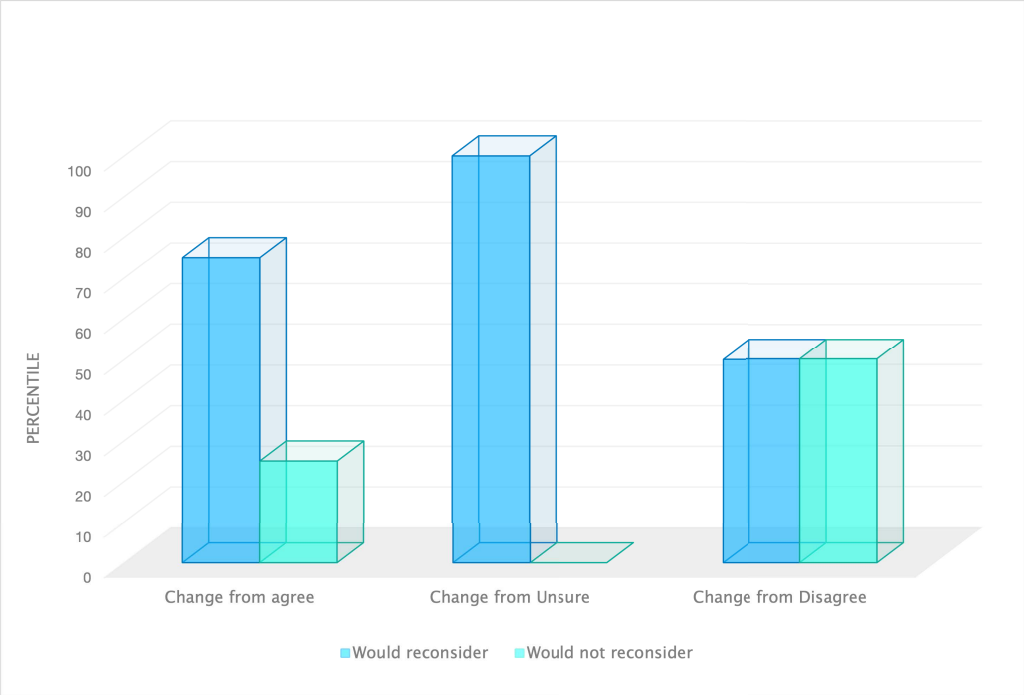

Figure 5.3: Participants perception: Change on Contrastive Explanations (H-bar visual).

Table 5.1 indicates that 44.45% of participants considered it fair to have a cut-off point, 22.22% disagreed, and 33.33% remained unsure. A supplementary question was asked of participants, if they would reconsider their response given new information on the defendant who was granted bail, having been done so after the offence "Shooting with intent to murder".

Note: confidence variable for Group fairness construct was split 'Confident' (3) and 'Extremely confident' (3).

Table 5.3: Participants' perceptions on Sankey and TreeMap visualisations (%).

Construct	Confident	Not confident	Unsure
Individual fairness (Sankey)	55.56	11.11	33.33
Group fairness (TreeMap)	66.67	11.11	22.22

As shown in Table 5.2, 77.78% of participants indicated they would reconsider their response, when 22.22% would not. Table 5.2 also identifies that all participants who initially said they were unsure, would change their initial response, and 75% who had agreed would change. Of those who had initially disagreed with the cut-off, the most notable change was that half would reconsider their response.

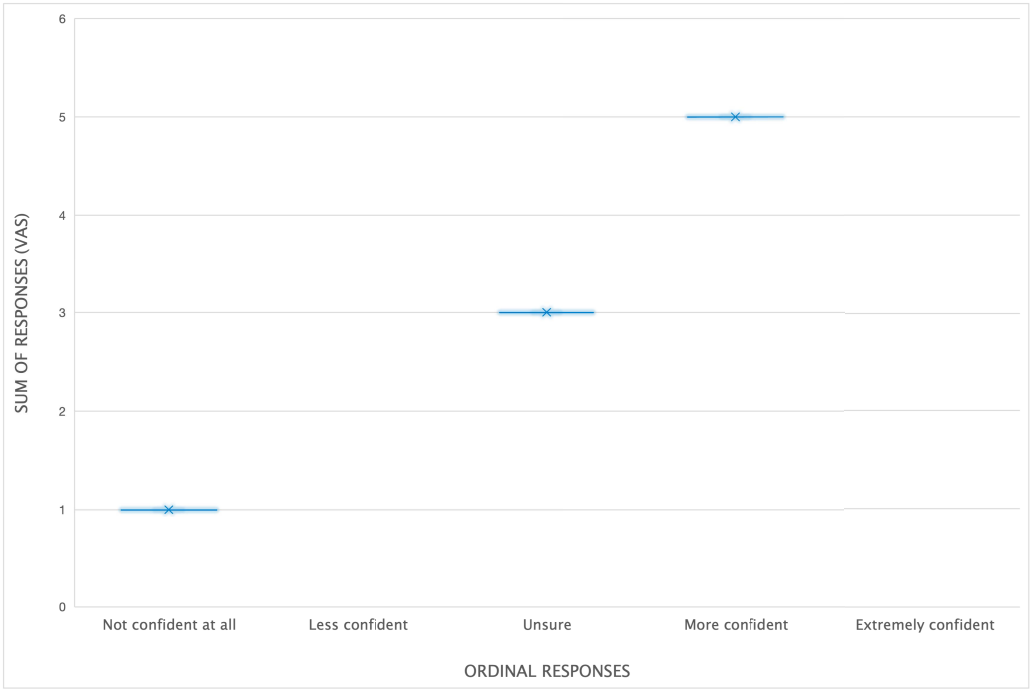

Figure 5.4: Participants' perceptions on Sankey visualisation in response to Confidence variable.

Confidence as a measurable variable was examined under the constructs of Individual fairness and Group fairness, the former accompanied by Sankey visual and the latter accompanied by TreeMap visual. Participants were asked to perceive their level of confidence in determining individual characteristics (e.g., age, gender, impairment) and group characteristics (e.g., race, culture, religion) based on these visualisations.

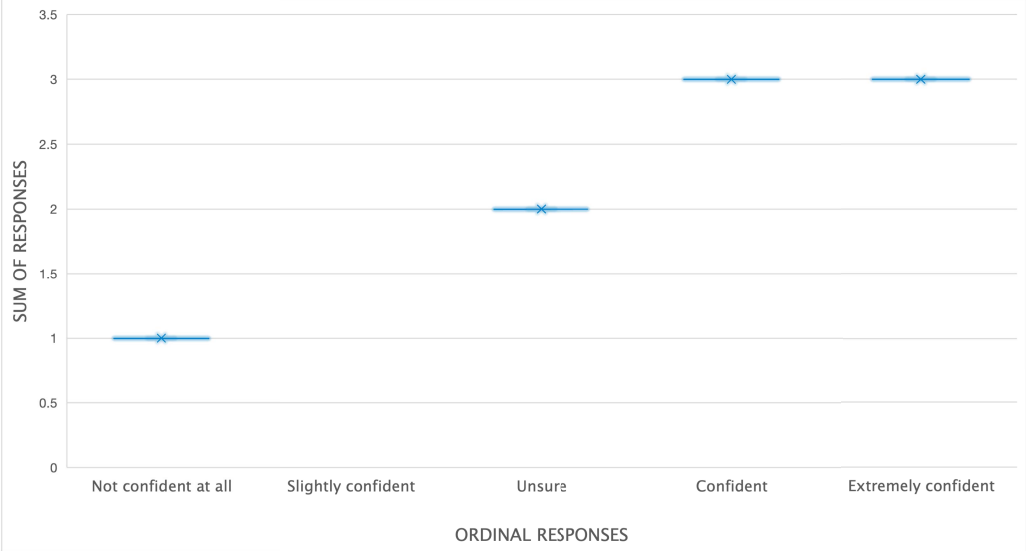

Figure 5.5: Participants' perceptions on TreeMap visualisation in response to 'Confidence' variable.

As shown in Table 5.3 under the construct for Individual fairness, just over half of the participants were in the confident range, and around a third were unsure. Whereas for Group fairness, two-thirds were within the confident range and 22% were unsure. 'Not confident at all' responses for both fairness constructs resulted in 11.11%.

Table 5.4: Participants' perceptions of Bail-tree process on 'Difficulty' variable.

Variable	*N*	Reasons (response no.)
Very easy/Easy	6	Visual appears straightforward (3) Clearly identifiable (1) Logical process (4)
Unsure	2	n/a
Very difficult/Difficult	1	Visual appears confusing (1)

Bail-Tree was the next visualisation presented to participants and they were asked to gauge their perceptions based on the process required to determine a defendants' bail. Participants were prompted with a basic iteration on the predictor variables and numerical values in direct reference to a defendant as a necessary step to build on for the accompanying question.

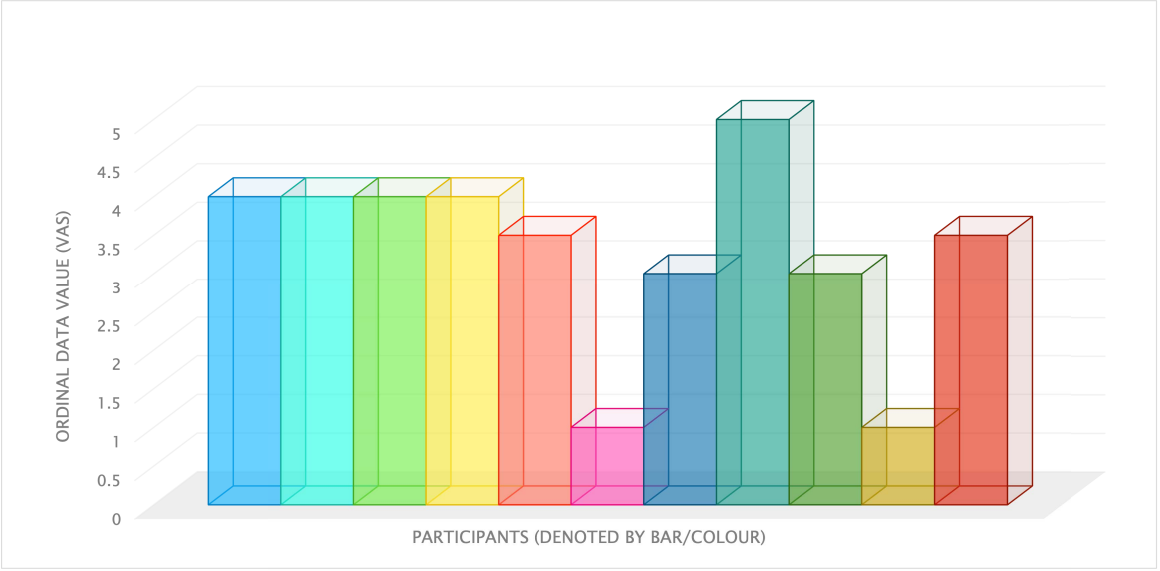

Figure 5.6: Participants' perceptions on Bail-tree in response to 'Difficulty' variable.

As Figure 5.6 displays, each coloured bar denotes a participant, and the majority of responses determined the process to be in the 'easy' range (VAS values 4 and 5), with the remainder being unsure (VAS value 3) or having found it very difficult to comprehend (VAS values 1 and 2).

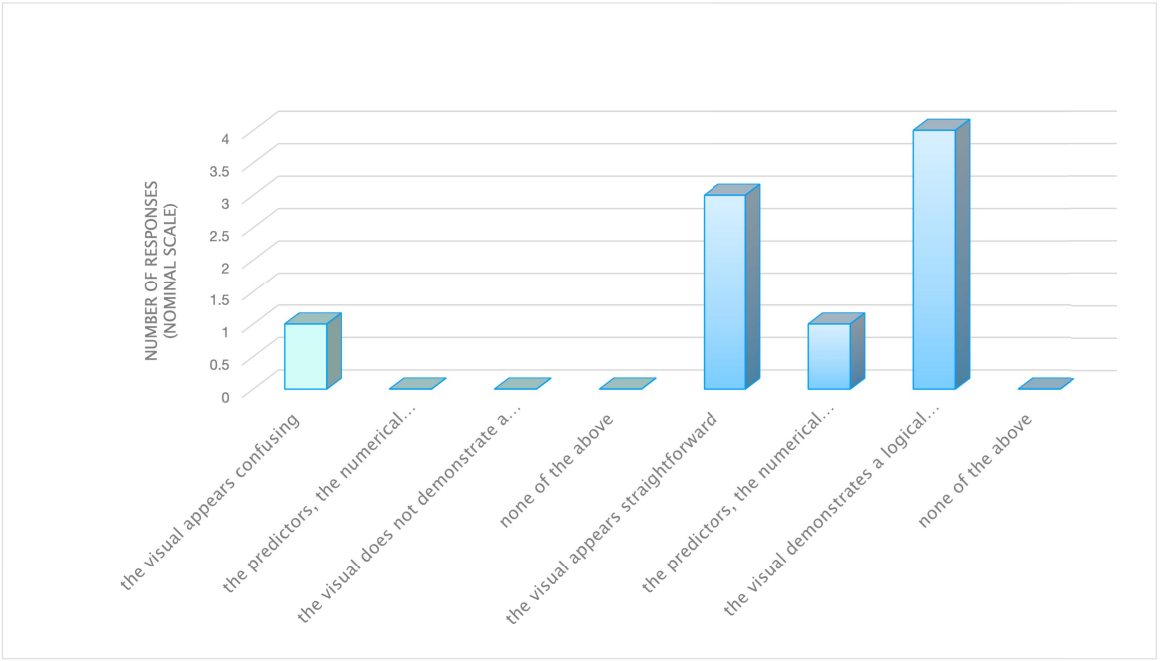

Figure 5.7: Participant's reasons for responses to Bail-Tree visual.

