TALLINN UNIVERSITY OF TECHNOLOGY
School of Business and Governance
Ragnar Nurkse Department of Innovation and Governance

Iren Irbe

Risk Prediction and Governance, Based on Machine Learning: A Criminal Justice Case Study

Master's thesis
Programme HAAM34, specialisation Innovation and Public Administration

Supervisor: Anu Masso, PhD, associate professor

Tallinn 2022

I hereby declare that I have compiled the thesis independently
and all works, important standpoints and data by other authors
have been properly referenced and the same paper
has not been previously presented for grading.
The document length is 15250 words from the introduction to the end of conclusion.

Iren Irbe ……
(signature, date)
Student code: 192043HAAM
Student e-mail address: iirbe@me.com

Supervisor: Anu Masso, associate professor:
The paper conforms to requirements in force

……
(signature, date)

Chairman of the Defence Committee:
Permitted to the defence
……
(name, signature, date)

TABLE OF CONTENTS

ABSTRACT

During the last hundred years, at an accelerating pace, the governance of modern society has become so complex that increasingly scientific and technical solutions are being used for governance. Among the main optimisation goals of governance is the efficient, targeted use of resources. As a result, this optimisation goal optimises the reduction of risks, reducing policy efficacy and increasing costs. The next evolutionary trend in social policy, evident in criminal justice policy, is how social governance is increasingly resembling more and more a medical field, which has led to the worldwide reduction of crime over the last thirty years: using an efficient toolbox of approaches from the fields of psychology, experimental criminology, goals of recidivism prevention, evidence-based reintegration, rehabilitation, desistance from crime, improvement of life course outcomes, victim assistance, crime prediction and reduction.

One of the central policy goals of most advanced societies should not only be the maintenance of civil order in the short term. Due to political cycles, short term goals are often a battlefield of cultural or populist political wars. A long-term goal that has become one of the central pillars of most countries' criminal justice policy is that those who break the rules of civil society will not recidivate are rehabilitated and reintegrated into society as productive citizens, and barring that, are prevented from causing harm. In contrast, crimes were mainly managed using capital punishment or exile hundreds of years ago. Today, capital punishment is infrequent, and exile is not a viable tool (although still sometimes used by totalitarian governments).

The reduction of recidivism and the total number of incarcerated people are two mutually inversely dependent parameters that can be optimised and measured. Thus, it is rational to optimise criminal justice policy to minimise recidivism within the social, legal, and political constraints. E-Governance systems are increasingly used to achieve these optimisation goals and provide a harmonised, standardised approach. These systems are guided by the predictions of assessment tools and the clinical assessment of offenders. Assessment tools can, for example, predict with various levels of efficacy recidivism risk (probability of recidivism weighed by concomitant harm) and criminogenic need (risk contributing factors that are amenable to change). These tools have

several drawbacks that are inherent to how the data for creating the model is being made: by the activity of the criminal justice system. Because the obtained data only partially explains the current state of criminality and the complexity of the human condition, there is a danger of perpetuating and amplifying existing societal problems. The threat is especially acute when the tools used are not static parametric statistical models but dynamic non-parametric algorithmic models: predicated on machine learning and artificial intelligence-based models.

This thesis adds empirical evidence to the hypothesis that dynamic risk factor-based recidivism prediction enables the self-consistent construction of non-biased and non-discriminatory risk analysis tools. Previous empirical research has shown that dynamic risk factors (factors amenable to change by the criminal justice system) have a similar predictive potential as static risk factors (factors that do not change). The practical part of this thesis validates some of the theoretical and methodological questions against the experience of practitioners using a survey. Factor analysis and formal concept analysis are used for analysing survey results. The provided theoretical framework, methodology, and empirical evidence are used for creating the survey and are informative for realising recidivism and criminogenic need prediction tools in particular and risk assessment tools in general. An additional result is that the experts value certain protective factors (factors that reduce the risk of recidivism) more than the current principal recidivism prevention methodologies like the risk-need-responsivity (RNR) model and the good-lives-model (GLM) emphasise. On the other hand, another result is a significant overlap between the expert opinions and RNR/GLM. In addition, experts wish to transition away from risk-based assessment toward a solution or a treatment-based assessment.

Keywords: risk assessment, risk, governance, criminology, recidivism, static factor, dynamic factor, machine learning, bias, critical theory

1. INTRODUCTION

The scientific hypothesis is that dynamic risk factor-based recidivism prediction enables the self-consistent construction of non-biased and non-discriminatory risk analysis tools. Empirical research has shown that dynamic risk factors (risk factors that the criminal justice system can change) have a similar predictive potential as static risk factors (factors that cannot be changed).

In addition, this thesis explores the questions on what a risk is, what are the public considerations of risk assessments, how to avoid the ratcheting effect inherent in risk assessments (biases are amplified by acting on risk assessments based on biased data), how technology interacts with sociocultural change, philosophical and ideological principles that help to align risk assessment with social realities and actionable risk assessments that are no iatrogenic. Also, the nature of data and prediction limits the applicability of risk assessment. The discussion section highlights various potential risk assessment problems on a policy level, along with proper policy framing for mitigating foreseeable issues.

The above hypothesis on dynamic risk factor-based recidivism prediction has been raised before and shown for algorithmic risk prediction, most notably by Kelly Hannah-Moffat (2005) and lately, for example, by Salo et al. (2019). This work contributes to the evidence that this hypothesis is valid also from experts' point of view, whom themselves use and create risk predictions.

For supporting the central scientific question and for exploring the ancillary scientific questions, a two-step programme is constructed:

1. A theoretical framework and methodology are described for creating the surveys for the second step.
2. Empirical evidence is obtained to support the framework and the hypothesis.

The theoretical framework and methodology will establish an epistemological, ethical, moral, and technological basis through which risk and criminogenic need assessment tools can be viewed. The empirical evidence on their construction and efficacy may be considered.
The empirical part will validate some of the theoretical and methodological questions against the experience of practitioners who deal day-to-day with criminal justice matters and have extensive hands-on knowledge of these matters. In addition to the theoretical treatment, a survey of professionals working in corrections was performed.
In future work, the third step would be to build such risk assessment tools to falsify or confirm the hypothesis empirically, but the current work will hopefully help achieve that goal.

In more detail, the first survey firstly captures the perception of the performance of static and dynamic risk and protective factors (factors that prevent recidivism) and criminogenic needs (treatment targets that reduce recidivism) using quantitative questions. Secondly, it poses qualitative questions to analyse and interpret the quantitative results.
The second operational and policy survey is for testing the theoretical analysis against the experience of professionals. Based on these surveys, ideas can be tested against the knowledge of these professionals. Also, scientific evidence from literature can be used as additional evidence.

These surveys give complementary information on algorithmic recidivism risk prediction tools and guidelines on how the datafication of risk assessments should further develop to avoid legitimacy issues in algorithmic risk prediction. Such a survey of professionals can inform algorithmic risk prediction because the algorithms are often modelled after the data used for clinical risk assessment. However, suppose even humans do not so much prioritise static risk factors for risk assessment in correctional or criminal justice settings. In that case, the nature of data used for algorithmic risk prediction should change. It is also a question of how to legitimise algorithmic risk assessment to the professionals in a way that would complement their work and incentivise the e-Governance of correct factors.

The theoretical, methodological, and empirical parts will be connected in the discussion section, where it is described how static risk factors can be used as test data to validate machine learning models. In addition, in the discussion section, a policy framework is described that takes the strengths of machine learning-based predictive models in criminal justice settings and describes how to renew evidence-based policies tailored for particular risks and needs continually.

The author hopes that this work will enable the introduction of additional low-violence-risk-preclusion strategies for reducing incarceration rates by improving the utility of various correctional policies (Reitz, 2020).

2. APPROACHES TO RISK ASSESSMENT

2.1. Approaches to risk

Risks can be defined in many ways. Still, for this thesis, an appropriate one is that **risk** is a probabilistic quantity that allows to base decisions that prevent some undesirable outcome (usually known as harm) due to some classification by crime or some other feature. In other words, it is a set of beliefs multiplied by the cost of undesirable outcomes (Berk, 2019, p. 60).

When looking at how risk is used, it becomes apparent what risk is in practice. After looking at applications and various definitions, a contextual understanding might arrive that can bring to understanding its social, ethical, philosophical, and legal aspects. The fact that risk is not a purely mathematically definable concept and requires an accompanying context does mean that it is socially constructed. In other words, the social context precedes its definition. As such, its meaning shifts through history and now, with the era of BigData, e-governance, and algorithmic learning, e.g., artificial intelligence (AI) and machine learning (ML), the exact meaning of risk shifts again (Kasapoglu & Masso, 2021).

In the criminal justice context, what is the utility of risk? Is it to establish the proper punishment for potential future misdeeds or the necessary services to facilitate reintegration and rehabilitation? Risk can be conflated with dangerousness (Dershowitz, 1970). Dangerousness usually refers to the risk of harm due to predicted violent behaviour, but risk assessment tools can also assess flight risk, parole violation risk, and other risks. If risk assessment tools' risk is defined as the probability of reoffending multiplied by the cost of the probable violation, it does not mean risk of violent offending. In practice, the risk is calculated for different crime classes or misbehaviour. The definition of risk can also depend highly on sentencing principles and the judge's discretionary space, and that is where evidence-based sentencing principles can help (Oleson, 2011). Prediction of dangerousness has the unfortunate possible side-effect that it is used as a way of preventive punishment. It thus has the potential of circumventing the proper judicial sentencing process and

of causing disproportionate harm, primarily due to the relatively poor false-positive and the true-positive ratio of risk assessment tools, which is often over 1.5 on average (Tonry, 2019). These issues can be mitigated by judiciously having different thresholds depending on the type of harm predicted by the risk assessment tool (Berk, 2019, p. 48).

MacArthur Violence Risk Assessment Study (Monahan, 2001, p. 129) thoroughly analysed violence risk assessment from the clinical perspective. It was pointed out that the belief that mental disorders often result in violent behaviour has persisted through history but has intensified of late. This interlinking has created a stigma around mental disorders, and because of this, people who need treatment deign to have it. This shift from a parentalist attitude toward treating mentally disturbed people to protecting the public from them occurred in the United States around the 1960s. The cold-war 1960s also saw the advent of the risk management industry, preventative treatment as public protection from harmful behaviour, from "dangerousness".

Development of risk management led to: "concern about potential liability (of clinicians) is leading some clinicians to participate in the creation of a system of preventive detention for persons thought likely to commit violent acts." (Appelbaum, 1988, p. 780) and a punitive system parallel to the sentencing-based system arose. Dershowitz (1974) describes how preventative detention is theoretically abhorred, especially if irrelevant to criminal guilt, but is accepted on a practical level. Preventative detention can also lead to misdemeanour statutes to confine people with suspicious behaviour. By using risk assessments as a tool in the criminal justice system, preventative detention is judged to be acceptable and is a form of preventative justice. Now, to what extent is it acceptable? Where is it acceptable to limit a person's human rights to protect the rights of others? Dershowitz (1974) describes how in the 18th century United States and the United Kingdom, the assessment of "dangerousness" was the prerogative of judges. However, as the justice system is handling more people with fewer resources, responsibility for the offenders' initial assessment and management devolved to various specialists: from pretrial and investigation to incarceration to prosecution to parole and probation, from social services to education. Risk assessment tools and clinical judgement protocols are there to standardise the assessments and provide a systematic approach.

Previously, preventative confinement was used to belay the risk of immediate harm. Here, the result-focus of the criminal system did not consider the "risk of harm" in punishing a felony. If the person were declared insane, they could be confined indefinitely, which the prosecution did, as it

wanted the maximum legally possible sentence. Thus, preventative confinement can be used as a workaround to punish more extensively than the law would allow if the person were not a felon.

2.2. Governing risk in public institutions

With the invention of modern public institutions, governance of the institutional processes became one of the central issues. Without governance, it will become virtually impossible to ensure that the processes work as intended and allocate resources properly according to need. It is also possible to design policies that target desired outcomes with good governance. In addition, the standardisation of said processes may enable structured scientific study and evidence-based governance modification due to the ability to control confounding variables better and gather higher-quality data. In other words, governance of public institutions must not be based on gut feeling, the prevalent fad of the moment or a bureaucrat's whim: it should be a result of systematic research and development. Former is at least how it should be, and as a goal, most modern public institutions should try to achieve.

Criminal justice is a case study, as an evidence-based approach has led to the realisation that evidence-based processes and tools must be used to figure out how to reduce crime and recidivism. Evidence-based practices lead to risk assessment at distinct stages of a criminal justice process, not just to standardise how to measure the risk of harm but also to measure the likely level of needed care or intervention and whether there is any improvement and thus what could be fine-tuned. In an ideal world, such tools would contain evidence-based models that help to optimise the process to achieve the best outcomes.

Risk assessment in particular and criminal justice forecasting began in earnest in the 1920s (Burgess, 1928) with the Illinois Parole Board Study (Bruce et al., 1928). It was also posited in a more general formulation: is prediction possible in social work using statistical methods (Burgess, 1929). The Illinois Parole Board study was the first to analyse over 20 years of data on how indeterminate sentencing works, based on over 3000 criminal and penal records. Furthermore, it gave the first extensive description of risk assessment.

After the Burgess study, an actuarial approach was rapidly adopted in criminal justice and other areas of governance and a slew of novel studies were made. Immediately caution was also advocated because it was noted that many factors could be either overlapping or irrelevant and had a correlation with recidivism outcome purely by happenstance (Tibbitts, 1932).

Historians of criminology and social institutions can be divided into two groups. Some use history to legitimise current approaches and the reverse: some believe that all social ills can be cured or mitigated and that current problems are more due to a lack of effort. Moreover, some say that the same institutions were used to enforce inequalities within society by dominant groups to enforce their dominance or hegemony. Of course, this polarised view is a bit of caricature, as there is a middle ground between these opposites.

2.3. Approaches to risk assessment

A critical data theory (Boyd & Crawford, 2012) is explored in this thesis to address the moral hazards of risk prediction instruments within e-governance. The theoretical treatment combines critical theory, pragmatism, and process philosophy: a delicate balancing act between the absolute and the relative. The theory is supplemented by the empirical measurement of views of the prospective users of such predictive instruments in the "Results" section 4.

2.3.1 Critical theory

This thesis aims to address the issues arising from the recognition that the diversity of society is not viewable as a one-dimensional system of categories. This fact seems elementary but has been explicitly recognised on a policy level in Western democracies only very recently, mainly during the decade starting from 2010 when social scientists recognised that immigrants cannot be categorised as singular blocks by ethnicity or by religion, for example, but that a multidimensional view of diversity is needed as a methodological concept. This concept has become known as superdiversity, named by sociologist Vertovec (2007), and is also reflected in data and thus in technologies that consume it (Taylor & Meissner, 2020; Masso, 2021). Diversity also means diversity of viewpoints and that there are several valid epistemological sources of knowledge. For example, standpoint epistemology is essential to understanding the sociological aspects of

diversity and the context of social phenomena. Standpoint epistemology is one of the primary sources of knowledge in critical theory, intersectionality and superdiversity studies, and ethnographic studies. Due to the philosophy and practice of science, some sources of knowledge are empirical, and some are the result of theoretical reasoning over abstract *Gedanken* experiments. It behoves not to disregard diverse sources of knowledge if this helps achieve policy aims such as reducing crime and creating a safer society. This linking of different schools of thought is relevant because it helps via non-ideological ways. For example, pragmatism allows many conversations from different viewpoints: pragmatists themselves range from the arch-conservative C. S. Pierce to socialist J. Dewey to Marxists. It is also interesting that pragmatism can serve as a theoretical basis for liberal democracy (Green, 2006, p. 306).

Critical theory is a social philosophy about a critical approach that emphasises that philosophy needs to investigate culture, human knowledge, and morality. In other words, practice is also an epistemological source of truth. It is a school of thought created by the Frankfurt school of philosophy. For example, Theodor Adorno, Max Horkheimer, Jürgen Habermas, and Herbert Marcuse have developed it, and initially, Theodor Adorno called Horkheimer's work with this moniker (Ghiraldelli, 2006, p. 202). In the literature, there is some confusion around critical theories because this term covers both philosophical and sociological claims in the work of different scholars. As a result, sometimes philosophical claims are taken as sociological and the reverse, leading to various misunderstandings.

As first expounded by Max Horkheimer, critical theory rose as criticism and revision of Marxism. *Marxism* can be defined as: "Marxism examines the historical, social, and economic conditions for the possibility of culture and knowledge" (Ghiraldelli, 2006, p. 202) and that a revolution by the working class is necessary to reorganise society, and production, culture. One of the central research subjects in Marxism is the tension between the laws of social organisation ("relations of production", "superstructure") and production ("productive forces", "structure") (ibid., p. 202). The critical theory focuses more on the "superstructure" (ibid., p. 202), while Marxism has traditionally focused more on the "structure" (ibid., p. 202).

According to Horkheimer, Marx made three fundamental mistakes (ibid., p. 203). Firstly, by trying to understand other cultures through the lens of Western civilisation, he did not draw proper lessons. He also fetishized progress and control of the physical world, which was, to be fair, the general goal of the Western civilisation during the 19th century. Thus " he did not realize that

technology belongs to the realm of necessity; it is a realm that sustains the suffering of nature, and what rests in the realm of freedom is our solidarity in favour of life, our demand for social justice and appreciation for nature." (ibid., p. 203). He also had the mistaken belief that class struggle would lead to social peace. Horkheimer, Adorno and Marcuse also rejected the idea of Marxist asceticism and political revolution due to having drawn their conclusions from what happened in Russia in 1917 and afterwards (ibid., p. 203). Thus, critical theory concentrated on critical discourse rather than the revolutionary course.

According to Adorno and Horkheimer, the principal methodological idea of critical theory in the interpretation of Ghiraldelli (ibid., p. 204) is the appreciation of the converse of the Hegelian notion that "rational is real and real is rational" (ibid., p. 204). What is real is a denial of the rational if a rational life is a life of goodness and freedom without unnecessary hardships. The context in critical theory is a "historical account about the irrationality of the rational during modernity" (ibid., p. 204).

Adorno and Horkheimer turned attention away from the Marxism fascination of class struggle and to questions about "Enlightenment" and modernity and focused on criticism of "Enlightenment" and modernity to protect the goals of "Enlightenment" (ibid., p. 203). The core axiom is "freedom in society is inseparable from enlightenment thinking" (Horkheimer et al., 2002, p. xvi). The second axiom "... institutions of society with which it is intertwined already contains the germ of the regression which is taking place everywhere today" (ibid., p. xvi).

Danah Boyd and Kate Crawford applied critical methods for studying various issues surrounding big data (Boyd & Crawford, 2012). As mentioned in the introduction, Chauncy Starr (1969) described the need to analyse different affected standpoints. Only utilitarian and positivist approaches are improper when technologies have significant social, political, legal, and safety impacts. In other words, it is necessary to avoid the rabbit hole of techno-utopianism, technological determinism, and scientism. As a reaction to techno-utopianism, technological determinism, and scientism during the 19th century, many new schools of thought arose alike: continental philosophy, activist philosophies like Marxism, critical theories, poststructuralism, postmodernism, pragmatism, existentialism, absurdism, hermeneutics, semiotics, and many others. They rose to prominence to balance positivism, analytical philosophy, structuralism and similar because, in their view, these schools did not provide constructive solutions to existing problems. Critics called this the crisis of analytical philosophy, and it continues today (Margolis,

2021). It can be said that critical data theory is a novel approach to address questions that were asked already during the Industrial Revolution. "It follows from Horkheimer's definition that a critical theory is adequate only if it meets three criteria: it must be explanatory, practical, and normative, all at the same time. It must explain what is wrong with current social reality, identify the actors to change it, and provide clear norms for criticism and achievable, practical goals for social transformation." (Bohman et al., 2021).

2.3.2 Phenomenology of risk assessment

When the risk is defined along with the applicability of risk assessment and the composition of the data for risk prediction will become dependent on each other not just on a practical level but also due to the context of the risk prediction and the categories being used. Thus, when thinking about how these categories are created and how the composition of training data of machine learning models is determined, it is necessary to also think about the metaphysical questions about the nature of knowledge and the cognitive processes of humans. To wit, "account of the general character of what we know must enable us to frame an account of how knowledge is possible as an adjunct within things known" (Whitehead, 1948, p. 158).

This leads to another school of thought, adjacent to pragmatism and in opposition to the classical speculative philosophy called process philosophy (speculative as in opposition to process), which is based on the statement that the former suffers from the "fallacy of misplaced concreteness" (Whitehead, 1948, p. 52). It is a critical view of a separation between the standpoint that investigation of the abstract provides a complete understanding of reality vs the study of the concrete. Classical thinkers and first scientists like Galen of Pergamon try to avoid the abstract by admitting that the full understanding and description of the concrete is a non-convergent process, "individuum est ineffabile", but that final perfect generalisations are not possible (Edelstein, 1952, p. 303). In addition, according to Edelstein (1952, p. 303) first instances of true empiricism became possible when the alliance between the abstract and concrete, an alliance between ontology and personal individuality, became possible. According to Whitehead (1948, p. 87), additional difficulties are that the concrete is ever-changing, and it is impossible to describe the concrete using the abstract, e.g., language, in a final manner. Thus, a metaphysical reality needed to be structured and ordered. The hope was that using the process of abstraction, the essentials of the concrete could be investigated.

Critical theories try to approach from the other direction. By drawing from the source of standpoint epistemology, these theories attempt to describe reality first and leave the abstractions relative and unfixed. Process philosophy draws upon constructivist epistemology. For deconstructionists like Derrida (Rorty, 1978), this leads to an infinite conversation about the radical singularity of reality about the infinitary properties of knowledge. In other words, classical philosophy believes in the stability, if not even immutability, of knowledge about the truth. Although, this is especially useful in discovering new patterns that can lead to new knowledge and new hypotheses for testing existing abstractions and finding new abstractions (e.g., models or algorithms). However, another potential fallacy here is that order (determinism) and simplification (preferably models with a fixed number of parameters) are necessary to find how seeming indeterminism of phenomena is deterministic. That order and simplification are more fundamental. During the 20th century, a whole field of non-linear and non-deterministic phenomena was found in applied sciences. Machine learning models are mostly non-linear, probabilistic, non-parametric, and use pre-knowledge (*a priori*) about what was before, e.g. Bayesian probability estimates *a posteriori*. The fear about machine learning models is also related to the control over the non-parametric nature of modern ML models about the implications of a priori knowledge used by the models. On the other hand, is the hidden structure of *a priori* knowledge and the world model of humans controlled? Complete control *ipso facto* assumes perfect knowledge.

Seeking deterministic, simplified, immutable essence and truth is fundamental to the modern scientific method, as it was, for example, for Plato and Aristoteles (among other things, Plato's theory of ideal forms or Kant's "thing-in-itself"). The ideal striven for by Popper's falsification principle that a scientific theory must be falsifiable is in a hidden way equivalent to the search for absolute truth as was demonstrated by Duhem-Quine thesis: apparent scientific falsifications are impossible, as there are always auxiliary assumptions and behind these, there are additional extra assumptions until there is a set of atomic principles, axioms, that are assumed to be true (Stanford, 2021). Alternatively, in other words, all theories face the epistemic challenge of underdetermination. Underdetermination can be divided into two categories: holist underdetermination and contrastive underdetermination. In the first case, if there is a need to abandon a hypothesis, it also means that the auxiliary hypothesis needs to be abandoned. In the second case, there might be equivalent theories that can be confirmed by confirming the hypothesis. Underdeterminism can threaten the rationality of science itself. While the issue of underdetermination might be taken to exclude rational defensibility of knowledge, a more

generous interpretation is that the improvement of the web of knowledge is a continuous optimisation process that requires constant belief revision that optimises toward rational defensibility theories (Stanford, 2021).

When reality is being analysed, the distinctness of perceptions is predicated on the ability to identify using the perception of reality. However, suppose contexts of experiences are applied to the perception process. In that case, no absolute distinctions exist and describe the fundamental problem behind experiences and when actuarial machine learning models are created based on data. Fundamentally, this is because modelling social phenomena is also a sociological phenomenon.

A consequence of the abstraction process is that the abstract can seem to be more accurate than reality itself. This is one of the main current criticisms of classical and analytic philosophy. Pragmatism and process philosophy is supposed to address this criticism: i.e., classically, the abstract sustains the concrete, while in process philosophy, the abstract or *eternal objects,* as per A. N. Whitehead (1948, p. 175), are sustained by the perception of reality.

Such a reversal is not a new idea, as displayed, for example, by Giambattista Vico's principle of *verum-factum*: "verum et factum reciprocantur seu convertuntur", which states that "truth needs to have a constructive property" (Honneth et al., 2008, p. 5). Vico argued that civil life also has constructive property. Still, that does not mean that what is true always has a *constructive property*. Honneth et al. (2008) describes Heidegger's notion of care in connection with the constructive property and that recognition must precede cognition and ontogenesis. In other words, according to Heidegger, "care" precedes "scientific" knowledge of behaviour (ibid., p. 47). Adorno adds that the preciseness of knowledge is dependent on the acceptance of as many perspectives as possible (ibid., p. 46). In conclusion, the constructive property is crucial for critical theories and process philosophy and the next topic, pragmatism.

Reification is a socially engineered pathology that, according to Lukács, is the result of a process that affects the following social relation dimensions (Stahl, 2018):

- Features of objects.
- Interrelationships of people.
- Intrarelationships of people.
- Interrelations of people with society.

Due to reification, these relations can be described quantitatively, and on the qualitative level, these acquire inanimate qualities and lose subjective qualities. This objectification of social relations isolates people. Objectification is one of the main arguments against risk assessment tools or for assessing the qualities and behaviour of people via algorithmic predictions (author's claim). This kind of commodification of subjectivity, reification, is a severe problem in critical data theory. Reification is a vital concept Frankfurt School of critical theory and even is connected to Heidegger's "Being and Time." In this work, Heidegger strove to break the dominant worldview, which relegates the human being (s. k. *Dasein*) to a "thing among things", and for that, the affirmation of people's characteristics is necessary (Honneth et al., 2008, p. 70). Adorno points out that it is needed to improve the preciseness of our knowledge for countering reification. That accuracy depends on the extent of recognition and acceptance of a multitude of perspectives. Such a stance is called a *recognitional stance,* and Heidegger calls it "care" (s. k. *Sorge*), "solicitude" (s. k. *Fürsorge*), and pragmatist John Dewey calls it "involvement" (ibid., p. 46).

Just like postmodernism and poststructuralism are part of French cultural tradition, so is pragmatism in the United States. It can also be understood as a particular set of social practices, strengths, and weaknesses of its society with a solid moral impulse: "…mode of thought that subordinate's knowledge to power, tradition to the invention, instruction to provocation, community to personality, and immediate problems to utopian possibilities" (West, 1989, p. 5). It is not a uniform school and thus is not easy to define, and it is diverse and heterogeneous, consisting of many complementary approaches. One of the reasons for being compatible with critical theory and postmodernism is its future-oriented instrumentalism and its activist bent on effective action. Compared to the analytical tradition of philosophy, pragmatism is a significantly more utilitarian school that dares to deal with social theory, cultural theory, and historiography. It is not only a mode of transcendental thought. Pragmatism deals with relations of knowledge and power, cognition and control, discourse, and politics (West, 1989, p. 3). It is interested in how social constraints are created through hierarchies due to race, class, gender, and sexual orientation.

According to the Aristotelian conception of nature, "everything which possesses any power of any kind, either to produce a change in anything or to be affected even in the least degree by the slightest cause, though it is only on one occasion, has real existence" (Plato, 1997, 247d–e, 269). During middle-ages, the theory of powers became the primary explanatory framework for different phenomena, and it was thought that God gave all creatures their essential natures, causal powers (Hill et al., 2021, p. 3). In contrast, in pragmatism, the epistemologies are constructive because, in

pragmatism, the competition between different narratives defines the truth (a weaker definition). In the author's view, constructive because they incorporate Popper's notion of a proof (falsification theory) while not excluding the Aristotelian epistemological approach.

2.4. Risk and needs assessment tools in criminal justice

In modern evidence-based criminal justice systems, risk-based assessment tools proliferate as parts of a more comprehensive integrative approach for handling people who have run afoul of the law. These tools help the criminal and social justice systems optimise prevention, harm, recidivism risk reduction, and protection of law and order. In theory, these instruments enable to tailor strictures against a common standard.