Supplementary questions presented to participants canvassed the rationale for the responses they chose. As detailed in Figure 5.7, the reasons registered mostly on Bail-Tree being in the "Easy" scale range were due to "the visual appears straightforward", "the visual demonstrates a logical process", and "the predictors, numerical values, and decision, are clearly identifiable". In the opposing range of "Difficult" one item was checked, which stated "the visual appears

confusing", corresponding with the difficulty variable.

Table 5.5: Participant response to Teleological & Deontological construct (%).

Construct	Agree	Did not agree	Unsure
Teleological	55.55	33.33	11.11
Deontological	22.22	44.44	33.33

Note: item imputation with arithmetic mean (x) due to one partially completed survey.

Concluding sections of the survey contained the situational example and vignettes, which coincided with the Circle Pack visualisation (refer to Appendix A.6). Participants were asked to read the situational example in conjunction to viewing the visualisation – the situational example addressed a moral scenario on AI determining bail. Following this, participants were asked to gauge their perceptions on these elements in response to teleological and deontological postulates.

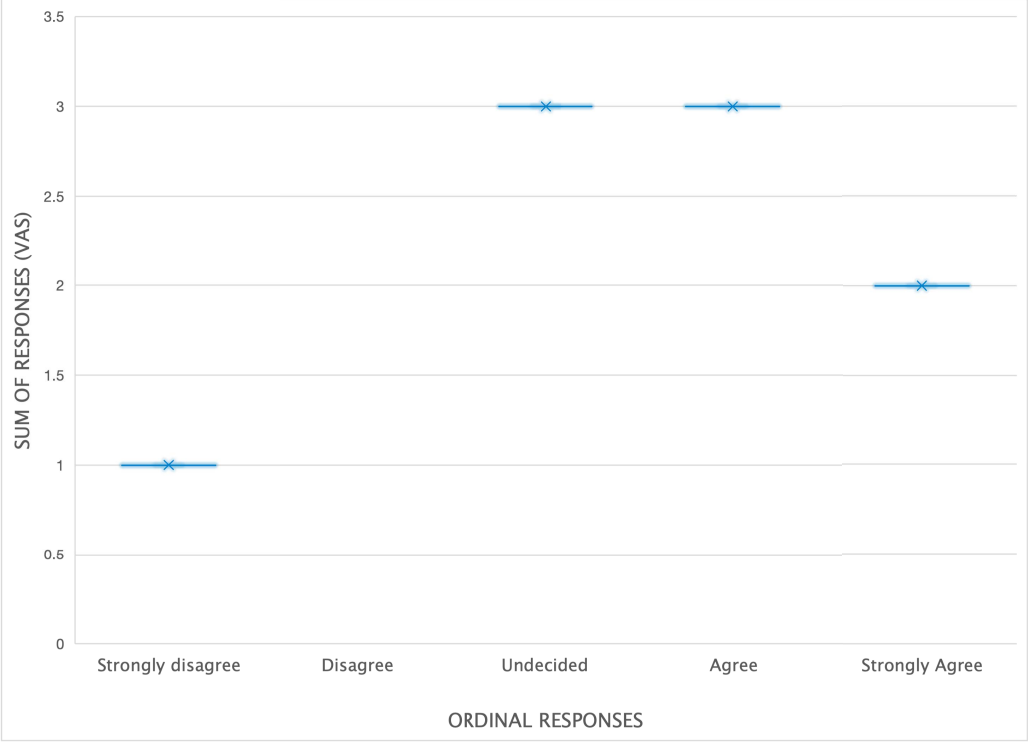

Figure 5.8: Sum of participants' responses to Teleological postulate.

As shown in Table 5.5, just over half of participants (55.55%) agreed with the teleological premise, while one-third (33.33%) were unsure, and 11% did not agree. Quite differently, participants were more unsure about the deontological premise (44.44%), one-third did not agree (33.33%), and 22.22% agreed.

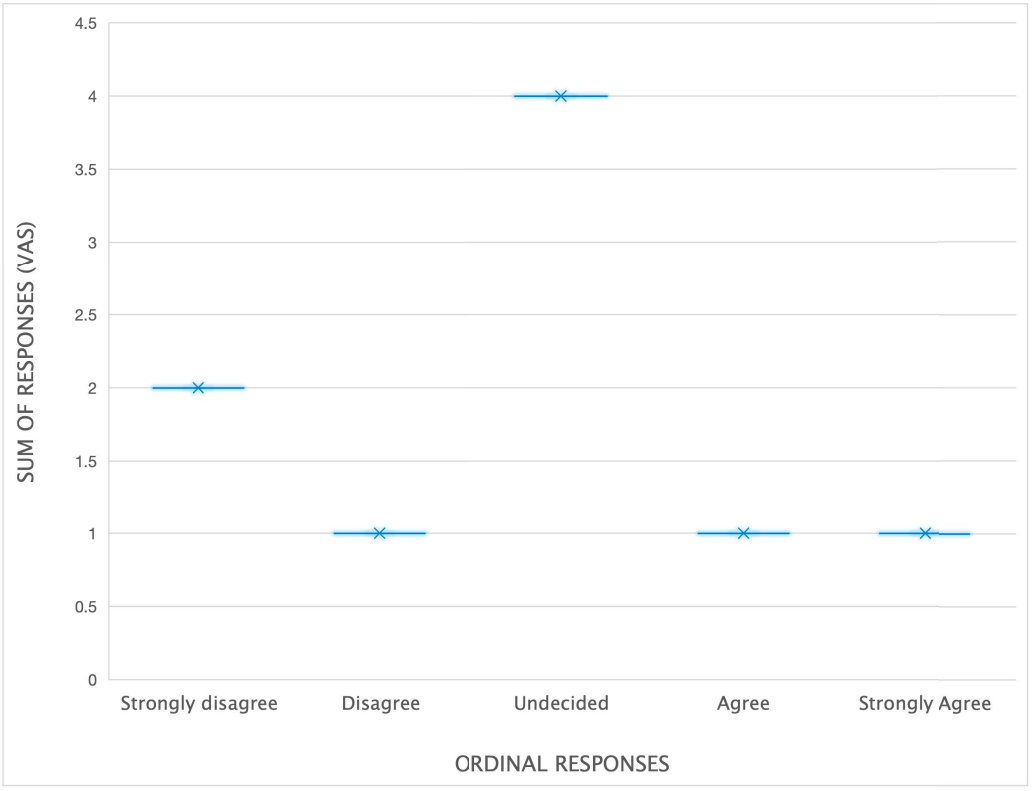

Figure 5.9: Sum of participants' ordinal responses to Deontological postulate

Statistical distributions were calculated by central tendency based measures, and as expected they were relatively unremarkable, itemised in Table 5.6 and Figure 5.10, and summarised in the following. The Teleological premise showed a mean of 3.5 with a standard deviation of 1.22. The Deontological premise showed a mean of 2.75 and a standard deviation of 1.30. Interpretively, the arithmetic mean or average, standard deviation (*SD*), and variability or variance (*Var (X)*) as shown in Table 5.6 and Figure 5.10, are calculations indicative that the participants in their responses did not deviate distinctly from the middle (or median) of being "undecided" on the two moral postulates.

Table 5.6: Outcomes of central tendency measures of Teleology and Deontology postulates.

	Mean	Median	Mode	*SE*	*SD*	*Var (X)*	ρ
Teleology	3.5	3.5	3	0.41	1.22	1.5	-0.28
Deontology	2.75	3	3	0.43	1.30	1.69	-0.28

To test the strength of association between the deontological and teleological constructs and ordinal variables, the Spearman's Rho (ρ) appeared to be the most adaptable [126]. Essentially, this test is to measure or observe the strength and direction the constructs and variables move, and is done so based on values ranging between +1 and –1: a positive direction towards +1 indicates the two variables have a strong positive relationship; a negative direction towards –1 indicates the two variables are negatively correlated yet still can have a strong relationship; and, the closer to zero indicates a weak or non-existent relationship [174]. As the association is seeking to observe a one-way direction only, a one-tailed test is applied [175]. Therefore, the correlative strength and direction of Teleology and Deontology resulted in the same value

of (ρ)= –0.28, interpreted as there being a negative correlation between the two constructs and having an inverse relationship, although as the values are close to zero, it is a reasonable assessment that the relationship is weak (refer to Table 5.6).

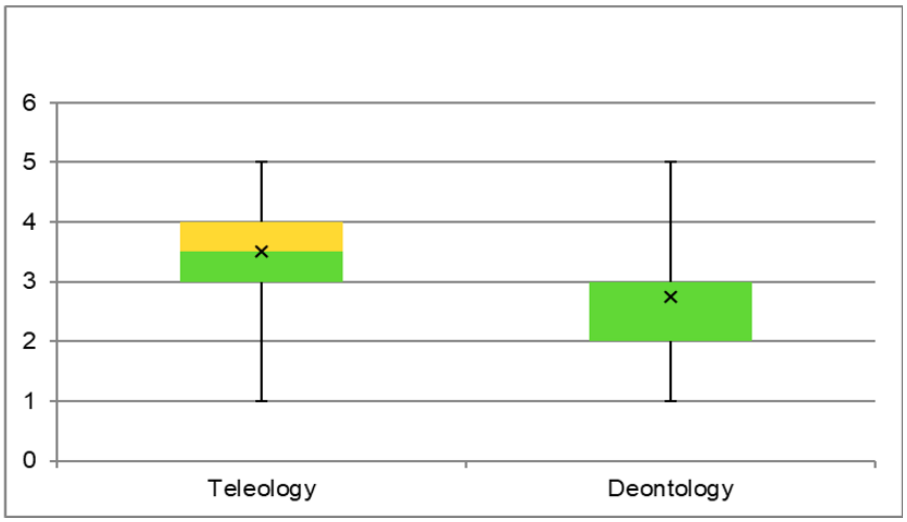

Figure 5.10: Central tendency measures of the Teleological and Deontological responses to situational example postulations.

Based on a "situational example" and Circle Pack visualisation, participants were asked questions based upon their rationale as to why they disagreed or agreed with the statement, and the response was then further separated into secondary nominal responses as to why they chose that response. Figures 5.11 and 5.12 displays the participant response distributions (including not providing any response). Regarding the Teleological postulate based on the situation example, one participant who disagreed reasoned this as being unfair for AI to decide bail for one defendant when measured against all other defendants. Participants who agreed with the Teleological postulate, reasoned this on the basis that failures made by AI are acceptable, despite benefits to the greater population (1); a good outcome is the right outcome (2); with one participant not selecting any of the reasons provided, and four not responding at all.

Using the situational example as the reference, the same process was given to participants in response to a Deontological postulate. Of those participants who disagreed, two said a defendants legal and moral rights should not override benefits to the greater population, and one selected none of the above. Contrastingly, those participants who agreed with the Deontological postulate, did so for the reason that a defendants legal and moral right should override benefits to the greater population (2), where the remainder did not provide any response.

Limitations As the approach to this study was based on EDA, there were not any predetermined expectations, and any identified limitations are made generally. Notwithstanding, *time* when leveled within several contexts, was overwhelmingly a considerable limitation, as discussed in the following.

Firstly, time in the sense of approaching prospective participants (e.g., defendants and/or

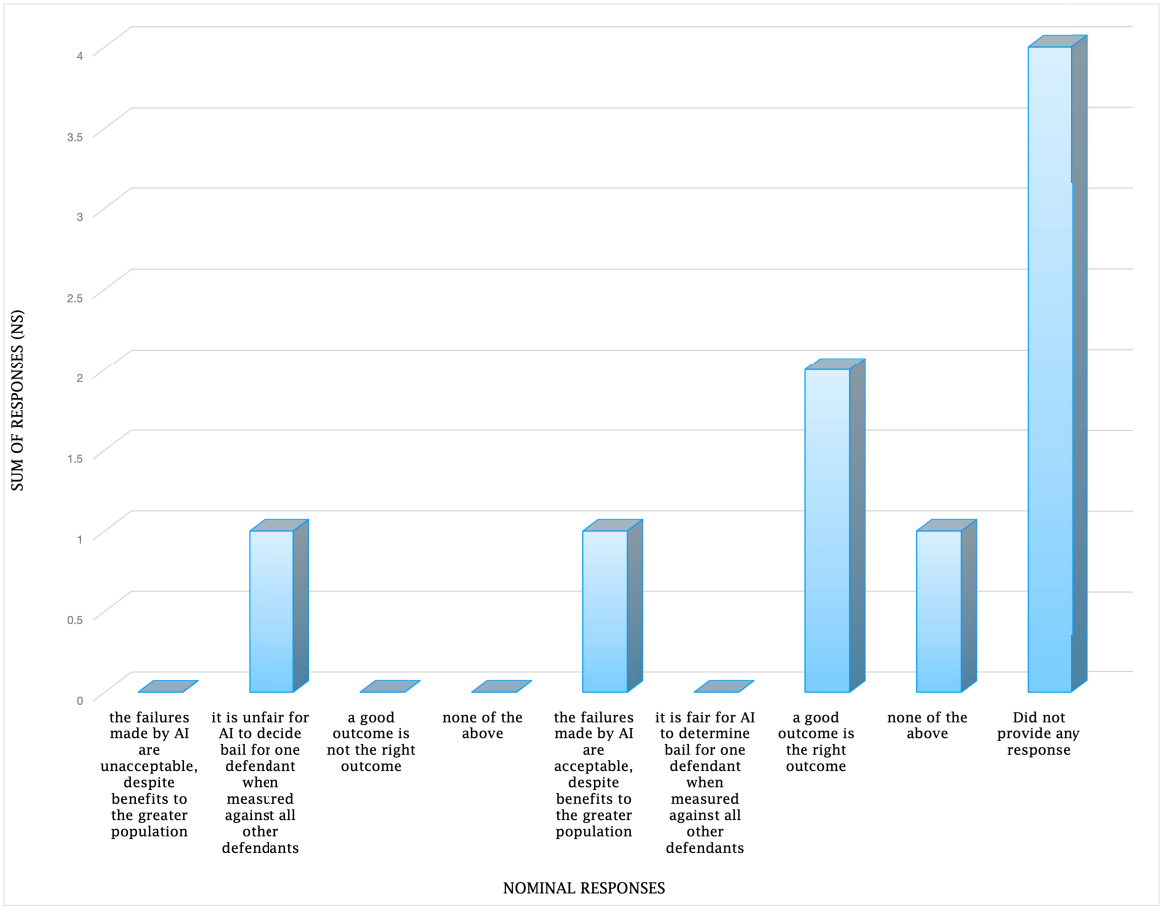

Figure 5.11: Participants responses to Teleological postulate secondary question.

victims) prior to their matters being heard in court seemed to have been met with apprehensiveness. Participant engagement was challenging in all locations – attempting to recruit and survey at court houses was ambitious, and not without its obstacles.