The justice system needs to be flexible to calibrate sentencing to optimal outcomes to enforce the law. Lay people usually view punishment from the just deserts perspective and apply the moral proportionality principle: punishment must be proportional to the blameworthiness of the offence. The issue is that if the aims of the justice system diverge and punishment is viewed by it from a utilitarian, deterrent perspective, the moral authority of law will be diluted (Carlsmith et al., 2002). When adding to punishment the rehabilitative and reintegrative goals, there will be a clash between various purposes of punishment (Focquaert et al., 2021, p. 157). Besides the prosecutorial process, sentencing, and punishment, the consequences of a conviction can resonate throughout the life course of offenders and their community in both legal and extra-legal sense.

In a way, a conviction can result in a perpetual extra-legal punishment and is a grave concern when criminal justice policy is designed (O'Reilly, 2018, p. 203). For the criminal justice system to maintain its moral authority, results must be shown. By being "tough on crime" or claiming, "nothing works", results can be quantified and easily demonstrated. Still, reintegrative/rehabilitative results are much harder to show and that, for example, led to the unfortunate hard-line turn in the U.S. criminal policy. One of the most regrettable results was the misreadings of Robert Martinson's criticisms of the U.S. prison system in his 1974 article (Martinson, 1974) that politicians used to claim that "nothing works" to rehabilitate criminals. By criminologists and social scientists, this was and still is considered a misreading of what he wrote (Martinson, 1979). Still, it resulted in a wave of political opportunism, and in the end, Robert Martinson committed suicide. This tragic event is considered one of the best examples of how

social science can be misrepresented and misused. It can be said that criminal justice policy is very much tied to the political climate, as it is not just a technocratic governance issue. Cullen (2006) has written how his surveys showed this at the height of US Martinson's "Nothing Works" doctrine (ibid. p. 666) that there was significant public support for rehabilitation and reintegration and still is.

So-called "just deserts" based on punishment *vis-à-vis* incarceration cannot be the primary vehicle for prevention of recidivism due to the "Iron Law of Imprisonment": "they all come back" (Travis, 2005, p. xxi), and thus it is not a matter of belief whether rehabilitation and reintegration work: it is an unavoidable necessity to make it work regardless of the current state of the art and to be "smart on crime", instead of "tough on crime" (Epperson & Pettus-Davis, 2017, p. 4). It behoves effective criminal justice that scarce resources must be allocated smartly, as poorly targeted efforts can lead to no net positive or even iatrogenic effects (Welsh et al., 2020). For example, it is widely known that incarceration can increase crime and that poorly calibrated intervention during the life course of youth can also have iatrogenic effects and is partly mediated by various life chances (Bernburg & Krohn, 2003). Thus, a life-course focused approach to criminal justice has become the predominant approach in many democracies, including Estonia, and the focus has moved towards a decarceration of society (Epperson & Pettus-Davis, 2017, p. 126). These instruments are tools in developmental and life-course criminology that focus on antisocial behaviour and offending using longitudinal research and influence these behaviours over time (Sampson & Laub, 1992).

Criminal justice policies and governance encompass numerous complex processes with decision points that have inherent risk: there always is a probability of harm if a decision is wrong. This thesis will contribute methodologically and policy-wise to designing these systems to reduce harm and improve outcomes when machine learning is used for governance. Motivation is ample evidence of such predictive instruments improving criminal justice outcomes, including sentencing quality (Etienne, 2009). However, the use of such instruments is also fraught with various risks: reinforcement of existing social processes that can further establish various inequities and thus make the social situation worse for multiple groups of people. Criminal justice involves the waiver of certain human rights. Still, a prediction can also involve the waiving of aspects of due process and can become its own shadow criminal justice process if implemented in an inconsiderate manner (Starr, 2014).

In the case of criminal justice, this means that various kinds of predictive instruments are used within the whole criminal justice process, depending on the process step, the nature of the infraction, and other circumstances. These systems are based on the concept of risk and risk of harm. If risk governance is used for nuclear energy, the data for said governance is impersonal. However, if used within criminal justice settings, the composition of said data matters differently because if an evidence-based understanding of the causes of crime and prevention of recidivism is incorrectly used, risk governance can go awry and cause serious social harm (Simon, 2007, p. 13). The problem can arise because data on individuals reflect the diversity among individuals and reflect their circumstances. If the risk is managed using variables mediated by social injustices, then using these variables poses the danger of reinforcing and perpetuating these same injustices (Starr, 2014). Risk assessment instrument prediction precision poses additional problems: even though false negatives reflect the imprecision of predictions in a neutral manner, false positives delegitimise the criminal justice system if these false positives are based on biased data (Oleson, 2011, p. 1368-1393).

The risk assessment methods and tools are classified according to the following maturity levels (Casey et al., 2011):

1. Clinical professional judgement (CPJ).
2. Evidence-based tools. From the beginning of the 1970s, these became into wider spread use in the United States and Canada. These tools quite reliably differentiate between high-risk and low-risk offenders. The data used for prediction is mainly historical and empirical without any theoretical underpinnings. These tools were insensitive to the changing circumstances of an offender.
3. Evidence-based and dynamic tools. These tools also consider dynamic risk factors like family relationships and are often referred to as risk-need assessment tools (RNA).
4. Integrating intervention and monitoring. These are based chiefly on the risk-need-responsivity (RNR) model. It has a relatively long development cycle that can take years.
5. Algorithmic machine learning and artificial intelligence-based tools are trained based on current and historical data and integrate prediction (risk assessment) and measurement (results, responsivity, monitoring) into a cycle that continuously refines and trains new risk assessment models to produce specialised assessment models as well. Prediction results can be calibrated to achieve political and policy decisions (Berk et al., 2021).

First, these can give a parole risk tailored for the stage of the criminal justice process and the type of risk (recidivism of crime, breaking parole rules, not showing up in court and similar). The others recommend treatment, rehabilitation, sentencing, reintegration, and parole rule creation. The second type is about predicting the ***dynamic risk factors*** that can be addressed. A dynamic risk factor broadly is a variable, like educational attainment, that can be changed vs a ***static risk factor*** like age or criminal history that is immutable (Hannah-Moffat, 2005).

The duals of risk assessment tools in criminal justice are ***criminogenic need assessment tools***, where the central concept is treatment and need, not risk (Hannah-Moffat, 2005).

2.5 Risk assessment, ethics, and ML

In recent decades, fairness, and transparency in the era of dataveillance (data surveillance) have become hot topics even though ML- and AI-based risk assessment tools are hardly more biased than the previous generations of risk assessment tools. When discussing the ethics of risk prediction using actuarial means, it must be considered that actuarial risk assessment is based on models calculated using various statistical procedures. If ML is considered statistical learning, then it can be said that statistical learning has been used for actuarial risk assessment for almost a century.

When the ethics of risk assessment is considered, it is necessary to understand that the concepts of fairness, transparency and bias have changed a lot with the evolution of the collective cultural psyche. It is a sign of the development of society that an understanding of how the collection and processing of data can radically affect lives, for better or worse. It is understood better how the ethical implications of risk assessments can have social implications that perpetuate various social issues or worsen these. Here again, the use of dynamic vs risk factors becomes relevant, as the use of static risk factors indirectly sanctions additional punishment based on what cannot be changed.

The heterogeneity of different processes and decision points also captures heterogeneous statistical phenomena. Thus, there is a risk that if these are not adequately understood, predictions and optimisations can lead to adverse outcomes. The following aspects can be measured: what is used for creating predictions, how predictions are used, and how prediction outcomes are balanced for different affected social groups vis a vis superdiversity and intersectionality. It is not only a

mathematical or technical problem, it is a social science and policy problem as well, and there are also ethical and moral questions that need resolution or at least balancing (Masso, 2021), and where an excellent theoretical framework is needed to ground those choices systematically.

Today the invention of better statistical methods combined with a much larger, more comprehensive dataset means that it is possible to create models that can better avoid bias. Human reasoning cannot be captured using classical propositional logic or logic that includes various modalities like certainty/ambiguity. Humans use common sense reasoning that combines knowledge about certainties that can be quantified using subjective probabilities (Minsky, 1986, p. 306). At the same time, humans suffer from various cognitive biases, use many shortcuts, are often inconsistent in their reasoning, and this can lead to large variabilities when assessing a person during their passage through the criminal justice system and this can lead to various forms of unfairness, including sometimes discriminatory behaviour (Kliegr et al., 2021). Due to ethical and regulatory concerns, a developing field of ML/AI is to create auditable, adjustable models, and their biases can be analysed and corrected. A handy feature of probabilistic ML/AI models is the constant adjustment of the models based on predictive ability and incoming data. Current systems are mostly actuarial and are fixed for many years, while ML/AI models are practical tools and constantly run adjustable scientific experiments. In other words, these give the ability to continually assess and improve its usefulness based on **evidence in real-time** (Berk, 2019, p. 151).

Risk and needs prediction models are updated based on correctional efficacy and provide the interpretability (auditability) of predictions (Rudin, 2019). As a result, ML models can, under certain conditions, perform at or better level than humans (Lin et al., 2020), reduce bias (Kleinberg et al., 2018), enforce the latest criminal policy standards (Hamilton, 2020), and can help combat heuristics and cognitive shortcuts in risk assessment (Hester, 2020). Multiple studies have shown that "evidence-based practices" can significantly (>10%) reduce recidivism (Etienne, 2009). In addition, risk assessment tools can help in increasing the average quality of sentencing (Harris, 2006). These tools can also provide additional scientific evidence for basing sentencing outcomes and finding the best evidence-based correctional measures with increasing predictive power. Of course, with any device, there is a danger of abuse if the predictions of these tools are employed without due caution. Besides recidivism reduction, risk assessment tools can also be used for reallocating prison-bound offenders to non-prison alternatives (Ostrom et al., 2002, p. 4). For example, in U.S., Pennsylvania and Wisconsin states, reduced sentencing programs take as an eligibility criteria risk-need-assessment (RNA), and offenders can choose to participate in these

(not mandatory), and the Department of Correction can take the results into account in selecting correctional methods (Carter & Sankovitz, 2014).

In 1928 E. W. Burgess showed how actuarial risk assessments could recommend how an offender should be treated. For example, the intelligence of an offender counterintuitively is not inversely proportional to recidivism risk (Burgess, 1928, p. 231). Already Burgess pointed out that risk assessment tools need to be adequately validated and evaluated, and here there are many problems with existing risk assessment tools. Some tools are not assessed, and sometimes, both how the instrument works and the instrument evaluation results are declared business secrets for proprietary tools. The execution is occasionally problematic during the evaluation process: for example, the same training and test data are used, leading to inflated accuracy numbers (Berk, 2019, p. 120).

Evidence-based tools, including risk-need-responsivity tools, help in answering many crucial questions and thus have become *de rigueur* in most democracies, especially in the U.S. (Garrett, 2019), Canada (Kroner et al., 2020), U.K. (Debidin, 2009), France (Hodgson & Soubise, 2016), Finland (Salo et al., 2019), to name a few. Due to this, there is a large body of research on the efficacy of these tools and the various models and underlying assumptions.

3. METHODOLOGY

The research design is non-experimental because analysis and measurement are done *post hoc*, and no clear control group can be analysed. This thesis analyses qualitatively via interviews with experts on criminal policy, corrections, risk, and needs prediction instruments. Both exploratory and descriptive questions are asked here as subject matter experts can answer these. Diagnostic and design questions can also be beneficial in making qualitative analysis better.

Two survey questionnaires were created to validate the theoretical approach: a factor survey and a service design survey. Survey questionnaires are based on the theoretical treatment and RNR (Andrews et al., 1990) and GLM (Ward & Brown, 2004) models. A more thorough discussion on these can be found further in Chapter 4. All qualitative questions were duplicated by ordinal valued questions (quantitative). Questions were combined with an ordinal scale (nominal variables) with qualitative questions (categorical variables) to weigh and validate the ordinal scale questions. Ordinal variables are encoded into the five-point Likert scale from -2.5 to 2.5. Likert scales are a set of ordinal scale grades; on Likert scales and scores and their use, see further F. P. Irwing (2018, p. 9-17). Some questions affirm the negative, and their results were flipped (for example, score 2 becomes -2). The questions were encoded into nominal boolean variables by subdividing each question into its conceptual terms. Before the final analysis, all scales were normalised to the unit interval [0,1]. Some questions capture dominance data where the user must compare different risk assessments.

For a pilot study, the author got a reasonable number of responses, with a request/response ratio of 2.57 (54 requests, 21 replies). According to the literature, the best approach for such a small sample size would be to use Bayesian Factor Analysis. Bayesian methods make fewer assumptions on the underlying distributions of variables and incorporate a priori qualitative information (Berk, 1995). On the other hand, that does not mean that the results are statistically highly significant, and thus it makes more sense to analyse responses to questions more qualitatively.

All quantitative analysis was done using google sheets for pre-processing and R programming language for the actual analysis. In the first step, formal concept analysis was done to understand qualitative properties and associations between questions, and in the second step, factor analysis was done. Formal concept analysis was done to help in having a context for interpreting the factor analysis results. For that, each question was subdivided into concepts, and thus for each question,

there was a boolean vector or row. Each concept was in a separate column, designated for that concept and combined into a matrix. Each row of that matrix then was multiplied by the Likert score of that question. A further enhancement was to weigh each row by the number of concepts to reduce the impact of questions with multiple questions.

Next, seriation of the matrix was performed to simplify and cluster the matrix for the subsequent analysis step. Seriation – is when rows and columns are shuffled so that non-zero values would cluster at the diagonal of the weighted concept matrix. Seriation does not change the qualitative properties of the matrix but would make its visualisation much more informative as it groups questions based on concepts (Liiv, 2010).

The R package "fcaR" (López-Rodríguez et al., 2021) was used for formal concept analysis (Ganter & Obiedkov, 2016), which generated a set of concepts and a set of associations between the concepts. These associations are represented as a hierarchy (Wille, 2009).

Next, the concept to question matrix was analysed using the R psych package (Revelle, 2017) to reduce the rank (number of linearly independent rows or equivalently the number of columns/dimensions) of the obtained matrix to uncover the structure of the concepts. There is latent variable exploratory factor analysis (EFA) using principal axes factoring (PAF) as an estimation procedure (Fabrigar et al., 1999). EFA was used because it does not assume any *a priori* assumptions and PAF. After all, it does not assume any distribution of the variables. In principle axes analysis, the eigenvalue decomposition of the matrix is done, and then the first n factors are used for estimating commonalities. The procedure is repeated after the commonalities are entered into a diagonal matrix until the diagonal sum does not vary more than a present amount.

In addition, a reliability analysis of the results was performed by calculating McDonald's omega coefficient (Revelle & Zinbarg, 2008) using the R programming language package "psych" for assessing the factor analysis construct validity. In addition to providing a reliability estimate omega, this analysis performed hierarchical factor analysis, rotating factors, and then performs second-level exploratory bi-factor analysis by performing Schmid Leiman transformation that extracts a general factor that will be used for calculating the omega reliability estimate, so-called general factor saturation coefficient (Revelle & Zinbarg, 2008). As a result, factor analysis discovered the factors that best fit the observed questionnaire. Based on factor analysis, factor clusters were found.

Another method called item cluster analysis (ICLUST) was used as a complementary method (Revelle & Zinbarg, 2008). The Likert scale enables negative and positive weights, which are reflected in factor analysis weights and enable the comparison of protective factors vs risk factor related questions, for example, as protective and risk factors are diametrically opposite.

For each survey quantitative results are briefly discussed along with a discussion of the answers to qualitative

4. RESULTS

The author has developed three questionnaires to assess three dimensions of the thesis.
First is "Factor Survey", which is on risk and protective factors relevant to risk and need assessment (factor survey). This questionnaire is necessary for assessing which factors can be deduced from existing data, which can be clinically assessed or obtained using self-surveys.
Second survey is "Service Design Survey" on service design that would give input into how the human-computer interaction could be designed so that the tool would be complementary for work with offenders and not interfere with professional discretion. It would also map the major policy issues.

All two surveys were combined into a single google form for easy input. To increase the probability of survey completion, all questions were labelled optional, and the most important but quickly answerable quantitative ones were pushed to the front of the survey. Questions were designed to establish the belief system of the interviewee regarding whether people can change, is desistance possible and their hope for rehabilitation and reintegration. It mainly comes out when one asks about the protective factors. Another vital topic is how the experts see risk, which leads to the subject of risk assessment. The experts were queried on different risk factors, protective factors, and criminogenic needs defined in the dominant RNR (Andrews et al., 1990) and GLM (Ward & Brown, 2004) models.

A subset of these problems was selected to avoid making the questionnaire too lengthy, and quantitative questions were scored on a linear scale from one to five and then normalised to -0.5 to 0.5. Questions could be commented upon, and the comments were considered during the analysis of the survey results.

The author combined the following concepts drawn from RNR and GLM models and based on author's service design experience into following topics:

- Measurement: How to measure risk and provide meaningful and useful predictions to the users. Quantitative and qualitative measures need to be consistent and comparable even if the predictive models change.
- Trust: The success of a correctional tool requires a reasonable level of trust from the public, officers, officials, and motivation from offenders to engage and not just to play the system.
- Users: Interaction between users and the predictive tools needs to be intuitive and helpful, as otherwise, there will be significant resistance to its adoption.
- Instrument: The predictive tools (see section 2.4 on principles).
- Non-criminogenic need: Needs that improve an offender's quality of life but do not significantly affect their probability of offending.
- Static risk factor: Risk factor that cannot be changed by treatment.
- Dynamic risk factor: Risk factor that can be targeted by treatment.
- Responsivity: Sensitivity to treatment
- Fairness: Balancing predictions with policy goals (Berk et al. 2021).

These concepts enable the Author to re-encode the questions into variables for formal concept analysis (see Appendix 2, Table 3, and Table 4). Based on the emerging correlations, questions can be clustered.

Autor is hoping that this survey will give insight into how to design and operationalise risk, need experts to view the static, protective, and dynamic factors and what are the most critical considerations for service design. Service design indirectly informs us about the role of factors and how they can be used. This is also about the central thesis of this work: using dynamic and protective factors for risk prediction.

Questionnaires were sent out to 54 people, and 21 people answered. Of these, five people answered the questions with explanations. The pilot's goal was to establish that the questionnaire and the methodology are sufficient to provide empirical evidence for demonstrating the central thesis of this study. For providing additional context, a qualitative approach was used in addition to the quantitative one: besides quantitative factor analysis, formal concept analysis was used for breaking questions into semantic constituent atoms. For establishing the context of ordinal questions and for getting insight into the experiences of experts, narrative questions were added.

4.1 Factor Survey

There is a significant overlap between offenders' risk and protective factors, as defined in different risk assessment instruments and academic controversies (Heffernan, 2020, p. 21), where a protective factor is the opposite of a risk factor. This survey helps to establish how important these factors are in improving desistance outcomes in the considered opinion of professionals in the criminal justice field. Such a survey gives context and additional evidence to the central research question of this thesis.

This survey draws upon Risk-Need-Responsivity instruments like LSR-C (Serin & Hanby, 2016), Finnish Risk and Needs Assessment Form (RISK) (Salo et al., 2019) based on England/Wales OASys, Good-Lives-Model questionnaire (Harper et al., 2020) and desistance principles (McNeill & Weaver, 2010). It is interesting to note that many of the static factors in the factor questionnaire cover some of Vertovec's super-diversity factors like age, gender, and location (Vertovec, 2007). In risk analysis tools, other static factors can also appear in risk analysis tools, such as race, citizenship, etc. Still, some were skipped to avoid distracting the interviewees with too many sensitive questions.

In Appendix 2, Table 3, each question was broken down into relevant concepts, and the scores from each question were averaged over respondents. The resulting averages were multiplied with a binary matrix of the concepts resulting in an updated matrix, where each row corresponded to one question and each column to a concept. A binary matrix is a boolean matrix where each row is a question, and each column corresponds to a question. A true value is when a question has a concept. This way, the components of the questions can be broken down, and it can be seen where a significant overlap between questions is that some questions may be removed or merged *post hoc*.

For figuring out the hierarchical structure of the concepts being asked, the factor survey questions were mapped against concepts, and dendrogram projections were used to show connections via the structure given by the concept lattice from formal concept analysis (see Figure 1 using R programming language formal concept analysis package called “fcaR”). For visualising clustering better also, seriation of the concept-to-question matrix was performed using R package seriation. Seriation is ordering rows and columns to minimise the distance between filled cells. Seriation

leads to a form of clustering and simplifies the visualisation of the factor analysis results. It should be noted that the Author does not do classical factor analysis due to the small number of respondents. Thus, a more qualitative approach is more appropriate.

Of the static factors, place of origin was considered one of the most critical factors in recidivism (4.7 points out of a maximum 5.0). Interestingly, former gang membership is not considered that important (4.1 points out of a maximal 5.0). Of protective factors, inner peace was considered especially important even though, at least according to statistics on the good lives model, it is unnecessary for desistance results. Also, creativity and sports were supposed to be particularly important.

Here the positive Likert scale values are blue. Interesting qualitative features are that respondents, on average, weighted static factors associated with offenders with an average of 0.72 points (scale [0,1]), while when related to dynamic factors, protective factors, or treatment, it was weighted with 0.8 points, while the maximum is 1. Non-criminogenic need questions got around 0.74 points when associated with protective factors. From qualitative multiple-choice and open questions and the quantitative analysis, the Author concludes that respondents consider dynamic factors relevant when working with offenders. In terms of predictive quality, they are deemed equivalent by the sampled population. For analysis of a similar survey, but with offenders and analysis of the predictive quality of various static, dynamic, and protective factors, see more in Harper et al. (2020).

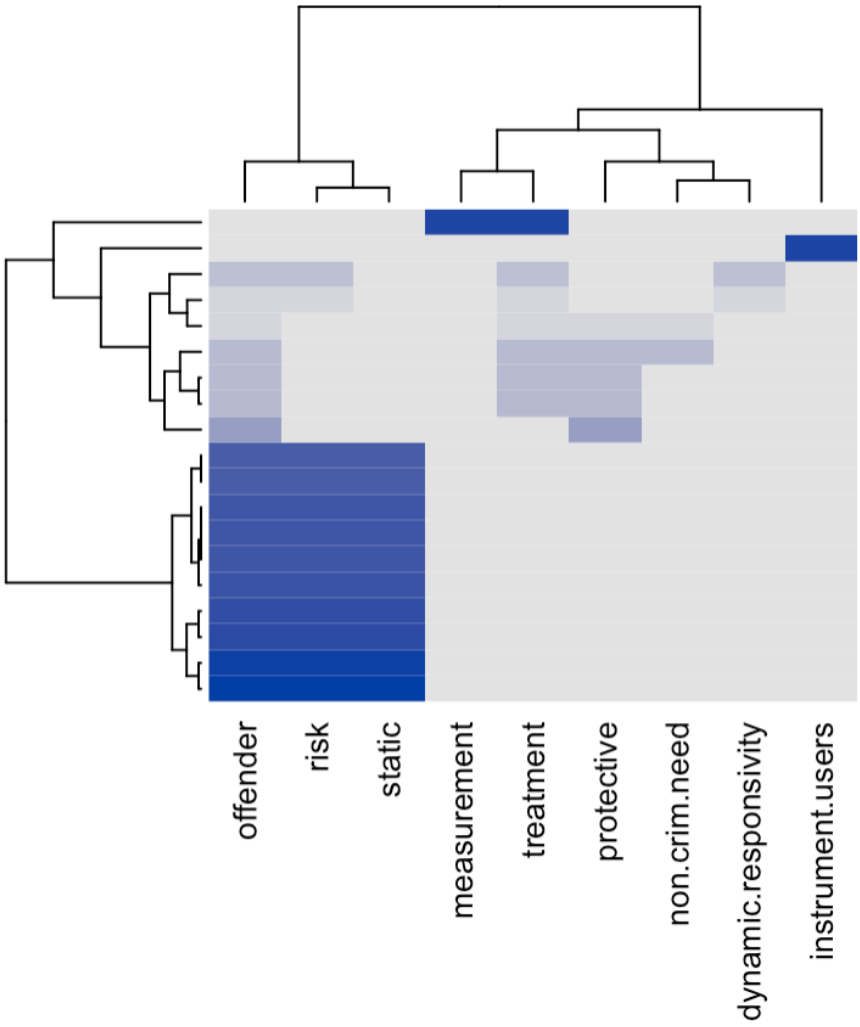

Figure 1. Formal concept analysis of factor questionnaire questions against concepts. Scale from 0 to 1.
Source: Created by the author using the fcaR R package (López-Rodríguez et al., 2021; Ganter & Obiedkov, 2016), based on a factor questionnaire.

The factor survey contains 30 ordinal questions, and the algorithm discovered three factors that account for the majority of the correlations. The root-mean-square of residuals is 0.04, Cronbach's alpha 0.77, total omega 0.92, and hierarchical omega 0.59. These were calculated to express the reliability of this factor analysis. Factor analysis with a Cronbach alpha of 0.77 is usually considered good (Revelle & Zinbarg, 2008).

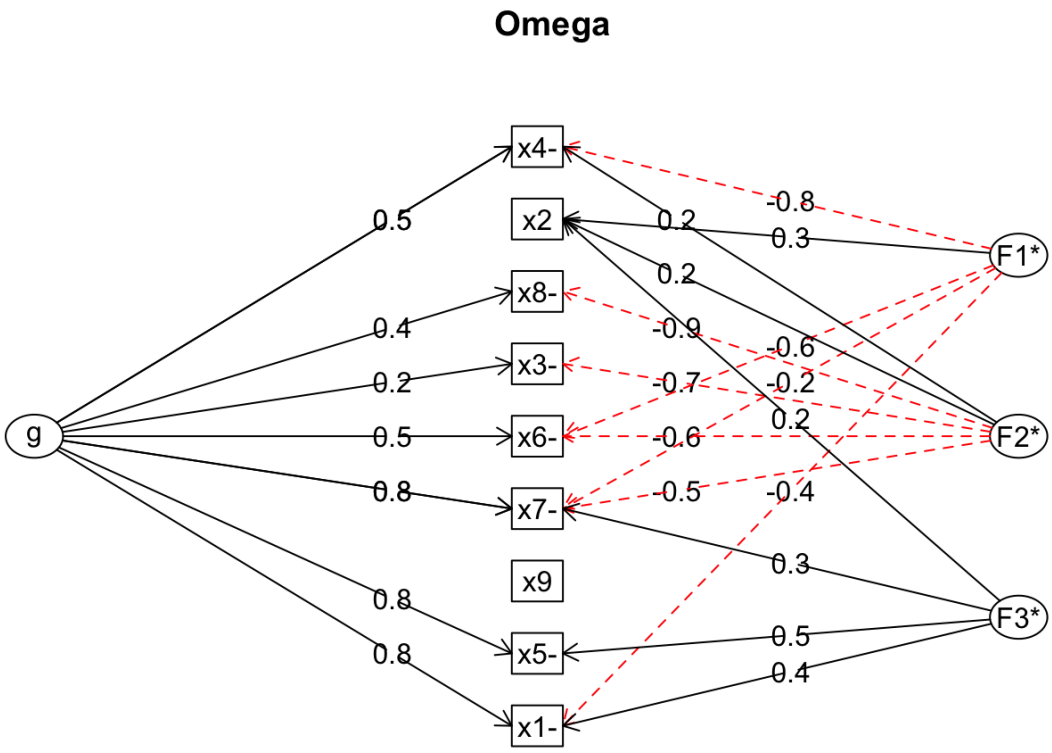

Figure 2. Hierarchical factor analysis of risk survey.
Source: Factor analysis of risk survey concepts using R programming language package "psych."

Following clusters of factors were found F3: (x1,x2,x5,x7) = (risk, measurement, static factor, offender), omega_total= 0.94; F2: (x2,-x3,x4,-x6,-x7-x8) = (measurement, non-criminogenic need, dynamic factor and responsivity, treatment, offender, protective factor), omega_total= 0.85; F1: (-x1,x2,-x4,-x6,-x7,-x8) = (risk, measurement, dynamic factor and responsivity, treatment, offender, protective factor), omega_total=0.28. See Figure 2. For variables, see Table 1.

Table 1. Variable names of concepts for factor questionnaire
Source: The Author defined the variables or concepts contained in the questions.

Variable	Concept	Variable	Concept	Variable	Concept
x1	risk	x5	Static factor	x8	Protective factor
x2	measurement	x6	treatment	x9	Instrument, users
x3	non-criminogenic need	x7	offender	x10	trust
x4	dynamic factor,responsivity				

In addition, item cluster analysis was performed (Revelle & Zinbarg, 2008), see Figure 3. It also revealed interesting patterns. Here alpha is mean-split-half-correlation, and the number under the alpha is called beta, the worst-split-half correlation. These also describe the quality of how factors fit the data. Root-mean-square (RMS) of residuals is 0.2.

Following clusters were found:

C1: (x1,x7)

C3: (C1,x5)

C4: (x4,x6)

C2: (x3,x8)

C6: (-x2,x9),

where dynamic factors and responsivity form a subcluster and are thus under C4, and static factor is under a separate subcluster C3.