Secondly, in the context of allowable time to complete the survey, interruptions such as discussions with legal representatives, court staff, and/or being called into the courtroom, was an inhibiting factor. Engaging individuals who may be defendants or victims in criminal matters might be best undertaken through different approaches, such as liaising with justice departments or private legal representatives to recruit, where the anxiety factor may be minimised.

Thirdly, low participant numbers equated to lesser than anticipated data collection. More time allocated to the survey might have been improved participant numbers. Even though this study did not have as its objective to sample a population as a evaluation metric in a conventional sense, it could have allowed for that option for a more robust basis for future research. For instance, questions on explainability – to gauge what was and was not understood in the visualisation. In a statistical sense, the lower sample number leads to questions on reliability and validity, despite the literature supporting the integrity of small sample sizes.

Fourthly, as a factor in the previously discussed time limitations, there were several circumstances beyond the control of this researcher. Approvals from external organisations were relatively slow due to required permissions from departmental management on safety and ethics, which was also dependent upon ethics approvals, which was also delayed for various reasons.

As signalled earlier, public areas within court environs are not the most appropriate location

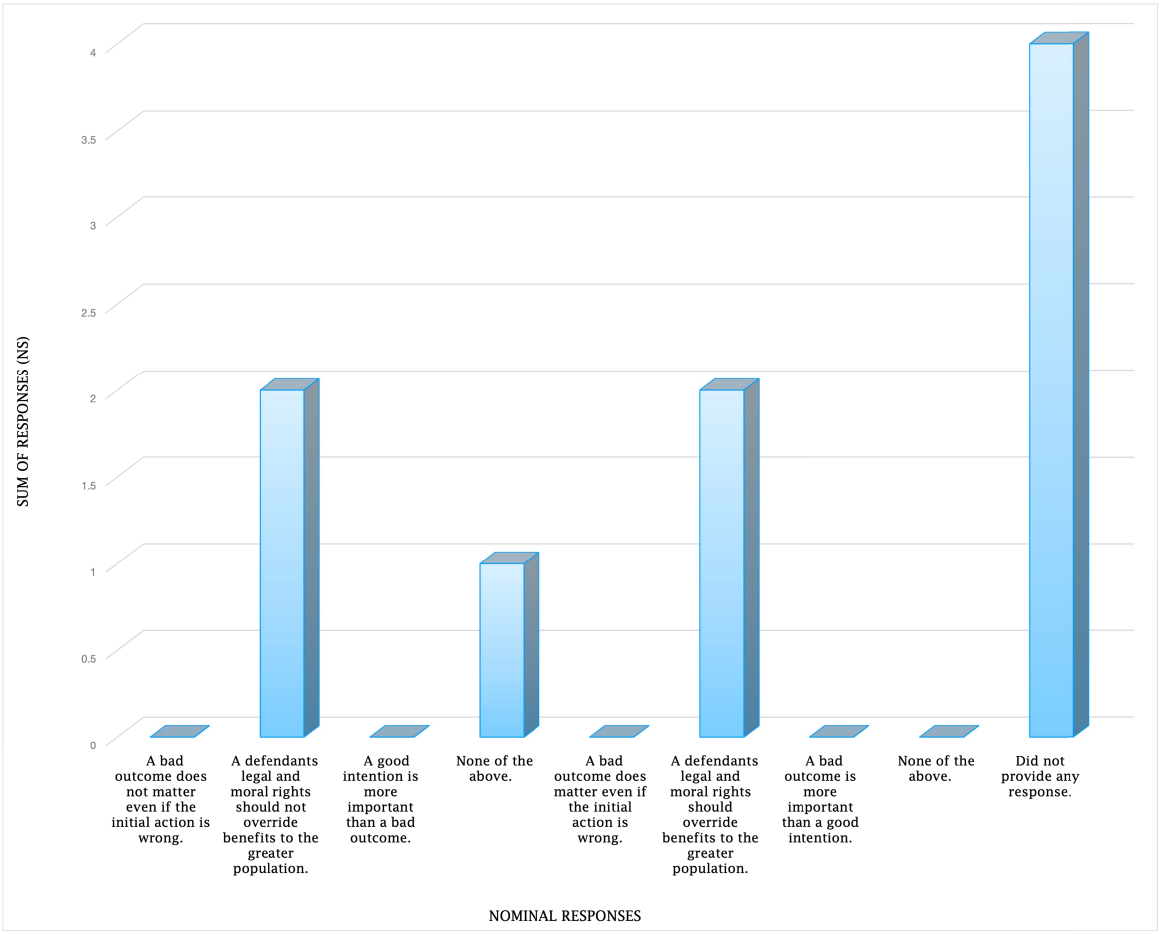

Figure 5.12: Participants responses to Deontological postulate secondary question.

to conduct surveys. The proximity to potential participants, for example, overhearing any discussions between researcher and participant may have had an impact on responses from an active participant and/or the prospective participant. Countering this circumstance led to a smaller participant pool to recruit from, as many prospective participants had to be excluded.

Even though there were different locations chosen to conduct the surveys, it is relevant to note two matters on generalisation of the results. The first matter is on the perceptions of participants in different State legal jurisdictions, and observed that this can have a positive or negative impact. A second matter to be noted is that regional courthouses, with the exception of Brisbane City, may generalise perceptions for stakeholders from cities.

Software access was considered a manageable limitation. Despite being freely available, there were some restrictions. Many of the visual analytical and data mining programs require paid subscriptions, and without this access level, it significantly delimits functionality. In a more specific evaluation, the quality of the visualisations themselves were not measured, and this potentially could have had some influence on the participants responses.

5.4 Discussion

Outlined in RO2 and RO3, summarily speaking, was to gauge the perceptions of court-users and other stakeholders on fairness and explainability, as ethical mitigators to AI-generated bail decisions, and resolve moral queries on deontology and teleology to AI-generated bail decisions.

Generally, results from the survey reflected a combination of uncertainty and pragmatism: the uncertainty, while not being dominant, was evident from the data; the pragmatists feature more so on progressive notions in decision-making. This was similarly reflected upon in the literature:

> Stakeholders can also be sceptical and reluctant to adopt AI systems without the ability to explain system decisions, even if the systems have been shown to improve decision-making performance [169].

This Primary Study featured various visualisations and iterative prompts, all of which could be said to correspond with xAI – a salient exponent in the 'ethical AI' genre – however the two visualisations specifically examining xAI were Bail-Tree and H-Bar. Even so, they are very different visualisations: Bail-Tree, as a tree classifier, was purposely prepared to explore visual literacy and analysis of a visualisation by highlighting the attributes or predictor variables used to decide a defendant's bail. Bail-Tree was observed to be consistent by means of comprehensibility, an observation exemplified in the literature based on its 'identifiability' to a class and intelligibility [81]. Conversely, H-Bar was prepared to draw on perceptiveness, decisiveness and justifiability, and the data was indicative of some uncertainty and disagreement, but ultimately a greater proportion found its justifiability. This theme on a contrastive outcome was persuasively expressed in a formulaic manner by [176], rephrased in the following:

> To justify the decision of X and not Y, understanding why X and not Y, contains the reason for X and of the failure for a simultaneous outcome of not Y.

Interestingly for H-Bar, the data showed a sharp swing toward a change in response, subsequent to additional information upon what crime the bail case was deciding – a likely reassessment of the justification in this data point being AI-generated. Comparably to the swing substantiated in the data from this survey, the research literature on AI-prediction and recidivism claimed that a small percentage of participants would reassess their opinion subsequent to new information given by AI [26].

Individual fairness and Group fairness constructs share a juxtaposition in that they both seek to be representative of those characteristics which define humans as singularly unique, while defining group membership within society. If the occasion arose where such characteristics become determinants in decision-making, a reasonable mind would suggest it as unfair unequivocally. As intended, the AI-generated decisions were presented in different visualised formats based upon the same data to interrogate participant perceptions upon whether an individual or a group of individuals, would be fairly assessed for bail. The results were indicative that AI-based decisions, including the two visualisations, were representative of fairness. While difficult to make direct comparisons, research outputs with some similarities on AI-decisions and people perceptions featured fairness prominently, where the balance between participant attitudes toward AI-based justice and fairness was not in favour of one or the other [30][32]. A veiled and somewhat inadvertent matter on punitive action and recourse has become more apparent in this study's results, where the data regarding erroneous decisions which test moral duty and ethical consequence are in the balance. These were put to participants in deontological and teleological postulations. This was expressed cogently in the literature:

> Participants must categorise a harmful action as either acceptable or unacceptable, thereby endorsing *either* the deontological *or* utilitarian principle [emphasis made in original quote] [43].

An interpretation of this is in the interchange of moral principles and bail decision-making. It conveys the resulting misclassification in predictive modelling leading to incorrect decisions, that is, a defendant could be refused bail and imprisoned for an undetermined period of time, or granted bail and be at liberty for an undetermined period of time thereby potentially posing a risk to a community. As such, the data resulted in a greater portion agreeing with the utilitarian argument on the greater good, meaning any incorrect decision having a minor impact was acceptable for a majority to benefit. However, the deontological position did not favour participants.

Contended in the literature is whether deontology and teleology are juxtaposed or opposing principles having an inverse relationship, and to remedy this contention, any relationship or lack thereof is to be tested mathematically [43]. As referred to in the results section, Spearman's Rho was the chosen test, and the outcome from this validated the contention of any relationship provided by the data is not evident. Moreover, by some measure of observable validation, a variation between the two principles was reflected.

While there was not any direct or similar correlation in the reviewed literature on the moral principles and AI-generated bail decisions, research that measured public attitudes on "algorithmic decision-making" to calculate "criminal risk" indicated a greater percentage of the public were approving of it for inmates awaiting parole [30].

Fundamentally, the culmination of the survey data demonstrated that participants were generally adept in analytical reasoning of the visualisations that depicted decisions, and were decisive on ethical and moral matters.

It became apparent during the course of this Primary Study, having attended some courthouses in the Queensland criminal jurisdiction, that a courtroom was designated solely for the purpose of bail matters. This gave me pause to consider a justice system where defendants could have the opportunity to put their bail case before an intelligible court – evaluated in the intelligible world of predictive models and machine learning – where decisions are formulated on fairness and explainability (among other ethical considerations). This is evidenced in the least by participants having responded comprehensibly to AI-generated bail decisions, inclusive of ML and VA, where as such, there is scope for a future intelligible justice system.

Chapter 6

Conclusions and Future work

6.1 Conclusion

This thesis aimed to examine the application and facilitation of AI-generated bail decisions and whether VA can mitigate *fairness* and *explainability*. As such, three research objectives were proposed, in summary: to identify ML and VA as instruments to facilitate bail decisions; to determine ML effectiveness at decision-making and VA at mitigating those decisions; third, were AI-generated decisions perceived within teleological or deontological principles. These were not intended to bring absoluteness to the outcomes through conventional statistical methods; rather, it was predicated on a heuristic approach, and the inception of ideas in the sense of a pilot study, as remarked upon in the literature [15][159]. Three chapters were individually apportioned for each study, as summarised in the following.

Chapter 3 was to provide a basic insight into bail as a subject matter and what importance it plays in the justice system. The impetus for this study was the change in bail laws in the state of NSW and the impact over a 10-year period to the court system. It employed conventional quantitative techniques and statistical analyses. For example, a descriptive analysis determined the yearly average of defendants who were refused bail had increased from 2014 onwards, as well as, breach of bail and court delays (remand time). Court delays had in fact increased by 52 percent. These results were similarly found in past research [177][178]. Applying other analysis methods, correlation was found between refused bail and court delays, and the linear models were a good fit to the data, although there was not any statistical significance. It is noted nonetheless that some scholars argue against significance value and explanatory analysis measures [136][137].