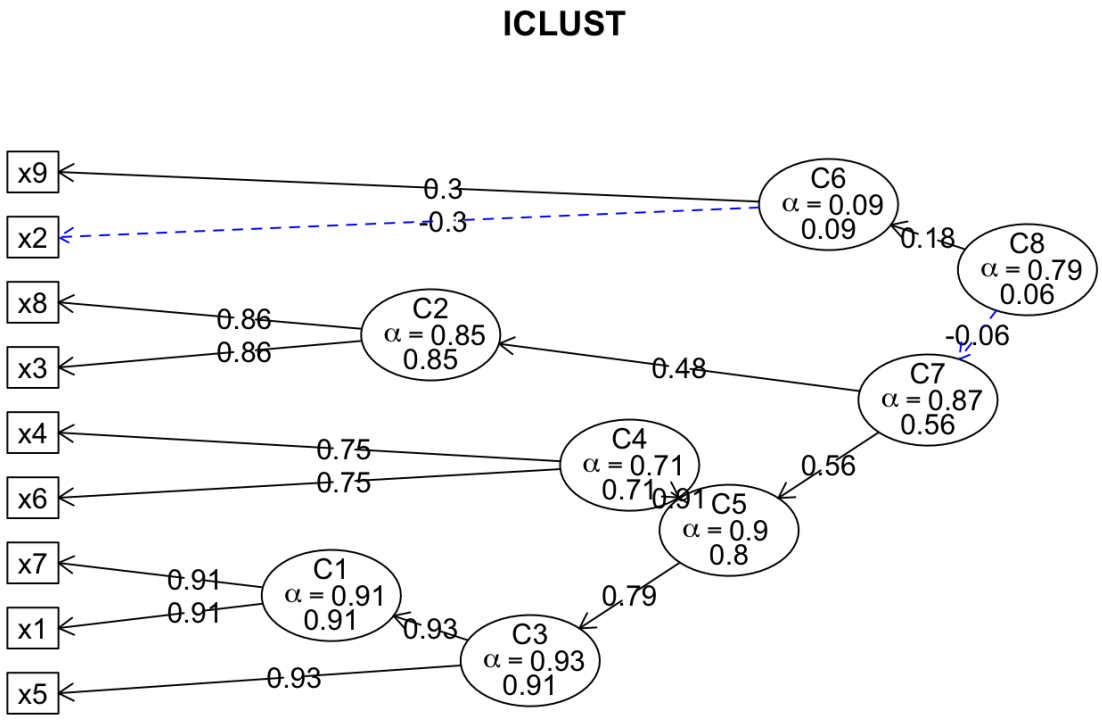

Figure 3. Item cluster analysis of factor questionnaire.
Source: Author's calculations for variables in Table 1.

Item clustering is considered less precise at accounting for the quality of factor analysis than McDonald's omega. Thus, factor analysis can be viewed as a better match, significantly as the RMS score was improved.

Qualitative questions were asked about various risk factors:

- Q18, (Appendix 1): Regarding educational history following aspects were mentioned: quitting, stopping education, shifting subjects frequently, age when left school, possible behavioural reasons for expulsion, lack of goals in life, lack of professional skills, lack of interest in development, low levels of education like lack of reading and writing skills, primary school education.

- Q20, (Appendix 1): Regarding employment history (from the RNR model), the following were mentioned: gaps in career, sudden changes, failures, resignations, motivation to be employed, long periods of unemployment, lack of a career ladder, or even never having been employed, short employment periods.

- Q26, (Appendix 1): On the protective aspects of leisure (from GLM model): respondents mentioned that leisure helps with personal development, reflection in finding meaning, develops thinking and creativity, and increases the sense of achievement and belonging. Also, it was mentioned that it could give hope or goal and increase social skills. If the leisure time is spent on a hobby or recreational activity, it can keep their brain working and give a goal that helps avoid reoffending. Avoiding a sense of isolation was also mentioned.

- Q28, (Appendix 1): Regarding criminal history, the following were mentioned: what are the patterns of behaviour or if there are any sudden changes, has there been violence, what was the age of first offence or conviction, the severity of the crimes, socio-economic factors, and neighbourhood.

- Q30, (Appendix 1): On age, respondents said that it depends on the nature of the offence, as sexual or domestic violence is less affected by age than, for example, gang-related offences. With age, the nature of offences also changes for habitual criminals, but in general, recidivism often declines.

- Q32, (Appendix 1): Regarding gender, respondents said that female offenders are more affected by the toxic three: domestic violence, mental health, and substance misuse. Also,

women tend to commit criminal acts more often from an urgent need, while men do these often more deliberately. Some respondents were unsure if gender is affecting recidivism.

4.2 Service Design Survey

Suppose predictive tools complement professional judgement of risk, criminogenic needs, and responsivity. In that case, it is essential to think firstly: how is risk defined, how is it presented (Crowson et al., 2007), how are protective factors and criminogenic needs defined and presented, and how is responsivity defined and presented. The last one is becoming more and more significant, as it shows whether there is a treatment effect or not and may be one of the hardest to measure. It is also related to human-computer interaction design and how such tools can complement the professional judgement of officials by being informative and how the judgement of professionals vs outcomes can be analysed for later policy adjustments.
Secondly, to think about how this survey helps outline policy and organisational issues that could negate any benefits of such tools.
Thirdly, to think about how to integrate such a tool into scientific experiment designs and thus facilitate continuous evidence-based improvements to criminal justice.
Fourthly, think how to implement such risk/need/responsivity assessment instruments so that the public would trust these and that they would be considered legitimate.

Questionnaire questions and the used terms are shown in Appendix 2 Table 4, where each question was broken down into relevant concepts, and the scores from each question were averaged over respondents. The resulting averages were multiplied with a binary matrix of the concepts resulting in an updated matrix. Questions were mostly created based on author's

Service design survey questions mapped against concepts and dendrogram projections show connections via formal concept analysis. In Figure 4, the positive Likert scale values are blue (the Author has mapped [-2.5,2.5] scores to [0,1] for formal concept analysis). Analysis shows that users' trust and feedback to offenders is essential (around 2.0 from [-2.5,2.5]). Also, the explainability and details of the predictions are considered important, especially when combined with treatment (about 1.9 from [-2.5,2.5]).

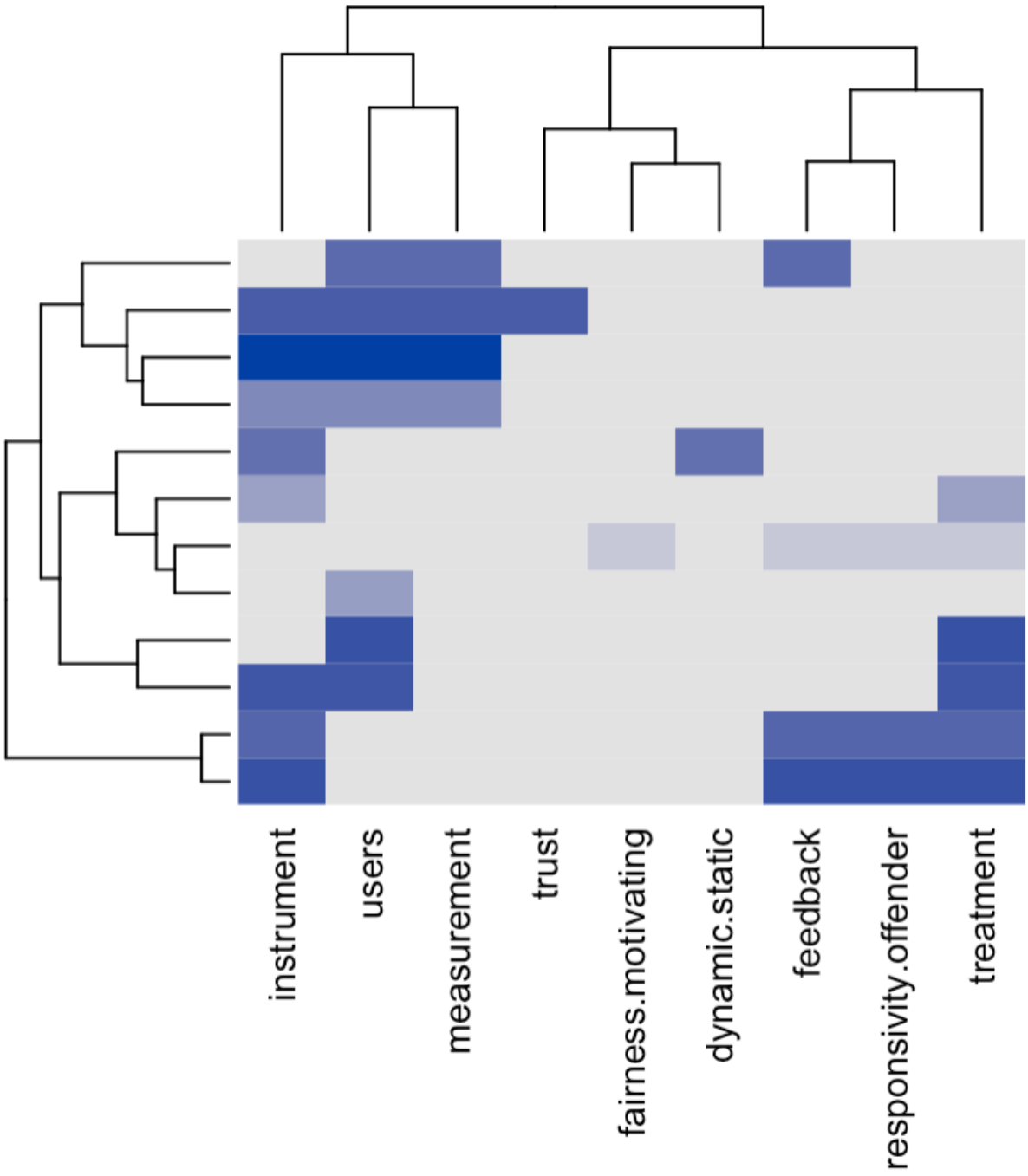

Figure 4. Formal concept analysis of service design questionnaire questions against concepts
Source: Created by the Author using the fcaR R package based on a factor questionnaire.

Just as for the factor survey above, McDonald's omega calculation was performed along with item cluster analysis to discover the optimal factors to match the survey results.

The service design survey has 24 ordinal questions, and the McDonald's omega calculation algorithm discovered two factors that account for most of the correlations. The root means square of residuals is 0.05, Cronbach's alpha 0.58, total omega 0.89, and hierarchical omega 0.38 (Revelle

& Zinbarg, 2008). The score is lower because the algorithm was trying to fit three factors, but the third one was not really needed, and thus this affected the total score. Failing to fit the third factor, it thus showed that the optimal match is using two factors.

This survey has the following clusters: F1: (x2,x4,x5,x6,x7,-x9) = (feedback, treatment, motivating, fairness, responsivity, offender, feedback, instrument, users), omega_total= 0.94; F2: (x1,-x4,x7,x8,x9)=(measurement, treatment, instrument, trust, users), omega_total= 0.77. For concepts behind the variables, see Table 2.

Table 2. Variable names of concepts for service design questionnaire
Source: The Author defined the variables and concepts contained in the questions.

Variable	Concept	Variable	Concept	Variable	Concept
x1	measurement	x4	treatment	x7	instrument
x2	feedback	x5	motivating, fairness	x8	trust
x3	dynamic and static factor	x6	responsivity, offender, feedback	x9	users

In addition, an item cluster analysis was performed, and two large clusters can be seen. Root-mean-square (RMS) of residuals is 0.09.

Using cluster analysis for the service design survey following clusters were found:
C9: (x1,-x3,x7,x8,x9)=(measurement, dynamic and static factor, instrument, trust, users)
C8: (x2,x4,x5,x6) = (feedback, treatment, motivating, fairness, responsivity, offender, feedback)
Here offender is clustered with feedback, treatment, motivation, fairness, and responsivity, while the instrument, along with measurement and professional user concerns, is clustered in the first cluster.

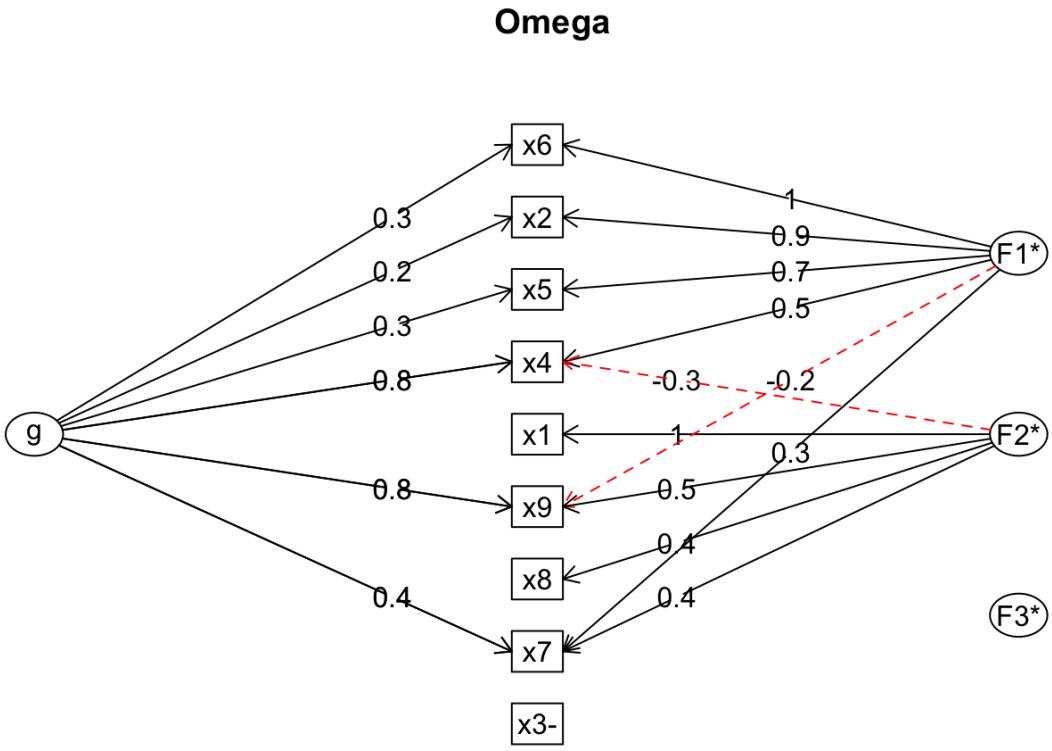

Figure 5. Hierarchical factor analysis of service design survey.
Source: Factor analysis of service design survey concepts using R programming language package "psych" (Revelle, 2017).