Chapter 4 was motivated by predictive modelling, regression and EDA, as quantitative and qualitative methods, devised in order to produce bail data in the development of visual analytics. While the logistic regression algorithms used in these models are well-established, the variables applied in these ML models were formulated out the bail legislation, having never been applied in this way to the knowledge of this writer (refer to Appendix B for the Model Predictor Information Table). As such, regression and classifier models substantiated reputable results as predictive mechanisms, which is consistent with the past literature [87][89][75]. The validity of these results gave both statistical and ethical assurances from a researcher perspective to

potential participants on the data and results validity.

Chapter 5 employed a qualitative and quantitative survey, which was evaluated by statistical and EDA techniques. A positive trend was observed in participant perceptions on AI-generated bail decisions toward a predictive model approach – a presumptive affect of the visualisations presented as part of the survey structure. An interesting finding was on the cut-off point/threshold – where the respective visualisation represented a clear indicator on where a threshold would lie – and where a majority of participants said they would reconsider their response in favour of a threshold after learning the defendant was charged with a serious offence yet bailed. A similar finding on participant perception and the offence seriousness was noted in other research [31]. Although participants were divided on the teleological and deontological moral perspectives, a utilitarian acceptance for harm to some to the benefit of others, was apparent. While it was not a measure in this study, the authenticity of using a real-world predictive model with real-world data is considered to be a worthwhile pursuit as observed through each participant engagement.

Corresponding to ethics in AI-generated decision-making in the Primary Study and Preliminary Study II was the latency in algorithm fairness, that is, limited exploration on fairness at any other stage that would address uncertainty expressed by some participants, or untested confidence in participants without esoteric domain advantage, for example. While the preceding literature on visual analytics to alleviate ethical concerns on fairness and explainability does not appear to have been explored specifically as a bail decision mechanism in any Australian criminal jurisdiction until now, visual analytics and ethics have proven to be essential to AI-generated decision-making for bail and the justice domain more generally [30][31].

6.2 Future work

While AI-generated decision-making continues to evolve, ethical and moral concerns will inevitably manifest. When applied in real-world scenarios, such as in bail – the determination by machine on one's liberty – serious implications and misgivings emerge based upon decision fairness and explainability. However, these are only two ethical principles, and as a result, further work into other principles is required. Following on from this, further exploration on facilitating ethical considerations using visual analytics as a corresponding technique may alleviate any implications and misgivings.

The user-study reported a limitation on the participant size. While there was strength in targeting participants within the court-user specified domain, any future study would benefit from a larger sample size in the provision of statistical rigour. This could include, but not be limited to, other individuals/groups beyond those attending court as defendants or victims, where judicial members and legal representatives, or members of the general public, would be surveyed.

A reoccurring subject matter was 'seriousness of the offence': it was one of the attributes for the predictive model, became the most prominent variable in the tree-structured classifier,

and was relevant to participants in the survey. It was also a subject of interest in the literature. Therefore, it is logical for any future work in predictive modelling on bail decision-making to feature it.

The NSW Government have existing policies, strategies and frameworks on AI – that includes decision-making and building public trust – and it is feasible for them to be advanced. Decision-making and analytical reasoning are not limited to bail, and various other types of legal decisions in the NSW criminal jurisdiction could benefit from predictive assessment, particularly with orders that are decided on the subjective standard: balance of probability. Moving outside the NSW jurisdiction, it was observed during the course of the Primary Study that within the Queensland criminal jurisdiction, some court houses had court rooms dedicated for bail proceedings. This evoked a future concept of dedicated AI-generated bail courts, where decisions could support the traditional human-based methods through ML and VA, for the benefit of intelligibility.

AppendixA Survey Visualisations

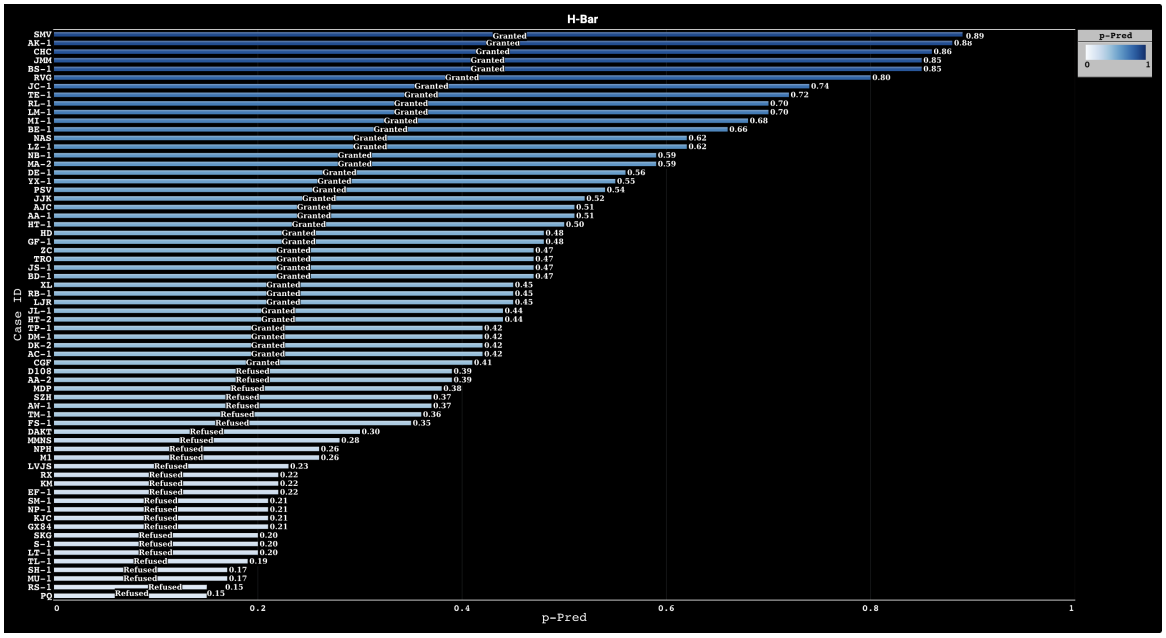

Figure A.1: H-Bar: defendants listed by code (left axis); AI-bail decision (middle of each bar); predictive values (end of each bar). Legend in the right top corner identifies predictive value by colour shade

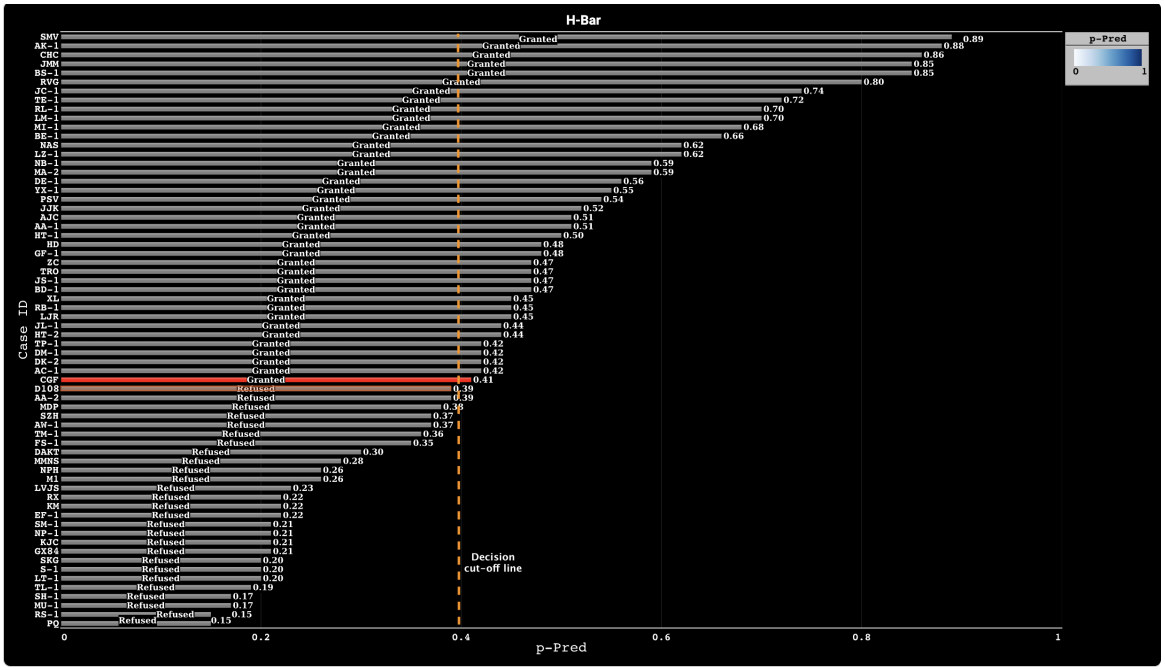

Figure A.2: H-Bar: contrastive explanation visual - example defendants horizontal bars illuminated in red; decision cut-off denoted by vertical dotted line

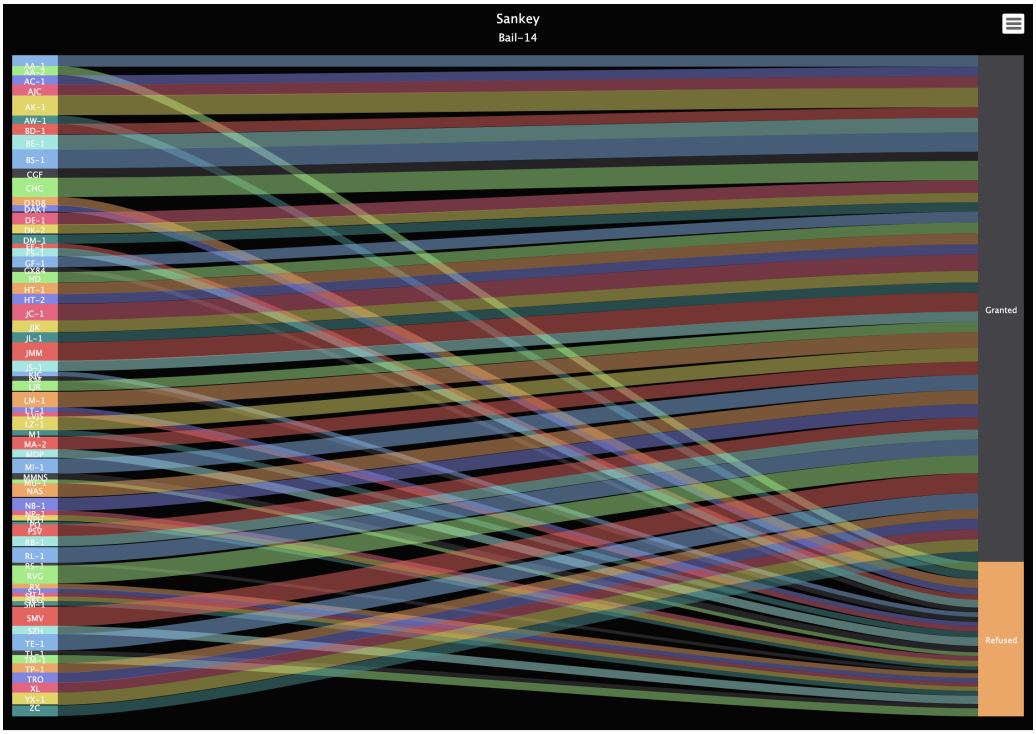

Figure A.3: Sankey: all defendants are listed on left axis; decisions travel on the illuminated channels to right axis to respective AI-bail decision of granted (top) or refused (bottom)

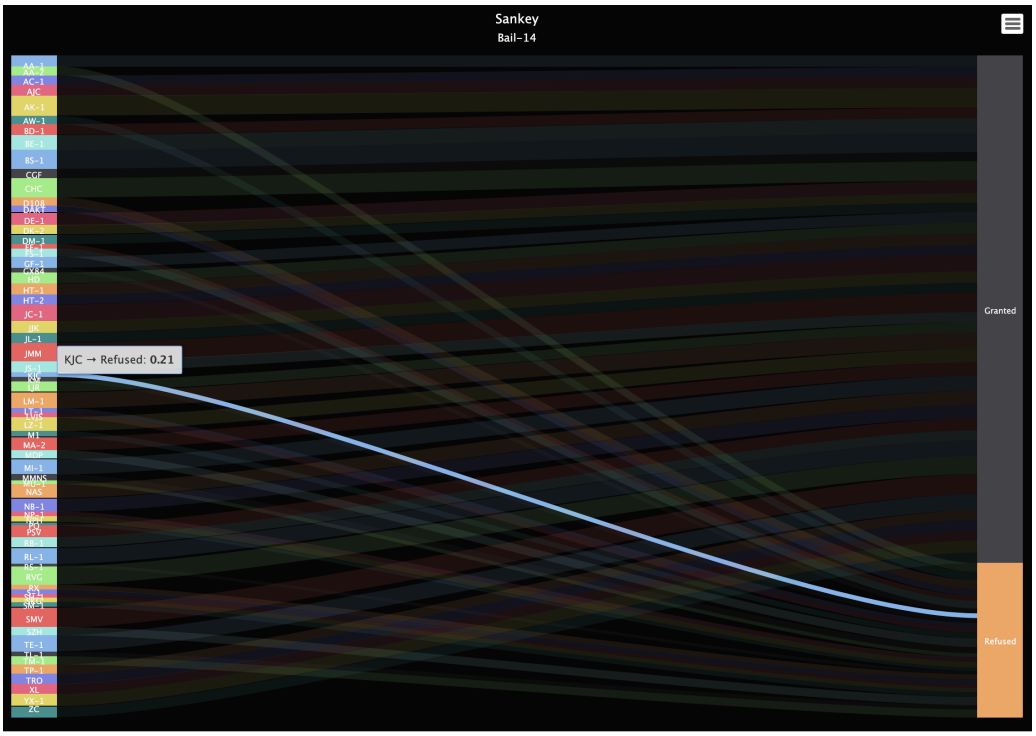

Figure A.4: Sankey: selected defendant on left axis "KJC" is illuminated (with literal and numeral notations of bail decision) and AI-bail decision channel illuminated to right axis, highlighting "Refused"