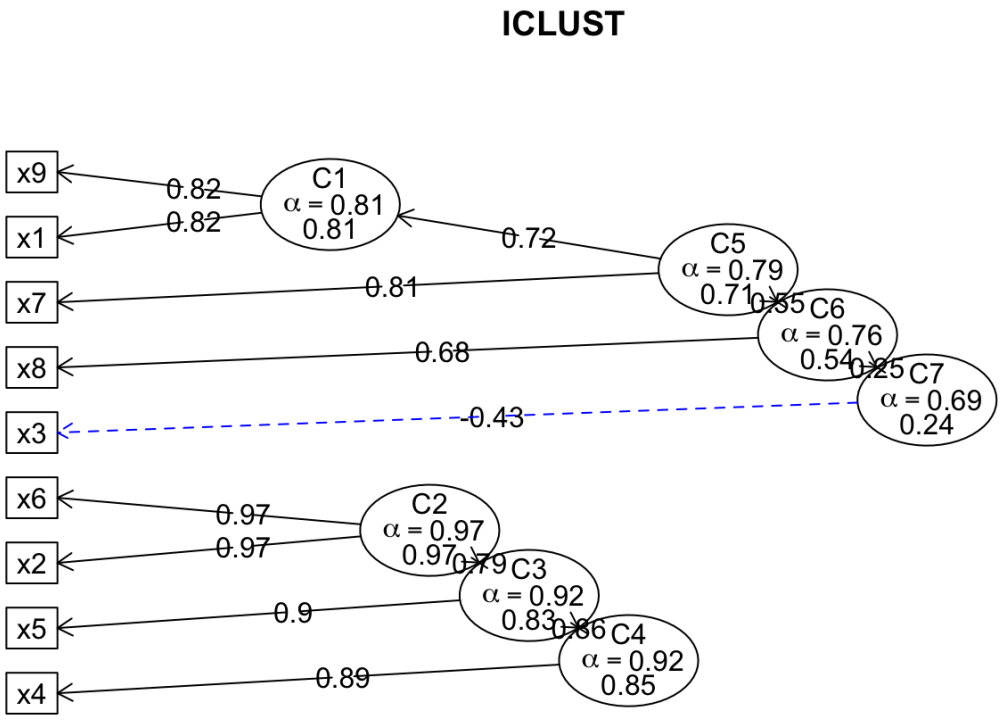

Figure 6. Item cluster analysis of service design questionnaire.
Source: Author's calculations for variables in Table 2.

It is noteworthy that when answering Q9 (Appendix 1), 4:1 of users prefers equalised outcomes (where the difference in predicted outcomes between groups is less than the difference observed in the training data) to equalised odds (where the true-positive rates and false-positive rates are equal across groups) for predictions (Berk et al., 2021), as this is what a specialist would do when assessing risk.

In Q35 (Appendix 1), it was asked what kind of prediction would be helpful for respondents professionally. One respondent said that it would be helpful if a prediction were to measure the success of interventions and treatments related to the offender profile. It was also mentioned that it is futile to predict a person's actions at a specific time, as they might not be in control of the situation or their emotions and behaviour.

An important question (Appendix 1, Q37) was asked what the legal ramifications of risk assessment, such as the presumption of innocence, are. Respondents agreed that this does affect the presumption of innocence, especially if risk assessments are done before sentencing and that the nature of the data used for risk assessment can affect their rights to privacy or even intimacy. It was also mentioned that solving crimes entails the factor of suspicion and that already is often in conflict with the presumption of innocence, and thus, risk assessment might further erode that protection.

Another critical topic is the presented predictions (Appendix 1, Q37). It was emphasised that all their clients are different and that averages are not conclusive. Also, the importance of necessary needs was emphasised: it leads to stress, provoking criminal acts commission.

For implementing any risk assessment tool, there is a great need for trust from all the involved parties, and thus it was asked how to increase it (Appendix 1, Q38). One suggestion was that its importance should be limited and that a stronger focus should be on opportunities for intervention, clients' abilities, environment, and the development of rehabilitative cultures. The need for the self-confidence of officers was also emphasised, along with self-reflection at all times. Also, before a new risk prediction instrument is launched, data on its efficacy must be gathered to show its effectiveness.

On the rights of the offenders to know what their risk score is (Appendix 1, Q39), respondents said that an assessment of risk is an opportunity for meaningful conversations and that the feedback

structure should motivate the offender to correct their course, as otherwise, this information could be iatrogenic.

The engagement of offenders is vital for achieving any success in reducing recidivism and creating meaningful change (Appendix 1, Q40). Respondents mentioned that at least one safe, rehabilitative, engaging, and trustworthy environment must be created. Secondly, it was said that unless an offender understands things from the position of their victims, they might often not be motivated to correct themselves.

Ethical aspects of risk assessment were also asked in Q41, Appendix 1, and only two answers were received. It was emphasised that humans need to be in the centre of decision-making and any activities that come out of that. Also, the A emphasises that ethics and morality are features of humans. While there is a chance to encode these into algorithms in some form, morality and ethics are not formalisable, and an algorithm cannot take responsibility for its actions. It was also emphasised that algorithms must be precisely contextualised to be only used in the narrow sense for which these were designed. The need for transparency and having modest initial goals was emphasised regardless of whether machines or humans made the decisions. It was also mentioned that even a professional psychologist often could not understand the logic by which a person behaves in particular situations. Thus, there is a need to study all factors, especially with high-risk crime (damage probability).

Transparency and interpretability of machine learning predictions is a critical aspect that ensures auditability of predictions and thus a precondition for a successful risk assessment implementation (Appendix 1, Q42). Respondents emphasised that transparency is crucial when it can affect human life decisions, but the most significant risk is that the algorithm predictions can be taken too seriously. In contrast, respondents feel that risk assessments should be interpreted as helpful indicators and assistance, if precision is less than perfect. It was also mentioned that assessments need to be accessible and transparent to offenders.

On the involvement of offenders in risk assessment (Appendix 1, Q43), it was also mentioned that offenders are often vulnerable people with a history of failure. Thus, there is a need to be very careful with it. Otherwise, the assessments can become self-fulfilling prophecies and therefore, whether a risk assessment is discussed or shown to an offender - it should be done individually. It was also mentioned that a specialist must determine if they can participate in the risk assessment

process for some psychological pathologies. Otherwise, this could give an impetus to psychological arousal, where the result of such arousal can be unpredictable. Also, it was emphasised that plans and targets must be composed together by the officer and the offender, as the offender must have a clear path for reaching said targets.

Also, a question on the merit-based approach to offender correction was asked (Appendix 1, Q44). It was said that the sanctions vs rewards approach can be potent but have its limits and, in some cases, can even be contra-productive. Also, offenders themselves need to choose between a hard and soft approach.

The ethics of e-Governance is an important topic and relates to governmentality and albocracy topics, and thus the author wanted opinions on that as well (Appendix 1, Q45). Respondents admitted that ethical boundaries are complex and more severe than in other contexts. It was admitted that criminogenic need prediction would be much more powerful; quote of the answer: "It is not the presumption of innocence as such that could be in danger, it goes much further: the presumption of repetitive behaviour is at the basis of predictions where there might be a need to. We want to focus on changing that behaviour and understand what works.... 'what works algorithms' would be much more powerful than risk assessment algorithms." (Appendix 1, Q45). It was also emphasised that risk assessments are a control tool, not a correction method and that understanding that difference is where the ethics of this topic lies.

One aspect of interest is whether risk assessment tools can also be used to assess correctional program quality (Appendix 1, Q47). Their respondents wanted to change the target: tools should be assessing abilities, opportunities, and capabilities instead of assessing risks and needs. Also, if the programs provide more opportunities in terms of learning, professional improvement, and creativity, the corrective model will be more efficient.

The topic of professional discretion is quite crucial as there is a possibility that, at some point, predictions could be mandated to set a certain baseline of management and treatment of the offender (Appendix 1, Q48). It was emphasised that it depends on the organisational culture and policies.

When asked about the main obstacles in changing the anti-social attitudes of offenders (Appendix 1, Q49), it was immediately countered that the main obstacle is putting them, per definition of an

anti-social environment such as a prison. It was also emphasised that being convicted as a criminal creates a stigma that is difficult to overcome until they adapt appropriately to society, and there, the state can help a lot in helping them adapt to society again, which can reduce recidivism.

Finally, it was asked how to mitigate the effect of anti-social peers (Appendix 1, Q50), and respondents said that there is a need to create trust and that grouping convicts by interests can help a lot: "a grouping of convicts by interests (for example, speciality, education or interest in creativity, etc.). This must be done so that a more authoritative criminal does not impose his interests on another weak person".

In conclusion, the respondents presented many exciting and valuable ideas that give context to their work and what they care about in their daily work. The emphasis from risk assessment to tools in assisting in correction, rehabilitation, reintegration, and desistance is a compulsive challenge. The Author would like to comment that here, as Berk (2019, p. 18) claimed, it is necessary to have different machine learning models for different purposes. The purposes must be clearly and narrowly defined. The respondents also emphasised this.

5. DISCUSSION

This thesis highlights and expounds on which data is used for risk prediction matters, and this chapter, will highlight how this relates to at which stage of, for example, criminal justice processes, a risk prediction are made. By being aware of this and by choosing different data compositions, it can help in defining precise ethical and human rights constraints. The discussion will cover how proper policy recommendations can be created. Based on industry best practices, some policy recommendations are analysed and synthesised consistently with the previously presented theory and empirical results. In its different shades of meaning, the risk is constantly used in many stages and layers of governance and policy. The contribution of this thesis is to highlight that the composition of data on which the risk is predicated matters. In machine learning, several models based on existing data can be built: 1. crime or policy area-specific models; 2. behavioural models based on the events captured by various systems concerning the risk assessee; 3. models based on dynamic risk factors; 4. models based on static risk factors. (Berk, 2019, p. 18).

As per Foucault, the definition of government is "government understood in the larger sense as a means of forming, transforming, and directing the conduct of individuals" (Foucault et al., 2014, p. 23). This definition also encompasses private organisations of sufficient scale. Risk assessment instruments are classified according to Foucault as governance technologies. The concept of governance technologies leads to the sociology of technology and to the governance of contemporary society, which is increasingly concerned with optimising policies for their effectiveness in policy aims and cost. One of the first to clarify the requirements of such technological solutions was Chauncey Starr in his 1969 article (Starr, 1969). He predicted how and why governance of risk using technology became embedded in the governance of modern society. In the case of criminal justice, the main reason for using predictive instruments, as the research has shown, is to improve outcomes which will improve the overall sentencing quality. If predictions can be made and measured, the governance of contemporary society may be affected to optimise policies both for the effectiveness of policy aims, allocate costs efficiently, and avoid counterproductive iatrogenic effects. In practice, measurement, prediction, and optimisation

capabilities can be integrated into modern e-Governance infrastructure. In addition, developments in artificial intelligence and machine learning have highlighted at a mechanistic or algorithmic level: risk, bias, fairness, belief and how this is related to ethics, morality, and human rights. Thus, there are broad implications for governance and policy by laying bare the phenomena that underlie these areas.

When considering the merits of technology and, by implication, how technology re, reasoning cannot just be a utilitarian comparison between technical performance and investment of societal resources. Understanding the relationship between social benefit and justified social cost is also imperative. If the latter requirement is satisfied, predictions of different potential outcomes can be used to decide the optimal policy approaches. E-Governance systems can orchestrate such optimisation processes, implying that predictive instruments must be embeddable within said systems. Here, Melvin Kranzberg's formulation of the so-called Six Kranzberg's laws (truisms) describes how technology interacts with sociocultural change (Kranzberg, 1986) and the patterns in technology's history. Those two laws are most relevant in any policy recommendation concerning risk prediction and related governance tools. The first law states that "technology is neither good nor bad, nor is it neutral" (Kranzberg, 1986), meaning that technology's creation and use involve value judgements. An algorithm or a dataset cannot be ethical, as ethics needs human value judgements. Same with many other technologies. In addition, it is crucial to understand that in different countries and cultures, there are different priorities. It is essential to understand the sensitivities of issues that technologies cause and what they solve. As the extant pain points change in time and space, it also means that this balance changes and thus, the ethics of how technology is used needs to be constantly re-examined. It also means that various disadvantages and benefits of these technologies are discovered as they are being brought into practice. The fourth law states that "although technology might be a prime element in many public issues, nontechnical factors take precedence in technology-policy conditions" (Kranzberg, 1986). Here, a crucial part will be the society's risk perception of the technology. Latter topics will be covered more in-depth in the rest of this discussion.

Without being critical of how data is being used for predicting recidivism and criminogenic need, technological determinism is accepted as a matter of certainty. However, the reality is that there are many different possible paths to reach goals, and part and parcel of engineering are to achieve these goals despite constraints. Without defining those constraints, technology can affect unwanted sociocultural change. These constraints can be captured by the Undesirability Principle, a

paraphrase of Heisenberg's uncertainty principle: "... the product of the costs of two or more conflicting courses of action is a constant. Society, therefore, can obtain one goal to whatever degree of desirability it wishes provided that it is willing to pay the price in loss of desirability in other goals" (Koshland, 1985, p. 4708).

Before policy recommendations can be given, it is necessary to consider several public policy considerations of risk assessment instruments in criminal policy, as is highlighted, for example, by Berk (2019, p. 16-17). The foremost policy consideration is ensuring public safety. Ancillary goals of public safety are reducing carceral population and efficiency of resource allocation. Another important goal of risk assessment instruments is the transparency of prediction: different stakeholders will require different levels of structured detail. Another politically and socially sensitive topic is fairness. How are race and other protected factors related to forecasting results? Which fairness goals are desirable, vis-a-vis equal outcomes and equal odds (Berk et al., 2021). Equal outcome criteria are where all groups are treated according to the same standards. Equal odds criteria balance this to account for social inequality, the bias of the criminal system, and other such standards to meet policy goals that adjust for different goals. For example, to reduce the proportion of a group in the carceral population due to iatrogenic effects of incarceration and its impact on the said population. In addition, it is essential to ensure the practicality of risk assessment instruments, as otherwise, they will not become widely used and has implication for the requirements for the precision and interpretability of prediction results. Previously when discussing static vs dynamic factors, the goal of forecasting risk vs forecasting need is also essential as the goals of forecasting risk and needs (treatment) are quite different and, according to Berk (2019, p. 17), can be mutually exclusive; thus, separate modelling efforts are recommended.

The service design survey showed that some people prefer a simple risk score, while others want the ability to understand what is behind the risk score. Some people prefer an absolute risk score, but the problem is that as the underlying models change, absolute scores will drift in time, and thus the real-life meaning of a quantitative score may shift. Therefore, relative risk ratios are recommended as a minimum, a widespread practice in medicine (Freudenburg, 1988). In behavioural economics, it has been discovered that people prefer certainty to risk and risk to ambiguity. It has also been shown that the amount of information on the distribution of probabilities for different outcomes leads to different decision-maker preferences: they prefer such an outcome to the situation where probabilities are uncertain (Kirchler & Hoelzl, 2017, p. 29).

Thus, if risk assessment seems precise, but the error is not correctly shown, it might be preferred to expert judgment, but these might be dismissed if there is not enough detailed explanatory information. In medicine, regarding data-driven ML prediction, it has been pointed out in the famous adage that correlation does not imply causation. Scientific and empirical research are needed to add more precision to predictions and vet the data points used for prediction (Prosperi et al., 2020).

In the criminal justice systems of the United Kingdom and the United States, it is common to run various experiments, for example, to see which interventions lead to which outcomes, and part and parcel of experimental criminology (Debidin, 2009). Having a breadth of different models based on distinct factors, diverse data sources, different data processing methodologies, and a constantly updating stream of new data provide an opportunity to run these experiments in situ and a real-time understanding of how policies work. When examining existing risk assessment models, the Author discovered that these are updated significantly in decade long cycles, but society changes fast and scientific understanding. Also, a risk prediction system, whether actuarial or based on machine learning or clinical assessments, should be continually refined and optimised: both in terms of the software systems and processes and end-user training. In other words, there needs to be: a global and local approach; algorithmic tools to inform and advise based on the global context; room for local approach expert discretion; human touch to customise approaches to avoid reducing humans into categories. Both parts are integral, as humans provide the necessary validation and measurement step for the algorithm's predictions locally.

The position of the author is that if the feedback of experts is also used to adjust the models, and if the training data is curated and constantly monitored by running updated previous versions against previous decisions, then the model can be aligned toward policy goals and thus, the machine learning model will become a better instrument for applying the goals and principles of a policy. On the other hand, a distance is needed between policy and the models. These models also need to be in concordance with evidence, with the latest results of experimental criminology. Considering that best practices constantly evolve, a lot will be learned.

At the same time, the system is being implemented, and it is best to specify principles, controls, and processes and iteratively enhance the former. In terms of organisation, cooperation between data scientists, process engineering, psychologists, psychiatrists, governance specialists, jurists, criminologists, correction officers, prosecutors, justices, politicians, and many others is needed.

For making this cooperation work, a process should be in place from the start, and it needs to be transparent (Berk, 2019, p. 116), and the predictions and use of said predictions, needs to be explainable. For example, in Illinois, the U.S., the use of risk assessment by the courts is opt-in for the defendants (Casey et al., 2011, p. 13). If chosen, a judge can be advised by the recidivism risk assessment in terms of not only risk but also appropriate care, and that can mean that in the case of mandatory minimums (a frequent case in the U.S.), the actual circumstances and resources assigned to the defendant during serving of their sentence, can be favourably adjusted, and reasoned. It has been shown that some judges do like to have additional information during sentencing and use the risk assessment to put their decision in a broader context. The issue with the GLM model is that it is not readily applicable to all people. Many criminals have multiple psychiatric disorders, and some derive enjoyment from sadistic activities. It is naive and hubris to think that somebody can fully reprogram people, especially if they do not wish. In corrections, the Good-Lives-Model GLM (Ward, 2004) appeals to better instincts, while risk-need-responsivity RNR is more utilitarian and does not focus on protective factors (Andrews et al., 1990). The problem with these models is that their effect is not that easy to measure, leading to a conflict between RNR (Andrews et al., 2011) and GLM (Ward et al., 2012) proponents. Even though there is criticism from the proponents of both models of correction towards the other, proponents of both agonise that the success of those models depends on the training and close adherence to the principle of said models (Duwe & Kim, 2018).

The danger in corrective models based on the different sets of risk and protective factors is that they can oversimplify things and degenerate into a solipsistic form of solutionism and scientism. On the other hand, if there are no comprehensive general evidence-based care standards, local distortions can lead to abuses, discrimination, and corrupt practices (Burgess, 1928). Nevertheless, various treatments, especially cognitive behavioural therapy, lessen recidivism. While for example, art therapy and similar therapies are not proven to lessen recidivism but are considered to contribute to general well-being (Ward et al., 2012). There are no magic solutions, but warehousing of undesirables in the long run only perpetuates problems. Short-term immediate criminal justice punishment goals hold up the criminal justice system's legitimacy and provide security and closure. However, retributionist goals do not optimise to reduce recidivism in the long term (Starr, 2014).

Many ethical concerns regarding risk assessment based on models or algorithms (ML/AI) are raised. One of the main arguments is that the only accurate risk assessment is expert-based and

clinical assessment in the case of criminal justice. The argument goes that it is unethical to make grave decisions about a person's future without considering all the individual aspects and classifying a person into specific groups. In reality, individual assessments are made in many stages of the criminal justice process. As the surveys presented in this thesis show, interviewees wish to have the discretion to make decisions based on individual characteristics. On the other hand, statistically, risk assessment tools have been shown to improve the aggregate situation significantly (Imrey & Dawid, 2015) and, as was shown previously already during the 1920s (Burgess, 1929): common standards and tools that provide predictions, based on statistics of the aggregate sample, help in addressing issues in only human-based processes (Goodman-Delahunty & Sporer, 2010). In practice, when predicting recidivism, state-of-the-art ML models are starting to, on average, outperform humans (Lin et al., 2020), and it is imperative to understand the reasons for that. In addition, it is also claimed in this thesis that if predictions are explainable (Hickey et al., 2021) and have both breadth (variety of specialised models) and depth (what factors are most significant in scores), a specialist can find aspects to turn their attention and have a context to what are the statistical properties of various risk factors. It is something to be further researched in practice. Another underused possibility is that several risk assessment models based on fairness criteria can be provided.

Unfortunately, fairness in recidivism risk predictions in particular and risk assessments in general (Berk et al., 2021) reflects the conflict between policy aims, existing social situations, diverse circumstances of people, and the fact that risk prediction systems are based on the data that has been measured not based on what is. In criminal justice, risk assessment models reflect the detected crime and the observed outcome, which can also reflect the social situation of diverse groups of people. Thus, on the one hand, it is unwanted to have a self-reinforcing feedback loop in a criminal justice process that hardens and enforces the current circumstances of certain groups. Still, on the other hand, it is a wish to treat people equally. Estimating heterogeneous externalities can become necessary when making predictions for policies based on models (Arduini et al., 2020). This is an ethical dilemma where some authors talk about the algorithm or ML or AI ethics, but that is fundamentally wrong: the responsibility for decisions is on people, from the creation of the software systems to the composition of data, to the policy with governance processes and the execution of said processes. Ethical, social, and moral problems are reflected in the risk assessment results, and it is the task of policy, process, and instrument creators to find reasonable compromise and to provide a balance between standardised and individual approaches to enhance the process instead of abdicating the discretion of officials to algorithms (Skeem et al., 2020).

In the case of criminal justice, it has been shown by various authors (Hannah-Moffat, 2005) that dynamic risk factor- and static risk factor-based models perform more or less the same and thus, in future work, it would be worthwhile to see if models 1, 2, 4 can provide nuance to the risk predictions of model 3 (see page 44). The Author can say that the diversity of different models offers a breadth of viewpoints, while the explainability of the predictions in terms of the significant factors that contribute most to the risk assessment (Rebane et al., 2020).

CONCLUSION

In the thesis, the Author has demonstrated that in the criminal justice settings, it is preferable to use dynamic risk factors for algorithmic predictions. While professionals can draw upon various data sources for clinical assessments, static factors like criminal history, educational history, race, and gender are not a prerequisite for a perfect prediction. A survey was done among criminal justice professionals that demonstrated how vital dynamic and protective factors are, factors that either are changeable by treatment or that protect against offending. These factors are separable also, in their opinion, from static factors. It was remarkable how important desistance principles and positive psychology have become among criminal justice professionals and how dynamic and protective risk factors have become as important as static factors for guiding their work. This survey validated that criminal justice professionals consider treatment, dynamic and protective risk factors, and rehabilitative and reintegrative goals equivalently or even more important than plain risk prediction via static factors.

By building algorithmic prediction models around dynamic and protective factors and using static factor-based predictions as a competing model for validating and tuning dynamic factor-based models, it is possible to arrive at a results-oriented e-Governance of criminal justice processes. It is possible because then measurement and acting upon predictions are done against factors that can be affected.

Another aspect mentioned in the thesis is that while the false-negative rate affects the risk to the security of various people, the false-positive rate disproportionately affects the vulnerable section of offenders. Thus, that will delegitimise the criminal justice system for certain parts of society. It is impossible to eliminate false-negative or false-positive risk predictions, but it is possible to base those predictions on data about what can be done to rehabilitate and reintegrate offenders. This way, criminal justice is not about retribution, deterrence, and risk management but rather about preventing crime, rehabilitating, and reintegrating. Criminal justice can also function as a gateway to social justice.

Current work can further be extended to repeat this survey for a larger sample size and combine the survey with in-depth cognitive interviews. In addition, the conclusions of this work can be formulated into clear policy goals and algorithmic risk assessment design requirements.

KOKKUVÕTE

Riskide masinõppe põhine ennustamine ja valitsemine kriminaalpoliitika näitel

Iren Irbe

Käesolevas magistritöös autor näitas, et kriminaalõiguse valdkonnas on mõistlik kasutada dünaamilisi riskitegureid algoritmiliste ennustuste tegemisel. Kuigi kliinilist hindamist tegevad professionaalid saavad kasutada mitmekülgseid andmeallikaid ning nende seas ka staatilisi riskitegureid nagu kriminaalregister, haridustee, rass, sugu jpm, siis need ei ole hea algoritmilise ennustuse eeldustingimused. Autor teostas küsitlust õigussüsteemi professionaalide seas ning see näitas kuivõrd oluliseks tegelikult peetakse nii dünaamilisi kui ka kaitsvaid tegureid, tegureid mis on siis kas kriminaalhoolduses muudetavad või retsidiivsust takistavad. On märkimisväärne kuivõrd oluliseks on muutunud kuritegevusest loobumise *(desistance)* teooria printsiibid ning positiivse psühholoogia roll kriminaalhoolduses ning kuidas dünaamilised riskifaktorid ja kaitsvad faktorid on muutunud vähemalt sama, kui mitte olulisemakski, kui staatilised faktorid nende töös. Lisaks see küsitlus näitas, et tõepoolest kriminaalõiguse ametnikud peavad ravi, dünaamilisi ja kaitsvaid riskifaktoreid, rehabiliteerivaid ja reintegreerivaid eesmärke olulisemaks kui riskide haldamist staatiliste riskifaktorite kaudu.

Ehitades algoritmilise ennustamise mudeleid eelkõige dünaamiliste ja kaitsvate tegurite abil ning samas staatilisi tegurite põhiseid mudeleid ennustuste valideerimiseks ja esimeste kalibreerimiseks kasutades, on võimalik jõuda palju tulemustele orienteeritumate e-Riigi kriminaalõigusprotsessideni. Seda seetõttu, et mõõdame ja tegutseme lähtuvalt teguritest mida saab mõjutada.

Käesolevas töös mainitakse ka aspekti, et kui valenegatiivsete tulemuste sagedus riskide ennustamises mõjutab inimeste turvalisust, siis valepositiivide sagedus disproprortsionaalselt mõjutab haavatavamat osa kriminaalõigussüsteemi sattunuist ning seega delegitimiseerib õigussüsteemi ühiskonna teatud osades. Ei ole võimalik elimineerida nii valenegatiivseid kui ka valepositiivseid riskide ennustusi, kuid on võimalik need baseerida andmetele, mis annavad

sisendit sellele, et kuidas saab rehabiliteerida ja reintegreerida kriminaalkaristusi saanuid. Kui sellest lähtuda, siis kriminaalõigus ei ole niivõrd keskendunud riskide haldamisele, heidutusele, karistamisele, vaid põhirõhk liigub kuritegevuse ennetamisele ning rehabiliteerimise ja reintegreerimisele. Sedasi saab kriminaalõigussüsteem kaasa aidata sotsiaalse õigluse edendamisele.

Käesolevat tööd saab täiendavalt laiendada korrates küsitlust suurema valimiga ning kombineerida küsitlused kognitiivsete intervjuudega. Lisaks saab töö järelmitest formuleerida selged poliitika eesmärgid ning algoritmilise riski hindamissüsteemidele täiendavad disaninõuded.

LIST OF REFERENCES

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for Effective Rehabilitation: Rediscovering Psychology. *Criminal Justice and Behavior*, 17(1), 19–52.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The Risk-Need-Responsivity (RNR) Model: Does Adding the Good Lives Model Contribute to Effective Crime Prevention? *Criminal Justice and Behavior*, 38(7), 735–755.

Appelbaum, P. S. (1988). The new preventive detention: Psychiatry's problematic responsibility for the control of violence. *The American Journal of Psychiatry*, 145(7), 779–785.

Arduini, T., Patacchini, E., & Rainone, E. (2020). Treatment Effects With Heterogeneous Externalities. *Journal of Business & Economic Statistics*, 38(4), 826–838.

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, 25, 421–458.

Berk, R. (2019). *Machine Learning Risk Assessments in Criminal Justice Settings* (1st ed. 2019). Springer International Publishing: Imprint: Springer.

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

Bernburg, J. G., & Krohn, M. D. (2003). Labeling, Life Chances, and Adult Crime: The Direct and Indirect Effects of Official Intervention in Adolescence on Crime in Early Adulthood. *Criminology*, 41(4), 1287–1318.

Bohman, J., Flynn, J., & Celikates, R. (2021). Critical theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/critical-theory/

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, 15(5), 662–679.

Bruce, A. A., Harno, A. J., Landesco, J., Burgess, E. W. (1928). *The Workings of the Indeterminate-Sentence Law and the Parole System in Illinois: A Report to the Honorable Hinton G. Clabaugh, Chairman, Parole Board of Illinois*. Portland, OR: Willan Publishing.