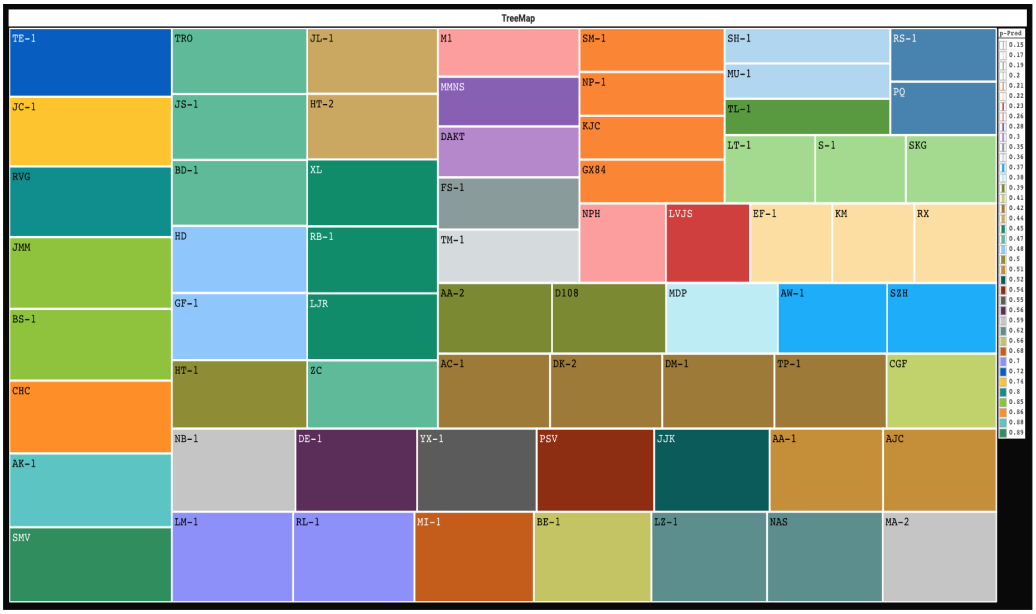

Figure A.5: TreeMap: bail decisions are identifiable by brick size and colour, and defendant code-identifiers are on each brick corner; predictive values are on the right side in Key. Note: bricks with same colour and size identifies same AI-bail decision.

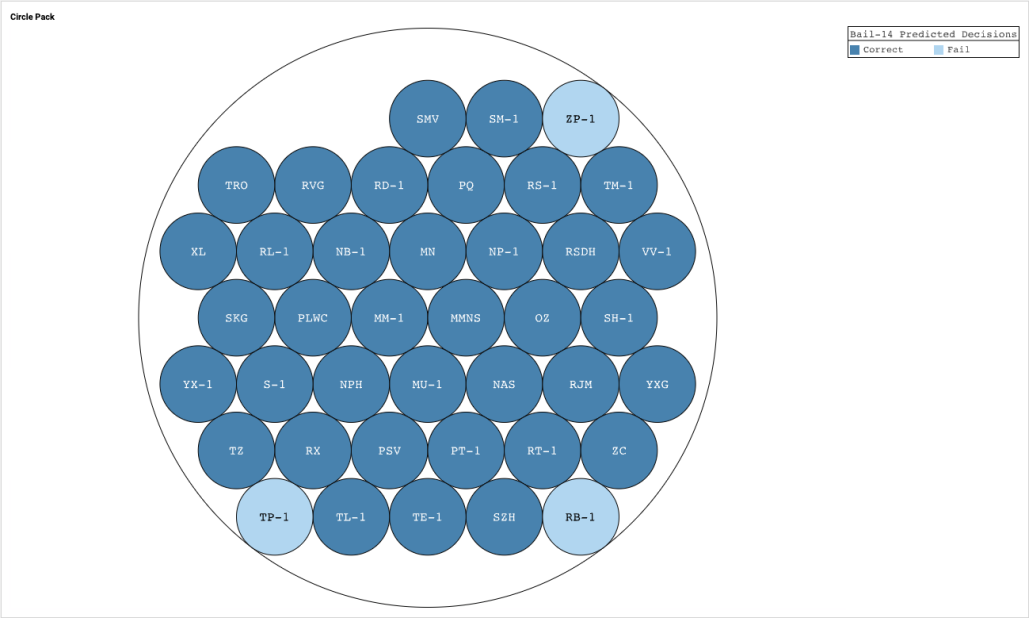

Figure A.6: Circle Pack: defendant identifier codes are on each circle; darker blue denotes correct predictions, lighter blue denotes failed predictions (as shown in Legend

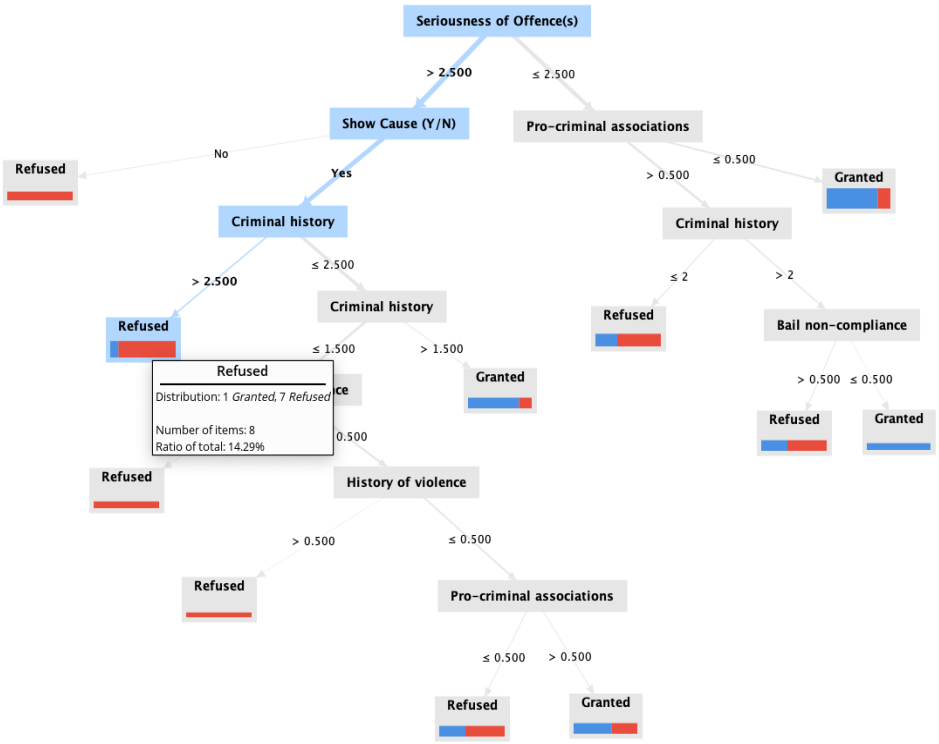

Figure A.7: Bail-Tree with a depth of 7. Defendant with code 'KJC' was used to demonstrate the predictors relevant to the decision were Seriousness of offence(s), Show Cause, Criminal history; values are identified on the branches between each node. Decision can be explained by starting at top (root node) and follow blue illuminated path along each branch and node (predictor classifier) to the AI-generated bail decision "Refused".

AppendixB Model Predictor Information Table

Model Predictor Information Table

Predictors and descriptions			Coding / Value
	Show Cause test		
	Criteria: New offence is a serious indictable offence (e.g., murder, child-sex offence). Defendant on bail, parole or subject of an active warrant. Non-compliance with current or prior supervision order. Current or prior serious personal violence. Firearm/weapon offences. Commercial, trafficking drug offences.	**B-LogR model** Determine if this is an offence under the criteria No [0] Yes [1] **TsC model** Determine if the defendant has shown cause why detention is not justified in reference to the criteria No [1] Yes [0]	
	Criminal history		
	Number of adult convictions (exclude summary offences). Defendant has a criminal conviction in any jurisdiction (exclude summary offences).	0 convictions No [0] ∴ Nil 1 conviction Yes [1] ∴ Low 2 convictions Yes [2] ∴ Moderate 3+ convictions Yes [3] ∴ High	
	Seriousness of offence(s)		
	Low Offences within this category are likely to be dealt with by a lower court (e.g., driving and summary offences which resulted in a fine) and as such the relative seriousness is low.	Penalty *=<3y* Yes [1] ∴ Low	
	Moderate Offences within this category are likely to be indictable, dealt with by the district or supreme court, and would carry a supervised parole period following release discretion by the SPA. Value and coding in this category on relative seriousness considers fraud/financial deception offences that are relevant to other deceptive behaviour that can impact bail determination (see flight risk/ failure to attend category).	Penalty *>3y to <=10y* Yes [2] ∴ Moderate	
	High Offences within this category are likely to contain violence or transnational trafficking and release discretion by SPA and SORC.	Penalty *>10y* Yes [3] ∴ High	

Table B.1: Model Predictor Information Table – page 1 of 2

Model Predictor Information Table

Predictors and descriptions		Coding / Value
History of violence		
Prior charges, convictions regarding violence Recorded information from custody status (matters of violence while in police or corrections custody)		No [0] Yes [1]
Prior charge or conviction while on bail (in NSW or another jurisdiction)		
Defendant has breached bail conditions on one or more occasions, even if it did not result in a court hearing.		No [0] Yes [1]
History of non-compliance with court-issued orders		
Defendant has demonstrated non-compliance with orders for community-based supervision (e.g., parole).		No [0] Yes [1]
Pro-criminal associations		
• Defendant is a member or linked to organised criminal network, gang affiliation. • Current offences were committed in company of co-accused. • Defendant has history of offending in company of other convicted offenders.		No [0] Yes [1]
Danger that may be posed to the public/victim(s) or other persons known to them		
• Defendant has previously demonstrated threats, did or had attempted to contact victim or witnesses. • Defendant has currently or previously demonstrated disregard for public safety.		No [0] Yes [1]
Flight risk/failure to attend		
• Flee a jurisdiction; ability to obtain falsified identification, documents, passports. • Defendant has prior occasions of failing to appear at court.		No [0] Yes [1]

Table B.2: Model Predictor Information Table – page 2 of 2

AppendixC Literature Review Synopsis

Author(s)	Focus	Methodology	Results	Limitations
[54]	Technology limiting bias in legal data	Quantitative	Bias was identified in data, particularly concerning race.	Low participation affected sample size
[41]	Testing a predictive algorithm adjusted for bias	Mixed-methods	No evidence of racial bias - some reduction noted in reducing bias against black offenders.	Overfitting - too many variables. Algorithm was trained one-way, testing black offenders against white, and would have benefited from both ways.
[45]	ML to inform bail decisions	Quantitative	Determined ML is more accurate than human-based decisions	There were a large number of input variables. Data is for one metropolitan jurisdiction which can affect reliability.
[112]	Computational decisions v human decisions	Qualitative	Participants favoured humans over algorithm-based decisions	Participants were influenced one particular way that could affect results. Mix sample of college students and crowdsourcing (MTurk)
[43]	Sentencing disparity in Victoria criminal courts	Quantitative	Sentencing disparity evident between courts for the same penalties	Re-sentencing for breach may not be reflected in records; No consideration for decision-maker rotations to other courts causing data inaccuracy; Unrepresented offenders can be dealt with differently by courts than represented offenders.

Table C.1: Literature Review Synopsis – page 1 of 4

Author(s)	Focus	Methodology	Results	Limitations
[68]	Public perception of ML risk assessments in criminal proceedings	Mixed-methods	Multiple sub-studies: participants expressed ethical concerns with ML risk assessments for bail and sentencing.	Public perceptions are subjective and sample group is not representative of the broader population
[50]	Predicting recidivism: what risk assessment is more statistically robust; what predictor domains are relevant	Quantitative Meta-analysis	LSI-R was most effective at predicting recidivism. Age and criminal history were strong predictor domains	Meta-analysis produces generalisations in the results due to studies being analysed. Database used for the meta-analysis showed bias in gender and race.
[51]	Clinical versus mechanical assessment	Meta-analysis	Mechanical (statistical) prediction more accurate than clinical (human)	Qualitative information may be limited due to clinical observations
[49]	Predict failure to appear in pretrial defendants using ML methods	Quantitative. Various ML methods	Indicative that ML is effective at assessing non-compliance. Using less variables is as effective as many variables.	Court records may have been missing information creating bias in calculation.
[77]	Evaluate various risk assessments used in pretrial matters	Qualitative	Demonstrated inconsistencies in the risk assessments	Critique of the statistical outcomes were not actually measured by the authors, and it would have benefited if they could have tested against actual court matters
[80]	Using algorithms for decision-making in criminal justice	Quantitative	Algorithm was of benefit in decisions, e.g., reduction in detainees following bail hearing.	Research affected by administrative matters e.g., information/data inconsistency

Table C.2: Literature Review Synopsis – page 2 of 4

Author(s)	Focus	Methodology	Results	Limitations
[110]	Prediction	Quantitative	Predictive visual analytics are effective	Small sample size affected outcomes
[108]	Assessing visual analysis against written analysis	Mixed-methods	No significant improvement with visual over text	Participants were obtained through 'crowdsourcing' through the internet
[91]	Predicting violence of inmates using ML techniques	Quantitative; Various ML techniques	ML techniques predicted violence better than conventional regression	Data may have been subjective and inaccurate based on prison records
[6]	Crime prediction using ML Linear Regression Techniques and Visualisation	Quantitative; Linear Regression	ML is effective in finding hidden patterns in crime data	Data from policing sources may be inaccurate or unavailable
[73]	Public attitudes on algorithmic decision making	Mixed-methods	Participants expressed concern for fairness on algorithm-based decisions and the implications in actual life events	Surveys can contain bias and errors due to phrasing of questions
[45]	Predicting decisions from legal cases using eight classifiers	Quantitative and qualitative; ML classifiers	All eight classifiers showed good statistical vigour in prediction	Methodology used 19 features that contained descriptors - it may have benefited the results by having less descriptors being more specific and objective (some of the features used in data collection and analysis seemed arbitrary)

Table C.3: Literature Review Synopsis – page 3 of 4

Author(s)	Focus	Methodology	Results	Limitations
[69]	Decision-making affected by Collapsing Choice Theory	Quantitative	Multiple sub-studies demonstrated a difference between choice and judgement: greater the number of alternative options does not impact on correct choices.	Bias impacted decision-making for participants
[46]	Comparison of various ML approaches for predicting recidivism	Quantitative; ML techniques	Results of research reported that a simple ML technique, SLIM, performs better than more recognised ML techniques	Based on the features used were from police records, there was not any mention of bias that has been well-reported in other studies
[132]	Examine predictors affecting failure to appear rates	Quantitative; Logistic Regression	Gender and impoverishment identified as predictors mostly affecting failure to appear at court	Information regarding social circumstances of sample could have affected results. Sample were from one jurisdiction.
[66]	Public attitudes toward algorithm fairness and effectiveness	Mixed-methods	Majority of participants found automated scoring of parole was between not very fair to somewhat fair	Qualitative aspects of this research approach can contain biases. Large survey with high number of questions.

Table C.4: Literature Review Synopsis – page 4 of 4

AppendixD Flow Charts 1 and 2

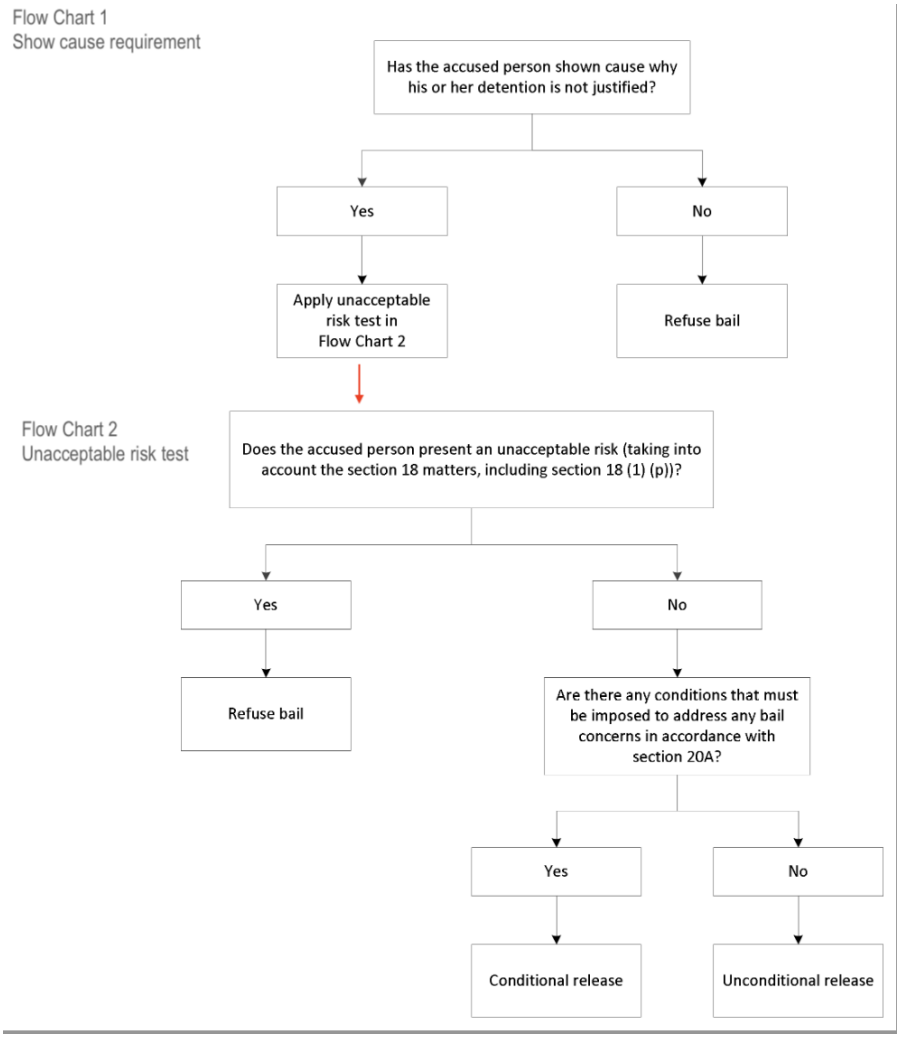

Figure D.1: Flow Charts 1 and 2–*Bail Act 2013* (NSW).
The red arrow is inserted to indicate the juncture between the two tests.

AppendixE BAILgram pseudocode

```
// === Nodes and Flows ===

Show Cause [1] Yes (FC1)
Show Cause [1] No (FC1)

No (FC1) [1] Refuse Bail (FC1)

Yes (FC1) [1] Unacceptable Risk - s.18 (FC2)
Unacceptable Risk - s.18 (FC2) [1] Yes (FC2)
Yes (FC2) [1] Refuse Bail (FC2)

Unacceptable Risk - s.18 (FC2) [1] No (FC2)
No (FC2) [1] Conditions - s.20A

Conditions - s.20A [1] No bail concerns - No conditions
No bail concerns - No conditions [1] Unconditional Release

Conditions - s.20A [1] Yes bail concerns - Impose conditions
Yes bail concerns - Impose conditions [1] Conditional Release

// === Moved Nodes ===

move Conditional Release 0.00216, -0.65161
move Unconditional Release -0.00229, 0.00136
move No (FC1) -0.00162, 0.90743
move Yes (FC1) -0.00296, -0.20335
move Refuse Bail (FC1) 0.02995, 0.93273
move Refuse Bail (FC2) 0.08092, 0.88287
move Yes (FC2) 0.06247, 0.69223
move Yes bail concerns - Impose conditions 0.04126, -0.47846
move No bail concerns - No conditions 0.03778, -0.16873
move Show Cause -0.19081, 0.3736
move Unacceptable Risk - s.18 (FC2) 0.05347, -0.18851
move No (FC2) 0.04562, -0.69639
move Conditions - s.20A 0.06066, -0.53701
```

Figure E.1: BAILgram pseudocode

Bibliography

[1] B. Waltl and R. Vogl. Explainable artificial intelligence the new frontier in legal informatics. *Jusletter IT*, 4:pp. 1–10, 2018.

[2] G. Oh. *Predicting Life-Course Persistent Offending Using Machine Learning*. PhD thesis, Sam Houston State University, 2021.

[3] P. Lussier, N. Deslauriers-Varin, J. Collin-Santerre, and R. Bélanger. Using decision tree algorithms to screen individuals at risk of entry into sexual recidivism. *J of Crim. Just.*, 63:pp. 12–24, 2019. DOI: 10.1016/j.jcrimjus.2019.05.003.

[4] K. Lum and C. Boudin and M. Price. The impact of overbooking on a pre-trial risk assessment tool. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020. DOI: 10.1145/3351095.3372846.

[5] N. Yeo. Most expensive suburbs to buy in 2020 – Houses, Sep. 2021. http://www.realestate.com.au/. Accessed Sep. 12, 2021 [online].

[6] I.C. Obagbuwa and A.P. Abidoye. South africa crime visualization, trends analysis, and prediction using machine learning linear regression technique. *App. Comput. Intell. and Soft Comput.*, 48(4):pp. 1–14, 2021. DOI: 10.1155/2021/5537902.

[7] A. Endert, L. Bradel, and C. North. Beyond control panels: Direct manipulation for visual analytics. *IEEE Comput. Grap. Appl.*, 33:pp. 6–13, 2013. DOI: 10.1109/MCG.2013.53.

[8] C. Gorg, Y. Kang, Z. Liu, and J. Stasko. Visual analytics support for intelligence analysis. *Computer*, 46(7):pp. 30–38, 2013.

[9] J. Wenskovitch, L. Bradel, M. Dowling, L. House, and C. North. The effect of semantic interaction on foraging in text analysis. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages pp. 13–24, Berlin, Germany, 2018. IEEE. DOI: 10.1109/VAST.2018.8802424.

[10] A. Endert, W. Ribarsky, C. Turkay, B.L.W. Wong, I. Nabney, I.D. Blanco, and F. Rossi. The state of the art in integrating machine learning into visual analytics: Integrating machine learning into visual analytics. *Computer Graphics Forum*, 36(8):pp. 458–486, 2017. DOI: 10.1111/cgf.13092.

[11] D. Soroko, N. Döge, A. Al Shafeei, and H. Heuer. Unpacking a model: An interactive visualization of a text similarity algorithm for legal documents. In *Proceedings*

of Mensch und Computer 2019, pages pp. 875–879, Hamburg, Germany, 2019. DOI: 10.1145/3340764.3345371.

[12] C. Steed, P.J. Fitzpatrick, J. Edward Swan II, and T.J. Jankun-Kelly. A visual analytics approach for correlation, classification, and regression analysis. In M.L. Huang and W. Huang, editors, *Innovative Approaches of Data Visualization and Visual Analytics: in Advances in Data Mining and Database Management*, pages pp. 24–45. IGI Global, 2014. DOI: 10.4018/978-1-4666-4309-3. Accessed Sep. 14, 2021 [online].

[13] New South Wales Bureau of Crime Statistics and Research. Criminal Court Statistics 2023. www.bocsar.nsw.gov.au/. Accessed Mar. 2023 [online].

[14] RapidMiner Inc. Rapidminer-studio (version 10.2), 2024. www.rapidminer.com.

[15] N. Stobbs and D. Hunter and M. Bagaric. Can sentencing be enhanced by the use of artificial intelligence. *Criminal Law Journal*, 41(5), 2017.

[16] N. L. Hillman. The use of artificial intelligence in gauging the risk of recidivism. *The Judges' Journal*, 58(1):pp. 36–39, 2019.

[17] D. Hogan-Doran. Computer says no: automation, algorithms and artificial intelligence in government decision-making. *The Judicial Review*, 13, Sep. 2017.

[18] Australian Human Rights Commission and World Economic Forum. Artificial intelligence: Governance and leadership (white paper). Aust. Human Rights Comm., 2019. http://ahrc-wef-white-paper-online-version-final20.pdf. Accessed Dec. 1, 2022 [online].

[19] M. Zalnieriute. Technology and the courts: Artificial intelligence and judicial impartiality. Aust. Law Reform Comm., 2019. http://alrc.gov.au/wp-content/uploads/2021/06/3-Monika-Zalnieriute-Public.pdf. Accessed Apr. 1, 2023 [online].

[20] U.K Gov. CDEI. CDEI AI Barometer: Independent Report, Centre for Data Ethics and Innovation. https://www.gov.uk/government/publications/cdei-ai-barometer/cdei-ai-barometer-criminal-justice. Accessed Feb. 7, 2021 [online].

[21] A. van Wynsberghe and S. Robbins. Critiquing the reasons for making artificial moral agents. *Sci. Eng. Ethics*, 25(3):pp.719–735, 2019. DOI: 10.1007/s11948-018-0030-8.

[22] P. Formosa and M. Ryan. Making moral machines: why we need artificial moral agents. *AI and Soc.*, 25(3):pp. 719–735, 2019. DOI: 10.1007/s00146-020-01089-6.

[23] C. Coglianese and D. Lehr. Regulating by robot: Administrative decision making in the machine-learning era. *The Georgetown Law Review*, 105(5):pp. 1147–1223, 2017. DOI: 10.1007/s00146-020-01089-6.

[24] G.V. Travaini and F. Pacchioni and S. Bellumore and M. Bosia and F. De Micco. Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction. *IJERPH*, 19(17):pp.1–13, 2022. DOI: 10.3390/ijerph191710594.

[25] M.T. Fischer and S.D. Hirsbrunner and W. Jentner and M. Miller and D.A. Keim and P. Helm. Promoting ethical awareness in communication analysis: Investigating potentials and limits of visual analytics for intelligence applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, volume June, pages pp. 877–889, Seoul, Republic of Korea, 2022. ACM. DOI: 10.1145/3531146.3533151.

[26] N. Grgić-Hlača and C. Engel and K.P. Gummadi. Human decision making with machine assistance: An experiment on bailing and jailing. *Proc. ACM Hum.-Comput. Interact*, 3:pp. 1–25, 2019. DOI: 10.1145/3359280.

[27] K. Nazemi and D. Burkhardt and A. Kock. Visual analytics for technology and innovation management: An interaction approach for strategic decision making. *Multimed Tools Appl*, 81(11):pp. 14803–14830, 2022. DOI: 10.1007/s11042-021-10972-3.

[28] B. Dupont, Y. Stevens, H. Westermann, and M. Joyce. Artificial intelligence in the context of crime and criminal justice, 2018. DOI: 10.2139/ssrn.3857367.

[29] A. Lara. Some apparent obstacles to developing a kantian virtue theory. *Análisis filosófico*, 30(2):pp. 187–219, 2010.

[30] A. Smith. Public attitudes toward computer algorithms, 2018. www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/. Accessed Dec. 27, 2021 [online].

[31] A.Fine. *Machine Learning Based Risk Assessments*. PhD thesis, Arizona State University, 2021.

[32] E. Treyger, J. Taylor, D. Kim, and M. Holliday. Assessing and suing an algorithm. Technical report, RAND Corp, 2023.

[33] Australian Productivity Commission. Australia's prison dilemma (Research Paper). www.pc.gov.au/research/completed/prison-dilemma/prison-dilemma.pdf. Accessed Feb. 25, 2023 [online].

[34] Audit Office of New South Wales. Managing growth in the NSW prison population, 2019. www.audit.nsw.gov.au/. Accessed Feb. 25, 2023 [online].

[35] A. Morgan. Managing growth in the NSW prison population, 2018. www.aic.gov.au/. Accessed Feb. 25, 2023 [online].

[36] B. Whitby. *On Computer Morality: An Examination of Machines as Moral Advisors*. Cambridge, Cambridge University Press, 2011.

[37] F. Lovett. *Rawls' a theory of justice: a reader's guide*. Continuum, London; New York, 2011.

[38] T. Pogge. Equal liberty for all? In O. Hoffe, editor, *John Rawls, A Theory of Justice*, pages pp. 95–110. Koninklijke Brill NV, Leiden, The Netherlands, 2013.

[39] O. Hoffe. An introduction to rawls' theory of justice. In O. Hoffe, editor, *John Rawls, A Theory of Justice*, pages pp. 1–20. Koninklijke Brill NV, Leiden, The Netherlands, 2013.

[40] J. Storrs Hall. Ethics for machines. In *Machine ethics*, pages pp. 28–46. Cambridge University Press, New York, USA, 2011.

[41] S. Freeman. Utilitarianism, deontology, and the priority of right. *Philosophy and Public Affairs*, 23(4):313, 1994.

[42] M. Timmons. *Moral theory: An introduction*. Rowman & Littlefield Publishers, 2012.

[43] P. Conway and B. Gawronski. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, 104(2):216, 2013.

[44] J. Avery. Legal data: Bias in the law, and how legal technology can be built to help correct for it. PhD thesis, Princeton University, 2021.

[45] G.F. Gaus. What is deontology? part one: Orthodox views. *The Journal of Value Inquiry*, 35:27 – 42, 2001.

[46] K.A. Hegtvedt and H.L. Scheuerman. The justice/morality link. In S. Hitlin and S. Vaisey, editors, *Handbook of Sociology and Morality*, pages pp. 331–360. Springer, New York, USA, 2010. DOI: 10.1007/978-1-4419-6896-8-18.

[47] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(Feb):pp. 1–38, 2019. DOI: 10.1016/j.artint.2018.07.007.

[48] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I.G. Penco, and A.C. Cosentini. A clarification of the nuances in the fairness metrics landscape, 2022. DOI: 10.1038/s41598-022-07939-1.

[49] M.H.L. Kaas. Raising ethical machines: Bottom-up methods to implementing machine ethics. In S.J. Thompson, editor, *Advances in Human and Social Aspects of Technology*, pages pp. 47–68. IGI Gloval, Hershey, PA, USA, 2021. DOI: 10.4018/978-1-7998-4894-3.ch004.

[50] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Soc. Methods and Research*, 50:pp. 3–44, 2021. DOI: 10.1177/0049124118782533.

[51] J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage. A survey on ethical principles of ai and implementations. *SSCI*, 50(1):pp. 3010–3017, IEEE 2020. DOI: 10.1109/SSCI47803.2020.9308437.

[52] Science Dept. of Industry and Resources. Australia's artificial intelligence ethics framework. Aus. Gov, 2023. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework. Accessed Dec. 2, 2021 [online].

[53] R. Berk and A.A. Elzarka. Almost politically acceptable criminal justice risk assessment. *Crim, and Pub. Policy*, 19(4):pp. 1231–1257, 2020. DOI: 10.1111/1745-9133.12500.

[54] H.W. Liu, C.F. Lin, and Y.J. Chen. Beyond state v loomis: artificial intelligence, government algorithmization and accountability. *Int. J of Law and Info. Tech.*, 27(2):pp. 122–141, 2019. DOI: 10.1093/ijlit/eaz001.

[55] S.M.G. Rankin. Technological tethereds: Potential impact of untrustworthy artificial intelligence in criminal justice risk assessment instruments. *Wash. and Lee Law Rev.*, 78(2):p. 647, 2021. DOI: 10.1093/ijlit/eaz001.

[56] B. Mittelstadt, C. Russell, and S. Wachter. Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages pp. 279–288, Atlanta GA USA, 2019. ACM. DOI: 10.1145/3287560.3287574.

[57] J.M. Stibel, I.E. Dror, and T. Ben-Zeev. The collapsing choice theory: Dissociating choice and judgment in decision making. *Theory Decis.*, 66(2):pp. 149–179, 2009. DOI: 10.1007/s11238-007-9094-7.

[58] B. Hedden. On statistical criteria of algorithmic fairness. *Philos. Public Aff.*, 49(2):pp. 209–231, 2021. DOI: 10.1111/papa.12189.

[59] F. Bell, P.L. Bennett Moses, P.M. Legg, J. Silove, and M. Zalnieriute. Ai decision-making and the courts: A guide for judges, tribunal members and court administrators, 2022. www.aija.org.au. Accessed Dec. 10, 2022 [online].

[60] A. Mannes. Governance, risk and artificial intelligence. *AI Mag*, 41(1):pp. 61–69, 2020. DOI: 10.1609/aimag.v41i1.5200.

[61] Z.(J). Lin. *Algorithmically Guided Human Decision Making in Criminal Justice and Beyond*. PhD thesis, Stanford University, 2021.

[62] Z.C. Lipton. The mythos of model interpretability, 2017. http://arxiv.org/abs/1606.03490. Accessed Jan. 29, 2023 [online].

[63] Y.F. Ng, M. O'Sullivan, M. Paterson, and N. Witzleb. Revitalising public law in the technological era: Rights, transparency and administrative justice. *UNSW Law Journal*, 43(3):pp. 1041–1052, 2020. DOI: 10.53637/YGTS5583.

[64] P. Attewell and D. Monaghan. *Data Mining for the Social Sciences: An Introduction*. Uni. Cal. Press, 2019. DOI: 10.1525/9780520960596.

[65] P.D. König and D.T. Krafft. Evaluating the evidence in algorithmic evidence-based decision-making: the case of us pretrial risk assessment tools. *Current Iss. in Crim. Just.*, 33(3):pp. 359–381, 2021.

[66] A.J. Larner. *The 2x2 Matrix: Contingency, Confusion and the Metrics of Binary Classification.* Springer International Publishing, Cham, 2021. DOI: 10.1007/978-3-030-74920-0.

[67] M. Ghasemi, D. Anvari, M. Atapour, J.S. Wormith, K.C. Stockdale, and R.J.Spiteri. The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Crim. Just. and Behav.*, 48(4):pp. 518–538, 2021. DOI: 10.1177/0093854820969753.

[68] J. Moor. The nature, importance and difficulty of machine ethics. In *Machine Ethics*, pages pp. 13–20. Cambridge University Press, New York, USA, 2011.

[69] J. Ryberg. Sentencing disparity and artificial intelligence. *J Value Inquiry*, July:n.p., 2021. DOI: 10.1007/s10790-021-09835-9.

[70] C.A. Heimer. The unstable alliance of law and morality. In S. Hitlin and S. Vaisey, editors, *Handbook of Sociology and Morality*, pages pp. 179–202. Springer, New York, USA, 2010. DOI: 10.1007/978-1-4419-6896-8-10.

[71] E. Heath. Part i, chapter iii: Of man's progressive nature. In E. Heath, editor, *Adam Ferguson:Selected Philosophical Writings*, page n.p. Andrews UK Ltd, Bedfordshire, UK, 2007. ISBN: 9781845404437.

[72] L. Battle and J. Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau, 2019. DOI: 10.1111/cgf.13678.

[73] J.P. Pinder. *Introduction to business analytics using simulation, Second edition.* Academic Press, London, 2020. ISBN: 9780323917179.

[74] C. Britt and D. Weisburg. Logistic regression models for categorical outcome variables. In A.R. Piquero and D. Weisburd, editors, *Handbook of Quant. Crim.*, pages pp. 649–682. Springer, 2010. DOI: 10.1007/978-0-387-77650-7.

[75] F.T. Ngo, R. Govindu, and A. Agarwal. Assessing the predictive utility of logistic regression, classification and regression tree, chi-squared automatic interaction detection, and neural network models in predicting inmate misconduct. *Am J Crim Just*, 40(1):pp. 47–74, 2015. DOI: 10.1007/s12103-014-9246-6.

[76] J.D.Wilkinson, M.A.Mamas, and E.Kontopantelis. Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study. *J of Clin. Epid.*, 152:pp. 176–184, 2022.

[77] D.G. Kleinbaum, L.L. Kupper, A. Nizam, and E.S. Rosenberg. *Applied Regression Analysis and other Multivariable Methods, Fifth edition.* Cengage Learning, Aust., 2014. ISBN: 9781285963754.

[78] S. Wijenayake, T. Graham, and P. Christen. A decision tree approach to predicting recidivism in domestic violence. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Revised Selected Papers 22*, pages pp. 3–15, Melb. Aus., Jun. 3, 2018. Springer.

[79] A. Nasridinov, S-Y. Ihm, and Y-H. Park. A decision tree-based classification model for crime prediction. In *Information Technology Convergence: Security, Robotics, Automations and Communication*, pages pp. 531–538. Springer, 2013.

[80] V.E. Lee, L. Liu, and R. Jin. Decision trees: Theory and algorithms. In C.C. Aggarwal, editor, *Data Classification:Algorithms and Applications*, pages pp. 87–120. Chapman and Hall/CRC, New York, US, 2014. DOI: 10.1201/b17320.

[81] B.S. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39:pp. 261–283, 2013.

[82] J. Zeng, B. Ustun, and R. Cynthia. Interpretable classification models for recidivism prediction. *J.R.Stat.Soc.(A)*, 180:pp. 689–722, 2017.

[83] J. Allsop AO. Technology and the future of the courts. *University of Queensland Law Journal*, 38(1):1–14, 2019. Copyright - Copyright University of Queensland, TC Beirne School of Law 2019.

[84] OECD Legal. Recommendation of the Council on Artificial Intelligence. *Org. Econ. Coop. and Dev.*, OECD Legal Instruments:pp.1–11, 2021. www.legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed Dec. 1, 2021 [online].

[85] C. Farmer, I. Parsons, and M. Bagaric. Sentencing inconsistencies: A case study. *Aust. Law J Rep.*, 92:pp. 517–528, 2017.

[86] M.J. Hall, D. Calabro, T. Sourdin, A. Stranieri, and J. Zeleznikow. Supporting discretionary decision-making with information technology:a case study in the criminal sentencing jurisdiction. *Uni. of Ottawa Law and Tech. Journal*, 2(1), 2005. www.papers.ssrn.com/sol3/papers. Accessed Oct. 2, 2021 [online].

[87] R.A. Shaikh, T.P. Sahu, and V. Anand. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:pp. 2393–2402, 2020. DOI: 10.1016/j.procs.2020.03.292.

[88] R.A. Berk and J. Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Crim. and Pub. Policy*, 12(3):pp. 513–544, 2013. DOI: 10.1111/1745-9133.12047.

[89] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. Goldstein. Simple rules to guide expert classifications. *J.R.Stat.Soc.(A)*, 183(Part 3):pp. 771–800, 2020.

[90] H.R. Zettler and R.G. Morris. An exploratory assessment of race and gender-specific predictors of failure to appear in court among defendants released via a pretrial services agency. *Crim. Just. Rev.*, 40:pp. 417–430, 2015. DOI: 10.1177/0734016815583350.

[91] G. Gendreau, T. Little, and C. Goggin. A meta-analysis of the predictors of adult offender recidivism: What works. *Criminology*, 34(4):pp. 575–608, 1996.

[92] W.M. Grove, D.H. Zald, B.S. Lebow, B.E. Snitz, and C. Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):pp. 19–30, 2000. DOI: 10.1037/1040-3590.12.1.19.

[93] T. Singh and Y. Jain and V. Kumar. Predicting parole hearing result using machine learning, 2023. ICETCCT, Dehradun. DOI: 10.1109/ICETCCT.2017.8280342.

[94] R.A. Berk, S.B. Sorenson, and G. Barnes. Forecasting domestic violence. a machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, 13(1):pp. 94–115, 2016. DOI: 10.1111/jels.12098.

[95] C. McCue. *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Butterworth-Heinemann, Oxford, U.K., 2014.

[96] New South Wales Government. *Bail Act 2013 (NSW)*, (No.26). www.legislation.gov.au. Accessed Nov. 16, 2021 [online].

[97] New South Wales Parliament in the House of Representatives. *Bail Amendment Bill 2014 (NSW), Second Reading*, 2014.

[98] P. Garling (Justice). *JM v R*, June 2015. New South Wales Supreme Court 978.

[99] L. McCallum (Justice). *M v R*, 2015. New South Wales Supreme Court 138.

[100] S. Bogart. Sankeymatic: Build a sankey diagram, 2023. sankeymatic.com/.

[101] S.Wykstra. Governance, risk and artificial intelligence. *AIMag*, 41(1):pp. 61–69, 2020. DOI: 10.1609/aimag.v41i1.5200.

[102] A. Wolfgang, A. Bertone, and S. Miksch. Tutorial: Intro to visual analytics. In *HCI and Usability for Medicine and Health Care: Third Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB*, pages pp. 453–456, Graz, Austria, 2007. Proceedings 3, 2007. Springer.

[103] D. Keim, J. Kohlhammer, F. Mansmann, T. May, and F. Wanner. Visual analytics. In D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors, *Mastering the information age: solving problems with visual analytics*, pages pp. 7–18. Eurographics Association, Goslar, Germany, 2010.

[104] G.D. Sun, Y.C. Wu, R.H. Liang, , and S.X. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *J.Comput.Sci.Technol*, 28(5):pp. 852–867, 2013. DOI: 10.1007/s11390-013-1383-8.

[105] M.T. Fischer, S.D. Hirsbrunner, W. Jentner, M. Miller, D.A. Keim, and P. Helm. Promoting ethical awareness in communication analysis: Investigating potentials and limits of visual analytics for intelligence applications. In *ACM Conference on Fairness, Accountability and Transparency*, pages pp. 879–889, Seoul, Korea, 2022. DOI: 10.1145/3531146.3533151.

[106] C.K. Reddy and Y. Li. A review of clinical prediction models. In C.K. Reddy and C.C. Aggarwal, editors, *Healthcare Data Analytics*. Springer International Publishing, Chapman and Hall/CRC, 2015. DOI: 10.1201/b18588.

[107] E. Wall, L.M. Blaha, C.L. Paul, K. Cook, and A. Endert. Four perspectives on human bias in visual analytics. In G. Ellis, editor, *Cognitive Biases in Visualizations*. Springer Nature, Cham, Switzerland, 2018. DOI: 10.1007/978-3-319-95831-6.

[108] T. von Landesberger, T. Schreck, D.W. Fellner, and J. Kohlhammer. Visual search and analysis in complex information spaces-approaches and research challenges. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages pp. 45–68. Springer, London, UK, 2012. DOI: 10.1007/978-1-4471-2804-5.

[109] J.B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94:pp. 42–53, 2019. DOI: 10.1016/j.artmed.2019.01.001.

[110] G. Alicioglu and B. Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers and Graphics*, 102:pp. 502–520, 2022. DOI: 10.1016/j.cag.2021.09.002.

[111] R.J. Crouser, L. Franklin, A. Endert, and K. Cook. Toward theoretical techniques for measuring the use of human effort in visual analytic systems. *IEEE Trans. Visual. Comput. Graphics*, 23:pp. 121–130, 2017. DOI: 10.1109/TVCG.2016.2598460.

[112] L. Micallef, P. Dragicevic, and J.D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):pp. 2536–2545, 2012.

[113] A. Arruda. An ethical obligation to use artificial intelligence: An examination of the use of artificial intelligence in law and the model rules of professional responsibility. *Am.J.Trial Advoc.*, 40:pp. 443–448, 2016.

[114] Y. Lu. *Methodologies in Predictive Visual Analytics*. PhD thesis, Arizona State University, 2017.

[115] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, and J.F. Bonnefon. Machine behaviour. *Nature*, 568:pp. 477–486, 2019. DOI: 10.1038/s41586-019-1138-y.

[116] B.J. Dietvorst, J.P. Simmons, and C. Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J of Exp. Psych.: General*, 144(1):pp. 114–126, 2015.

[117] New South Wales Parliament in the House of Representatives. *Bail Amendment Bill 2015 (NSW), Second Reading*, 2015.

[118] R. Beech-Jones (Justice). *DPP v Tony Mawad*, 2015. New South Wales Court of Criminal Appeal 227.

[119] N.J.Salkind. *Encyclopedia of Measurement and Statistics*. SAGE publications, 2006.

[120] E.Scarbrough and E.Tanenbaum. *Research Strategies in the Social Sciences: A Guide to New Approaches*, chapter Ordinary Least Squares and Logistic Regression Analysis. Oxford University Press, USA, 1998.

[121] R. Nau. Statistical forecasting: notes on regression and time series analysis, 2020. https://people.duke.edu/ rnau/411home.htm.

[122] C.E. Dismuke and R. Lindrooth. Ordinary least squares. In E.C.G. Chumney and K.N. Simpson (eds.), editors, *Methods and Designs for Outcomes Research*, pages pp. 93–104. ASHP, 2006.

[123] A. Bhattacherjee. *Social science research: Principles, methods, and practices*. University of South Florida, 2012.

[124] P.Deeprasertkul and K.Sarinnapakorn. A water level forecast of pattani river in the southern of thailand by deep learning. *Journal of Computer and Communications*, 11(Aug):14–28, 2023. DOI 10.4236/jcc.2023.118002.

[125] P.Mishra, C.M.Pandey, U.Singh, A.Keshri, and M.Sabaretnam. Selection of appropriate statistical methods for data analysis. *Annals of Cardiac Anaesthesia*, 22(3):297–301, 2019.

[126] P.N. Nardi. *Doing survey research: A guide to quantitative methods*. Routledge, 2018.

[127] A.G. Asuero, A. Sayago, and A.G. Gonzalez. The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36(1):41–59, 2006. doi: 10.1080/10408340500526766.

[128] C. Zaiontz. Real statistics resources pack for excel, copyright 2024. first downloaded 2022.

[129] M. Alvo. Tools for machine learning. In *Statistical Inference and Machine Learning for Big Data*, pages pp. 277–327. Springer, 2022. DOI:10.1007/978-3-031-06784-6-11.

[130] P. Garling (Justice). *JM v R*, 2015. New South Wales Supreme Court 978.

[131] I. Harrison (Justice). *R v Peter Tsallas*, 2017. New South Wales Supreme Court 64.

[132] R.A. Hulme (Justice). *DPP v Zaiter*, 2016. New South Wales Court of Criminal Appeal 247.

[133] G. Shmueli and O.R. Koppius. Predictive analytics in information systems research. *MIS Quarterly*, pages pp. 553–572, 2011. DOI: 10.2307/23042796.

[134] L. Rutkowski, M. Jaworski, P. Duda, and M. Jaworski. Decision trees in data stream mining. *Stream data mining: Algorithms and their probabilistic properties*, pages pp. 37–50, 2020.

[135] I. Lotfi and A. El Bouhadi. Artificial intelligence methods: toward a new decision making tool. *Applied Artificial Intelligence*, 36(1):1992141, 2022. 10.1080/08839514.2021.1992141.

[136] G. Shmueli. To explain or to predict? *Statist. Sci.*, 25(3):289–310, 2010.

[137] T. Dahiru. P-value: a true test of statistical significance? a cautionary note. *Ann.Ib.Postgrad Med*, 6(1):21–26, 2011.

[138] S.J. Clipper. *Predicting Failure to Appear: A Comparison of Statistical Techniques*. PhD thesis, University of Texas, 2016.

[139] A. Zaidi and A.S.M. Al Luhayb. Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Math. Problems in Eng.*, 2023(Article ID 5525675):pp. 1–11, 2023. DOI: 10.1155/2023/5525675.

[140] C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*, pages 62–76. PMLR, 2018.

[141] J.C.R. Martino. *Hands-On Machine Learning with Microsoft Excel 2019: Build complete data analysis flows, from data collection to visualization*. Packt Publishing Ltd, 2019.

[142] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, 2013.

[143] D. Kisselev. A Simple Interpretation of Logistic Regression Coefficients. Towards Data Science, Sep. 16 2021. https://towardsdatascience.com/a-simple-interpretation-of-logistic-regression-coefficients, Accessed Aug. 28, 2023 [online].

[144] Department of Statistics Online Programs. Applied Regression Analysis. Pennsylvania State University, 2018. https://online.stat.psu.edu/stat462/, Accessed Sep. 3, 2023 [online].

[145] Colectica. Colectica for excel 2024. https://docs.colectica.com/excel/.

[146] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J.E. Gilbert. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2022.

[147] L.M. Helmus and K. M. Babchishin. Primer on risk assessment and the statistics used to evaluate its accuracy. *Criminal Justice and Behavior*, 44(1):8–25, 2017.

[148] Y.H. Chan. Basic statistics for doctors. *Singapore Med J*, 45(4):149–143, 2004.

[149] J. Skeem and C. Lowenkamp. Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, 38(3):259–278, 2020.

[150] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[151] N. Ramakrishnan. C4.5. In X. Wu and V. Kumar (eds.), editors, *The top ten algorithms in data mining*, pages pp. 1–19. CRC Press, 2009. DOI: 10.1201/9781420089653.

[152] Department of Statistics. Probability distributions. Pennsylvania State University, 2024. www.stat.psu.edu/stat500/.

[153] D.J. Lytle. *Decision Making in Criminal Justice Revisited: Toward a General Theory of Criminal Justice*. PhD thesis, University of Cincinnati, 2013.

[154] T. Sourdin. Judge v robot: Artificial intelligence and judicial decision-making. Judicial Commission of NSW, 2018. www.judcom.nsw.gov.au/publications/benchbks/judicialofficers/judgevrobot.html.

[155] D.Fagan (Justice). *State of New South Wales v Ryan*, 2023. New South Wales Supreme Court 236.

[156] E. Firat, A. Joshi, and R.S. Laramee. Interactive visualization literacy: The state-of-the-art. *Information Visualization*, 21(3):285–310, 2022.

[157] A. Gelman. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779, 2004.

[158] A.T. Jebb, S. Parrigon, and S.E. Woo. Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2):265–276, 2017.

[159] R.A. Parker and N.G. Berman. Sample size: more than calculations. *The American Statistician*, 57(3):166–170, 2003.

[160] P. Bazeley. Integrating analyses in mixed methods research. *Integrating Analyses in Mixed Methods Research*, pages pp. 1–344, 2017.

[161] A.R. Piquero and D. Weisburd. *Handbook of quantitative criminology*, chapter Mixed Method Research in Criminology: Why Not Go Both Ways? Springer, 2010.

[162] P.F. Velleman and C. Hoaglin. *APA Handbook of Research Methods in Psychology*, chapter Exploratory Data Analysis. American Psychological Association, 2012.

[163] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde. Research electronic data capture (redcap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.*, 42(2):377–381, 2009.

[164] P.A. Harris, R. Taylor, B.L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, S.N. Duda, et al. The redcap consortium: Building an international community of software partners. *J Biomed Inform.*, 95(n.a):103208, 2019. doi: 10.1016/j.jbi.2019.103208.

[165] J.R. Edwards and R.P. Bagozzi. On the nature and direction of relationships between constructs and measures. *Psychological methods*, 5(2):155, 2000.

[166] A. Colley, S. Mayer, and N. Henze. Investigating the effect of orientation and visual style on touchscreen slider performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2019.

[167] J. Matejka, M. Glueck, T. Grossman, and G. Fitzmaurice. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5421–5432, 2016.

[168] J. Al-Hindawe. Considerations when constructing a semantic differential scale. opal.latrobe.edu.au, 1996.

[169] C. Fernandez-Loria, F. Provost, and X. Han. Explaining data-driven decisions made by ai systems: the counterfactual approach. *arXiv preprint arXiv:2001.07417*, 2020.

[170] A-M. Karimi, G. Barthe, B. Balle, and I. Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics*, pages 895–905. PMLR, 2020.

[171] M. Bancilhon, L. Padilla, and A. Ottley. Improving evaluation using visualization decision-making models: A practical guide. In *Visualization Psychology*, pages 85–107. Springer, 2023.

[172] E. Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

[173] C.E. Lawrence, L. Dunkel, M.McEver, T. Israel, R. Taylor, G. Chiriboga, K.V. Goins, E.J. Rahn, A.S. Mudano, E.D. Roberson, et al. A redcap-based model for electronic consent (econsent): moving toward a more personalized consent. *Journal of clinical and translational science*, 4(4):345–353, 2020.

[174] H. Knapp. *Introductory statistics using SPSS*. Sage Publications, 2013.

[175] D. Lane. *Introduction to Statistics: An e-book*. Apple iBooks, 2013.

[176] P. Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[177] S. Rahman. The marginal effect of bail decisions on imprisonment, failure to appear, and crime. *Crime and Justice Bulletin*, Nov.(224):1–12, 2019. ISBN 978-1-925343-73-1, Accessed [online]: July 2024.

[178] S. Yeong and S. Poynton. Did the 2013 bail act increase the risk of bail refusal? *Crime and Justice Bulletin*, April(212):1–12, 2018. ISBN 978-1-925343-61-8, Accessed [online]: July 2024.