Burgess, E. W. (1928) Factors determining success or failure on parole. In A. A. Bruce, A. J. Harno, J. Landesco, E. W. Burgess (eds.) (1928). *The Workings of the Indeterminate-Sentence Law and the Parole System in Illinois: A Report to the*

Honorable Hinton G. Clabaugh, Chairman, Parole Board of Illinois. Portland, OR: Willan Publishing.

Burgess, E. W. (1929). Is prediction feasible in social work? An inquiry based upon a sociological study of parole records. *Social Forces*, 7(4), 533-545.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284–299.

Carter, M. & Sankovitz, R.J. (2014). *Dosage Probation: Rethinking the Structure of Probation Sentences*. Washington, DC: National Institute of Corrections.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011). *Using offender risk and needs assessment information at sentencing: Guidance for courts from a national working group*. National Center for State Courts.

Crowson, C. S., Therneau, T. M., Matteson, E. L., & Gabriel, S. E. (2007). Primer: Demystifying risk—understanding and communicating medical risks. *Nature Clinical Practice Rheumatology*, 3(3), 181–187.

Cullen, F. T. (2006). It's Time to Reaffirm Rehabilitation Public Preference for Rehabilitation: Reaction Essay. *Criminology and Public Policy*, 5(4), 665–672.

Debidin, M. (2009). *A compendium of research and analysis on the Offender Assessment System (OASys) 2006-2009*. Ministry of Justice.

Dershowitz, A. M. (1970). Preventive Detention and the Prediction of Dangerousness--The Law of Dangerousness: Some Fictions about Predictions: Chapter 2 Social Research and the Law: The Administration of Justice. *Journal of Legal Education, 23*(1), 24–47.

Dershowitz, A. (1974). The Origins of Preventive Confinement in Anglo-American Law - Part II: The American Experience. *University of Cincinnati Law Review*, 43(4), 781–846.

Duwe, G., & Kim, K. (2018). The Neglected "R" in the Risk-Needs-Responsivity Model: A New Approach for Assessing Responsivity to Correctional Interventions. *Justice Evaluation Journal*, 1(2), 130–150.

Edelstein, L. (1952). The Relation of Ancient Philosophy to Medicine. *Bulletin of the History of Medicine*, *26*(4), 299–316.

Epperson, M., & Pettus-Davis, C. (2017). *Smart decarceration: achieving criminal justice transformation in the 21st century*. Oxford University Press

Etienne, M. (2009). Legal and Practical Implications of Evidence-Based Sentencing by Judges Symposium: Evidence-Based Sentencing: The New Frontier in Sentencing Policy and Practice: Perspectives of Prosecutors, Judges, and Defense Attorneys. *Chapman Journal of Criminal Justice*, 1, 43–60.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299.

Focquaert, F., Shaw, E., & Waller, B. N. (Eds.). (2021). *The Routledge handbook of the philosophy and science of punishment*. Routledge.

Freudenburg, W. R. (1988). Perceived risk, real risk: Social science and the art of probabilistic risk assessment. *Science*, 242(4875), 44–49.

Foucault, M., Brion, F., Harcourt, B. E., & Sawyer, S. W. (2014). *Wrong-doing, truth-telling: The function of avowal in justice.* University of Chicago Press ; Presses Universitaires de Louvain.

Ganter, B., & Obiedkov, S. (2016). *Conceptual Exploration (1st ed. 2016)*. Springer.

Garrett, B. L. (2019). Federal Criminal Risk Assessment. *Cardozo Law Review*, 41(1), 121–150.

Ghiraldelli, P. (2006). Marxism and critical theory. In J. R. Shook & J. Margolis (Eds.), *A companion to pragmatism.* Blackwell.

Green, J. M. (2006). Pluralism and deliberative democracy: A pragmatist approach. In J. R. Shook & J. Margolis (Eds.), *A companion to pragmatism.* Blackwell.

Goodman-Delahunty, J., & Sporer, S. L. (2010). Unconscious influences in sentencing decisions: a research review of psychological sources of disparity. *Australian Journal of Forensic Sciences*, 42(1), 19–36.

Hamilton, M. (2020). Judicial gatekeeping on scientific validity with risk assessment tools. *Behavioral Sciences & the Law*, 38(3), 226–245.

Hannah-Moffat, K. (2005). Criminogenic needs and the transformative risk subject: Hybridizations of risk/need in penality. *Punishment and Society*, 7(1), 29–51.

Harper, C. A., Lievesley, R., Blagden, N., Akerman, G., Winder, B., & Baumgartner, E. (2020). Development and validation of the Good Lives Questionnaire. *Psychology, Crime & Law*, 0(0), 1–26.

Harris, P. (2006). What Community Supervision Officers Need to Know about Actuarial Risk Assessment and Clinical Judgment. *Federal Probation*, 70(2), 8–14.

Heffernan, R. (2020). *Beyond dynamic risk factors: Towards a comprehensive explanation of offending*. Victoria University of Wellington.

Hester, R. (2020). Risk assessment savvy: The imperative of appreciating accuracy and outcome. *Behavioral Sciences & the Law*, 38(3), 246–258.

Hickey, J. M., Di Stefano, P. G., & Vasileiou, V. (2021). Fairness by Explicability and Adversarial SHAP Learning. In F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 174–190). Springer

International Publishing.

Hill, B., Lagerlund, H., & Psillos, S. (Eds.). (2021). *Reconsidering causal powers: Historical and conceptual perspectives*. Oxford University Press.

Hodgson, J., & Soubise, L. (2016). Understanding the Sentencing Process in France. *Crime and Justice*, 45(1), 221–265.

Honneth, A., Butler, J., Geuss, R., Lear, J., Jay, M. (2008). *Reification: A new look at an old idea.* Oxford University Press.

Horkheimer, M., Adorno, T. W., & Schmid Noerr, G. (2002). *Dialectic of enlightenment: Philosophical fragments*. Stanford University Press.

Imrey, P. B., & Dawid, A. P. (2015). A Commentary on Statistical Assessment of Violence Recidivism Risk. *Statistics and Public Policy*, 2(1), 1–18.

Irwing, F. P. (Ed.). (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (First Edition). Wiley Blackwell.

Kasapoglu, T., & Masso, A. (2021). Attaining Security Through Algorithms: Perspectives of Refugees and Data Experts. In: Julie B. Wiest (Ed.). *Theorizing Criminality and Policing in the Digital Media Age*. (pp. 47–65). Emerald Publishing. (Studies in Media and Communications).

Kirchler, E., & Hoelzl, E. (2017). *Economic psychology: An introduction*. Cambridge University Press.

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10, 113–174.

Koshland, D. E. (1985). The Undesirability Principle. *Science*, 229(4708), 9–9.

Kranzberg, M. (1986). Technology and History: "Kranzberg's Laws." *Technology and Culture*, 27(3), 544–560.

Kroner, D. G., Morrison, M. M., & M. Lowder, E. (2020). A Principled Approach to the Construction of Risk Assessment Categories: The Council of State Governments Justice Center Five-Level System. *International Journal of Offender Therapy and Comparative Criminology*, 64(10–11), 1074–1090.

Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2), 70–91.

Lin, Z. "Jerry", Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, 6(7).

López-Rodríguez, D., Cordero, P., Enciso, M., & Mora, Á. (2021). Clustering and Identification of Core Implications. In A. Braud, A. Buzmakov, T. Hanika, & F. Le Ber (Eds.), *Formal Concept Analysis (pp. 138–154)*. Springer.

Margolis, J. (2006). Introduction: Pragmatism, retrospective, and prospective. In J. R. Shook & J. Margolis (Eds.), *A companion to pragmatism*. Blackwell.

Martinson, R. (1974). What works? - Questions and answers about prison reform. *The Public Interest*, 35, 22–54.

Martinson, R. (1979). New Findings, New Views: A Note of Caution Regarding Sentencing Reform. *Hofstra Law Review*, *7*(2), 243–258.

Masso, A. (2021). Automating Super-Diversity: The Mechanisms of Machine Prejudice. *Article Submitted for Publication Consideration at Big Data and Society*.

McNeill, F., & Weaver, B. (2010). *Changing lives? Desistance research and offender management*. Scottish Centre for Crime and Justice Research, University of Glasgow.

Minsky, M. (1986). *The society of mind*. Simon and Schuster.

Monahan, J. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. Oxford University Press.

Oleson, J. C. (2011). Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing. *SMU Law Review*, 64(4), 1329–1404.

O'Reilly, M. F. (2018). *Uses and consequences of a criminal conviction*. Springer.

Ostrom, E., Dietz, T., Dolšak, N., Stern, P. C., Stonich, S., & Weber, E. U. (Eds.). (2002). *The drama of the commons (pp. xi, 521)*. National Academy Press.

Plato (1997). *Complete works*. Hackett Pub.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7), 369–375.

Rebane, J., Samsten, I., & Papapetrou, P. (2020). Exploiting complex medical data with interpretable deep learning for adverse drug event prediction. *Artificial Intelligence in Medicine*, 109.

Reitz, K. R. (2020). The compelling case for low-violence-risk preclusion in American prison policy. *Behavioral Sciences & the Law*, 38(3), 207–217.

Revelle, W. (2017). *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston. R package version 1.8.6

Revelle, W., & Zinbarg, R. E. (2008). Coefficients Alpha, Beta, Omega, and the glb: Comments

on Sijtsma. *Psychometrika*, 74(1), 145-154.

Rorty, R. (1978). Philosophy as a Kind of Writing: An Essay on Derrida. *New Literary History*, 10(1), 141.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Salo, B., Laaksonen, T., & Santtila, P. (2019). Predictive Power of Dynamic (vs. Static) Risk Factors in the Finnish Risk and Needs Assessment Form. *Criminal Justice and Behavior*, 46(7), 939–960.

Sampson, R. J., & Laub, J. H. (1992). Crime and Deviance in the Life Course. *Annual Review of Sociology*, 18(1), 63–84.

Serin, R. C., & Hanby, L. J. (2016). Client-Based Assessment of Need and Change. *In The Wiley Handbook on the Theories, Assessment and Treatment of Sexual Offending (pp. 1575–1592)*. American Cancer Society.

Simon, J. (2007). *Governing through crime: How the war on crime transformed American democracy and created a culture of fear.* Oxford University Press.

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior*, 44(1), 51.

Stahl, T. "Georg [György] Lukács", In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/lukacs/

Stanford, K. "Underdetermination of Scientific Theory", In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/scientific-underdetermination/

Starr, C. (1969). Social Benefit versus Technological Risk: What is our society willing to pay for safety? *Science*, 165(3899), 1232–1238.

Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review*, 66(4), 803–872.

Taylor, L., & Meissner, F. (2020). A Crisis of Opportunity: Market-Making, Big Data, and the Consolidation of Migration as Risk. *Antipode*, 52(1), 270–290.

Tibbitts, C. (1932). Reliability of Factors Used in Predicting Success or Failure in Parole. *Journal of Criminal Law and Criminology (1931-1951)*, 22(6), 844–853.

Tonry, M. (2019). Predictions of Dangerousness in Sentencing: Déjà Vu All Over Again. *Crime and Justice,* 48, 439–482.

Travis, J. (2005). *But they all come back: Facing the challenges of prisoner reentry (1st ed.)*. Urban Institute Press.

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies*, *30*(6), 1024–1054.

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

Ward, T., Yates, P. M., & Willis, G. M. (2012). The Good Lives Model and the Risk Need Responsivity Model: A Critical Response to Andrews, Bonta, and Wormith (2011). *Criminal Justice and Behaviour*, 39(1), 94–110.

Welsh, B. C., Yohros, A., & Zane, S. N. (2020). Understanding iatrogenic effects for evidence-based policy: A review of crime and violence prevention programs. *Aggression and Violent Behaviour*, 55, 101511.

West, C. (1989). *The American evasion of philosophy: A genealogy of pragmatism*. Macmillan.

Whitehead, A. N. (1948). *Science and the modern world*. Pelican Mentor Book

Wille, R. (2009). Restructuring lattice theory: an approach based on hierarchies of concepts. In S. Ferré & S. Rudolph (Eds.), *Formal Concept Analysis* (pp. 314–339). Spring

APPENDICES

Appendix 1. Questionnaires

Offender treatment service design questionnaire

Thank you for filling this questionnaire. Your aid will greatly help me with my Master's Thesis and also when designing new and useful services.

1. Which kind of risk measure do you prefer?

Märkige ainult üks ovaal.

- ◯ Absolute risk scale: X% of offenders with a score of X are reconvicted within X years"
- ◯ Relative risk scale: "offender with a score of Y is in the top X% in terms of risk for sexual recidivism"
- ◯ Rate ratio: " offenders with a score of X have Y times the recidivism rate of offenders with a score of Z
- ◯ Odds ratio: "the odds of recidivism for offenders with a score of X is Y times the odds of recidivism for offenders with a score of Z

2. How important is it to describe the precision of an assessment to the user?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Not important	◯	◯	◯	◯	◯	Very important

3. How detailed should scoring be?

Märkige ainult üks ovaal.

- ◯ One risk score
- ◯ One risk score, one criminogenic needs score, one responsivity (treatment effect) score
- ◯ Break-down of score by assessment instrument + professional judgements
- ◯ Break-down of scores, judgements, assessment instrument score components
- ◯ All of the above and that the predictions are weighted to correspond to the current step of the judicial process

4. Please assess the importance of following types of collaboration:

Märkige kõik sobivad.

	Request professional clinical judgement	Request additional assessments	Give feedback on the quantitative assessments	Scheduler for managing work and interacting with offender or victim	Automated treatment recommendation
Strongly disagree	☐	☐	☐	☐	☐
Disagree	☐	☐	☐	☐	☐
Neutral	☐	☐	☐	☐	☐
Agree	☐	☐	☐	☐	☐
Strongly agree	☐	☐	☐	☐	☐

5. How important is it to give feedback to offenders on the risk, need, and responsivity assessments?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly disagree	⬭	⬭	⬭	⬭	⬭	Strongly agree

6. Is it morally more correct to assess risk via dynamic risk factors than via static risk factors?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly disagree	⬭	⬭	⬭	⬭	⬭	Strongly agree

7. Involvement of offenders

If an offender can see their assessments in the hypothetical system along with why their risk scores are high and what they can do to lower them, would this be helpful in actually getting them more involved in treatment?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly disagreee	◯	◯	◯	◯	◯	Strongly agree

8. Merit based approach

Would it make sense to show what would be the rewards for improvement in scores and what would be the potential sanctions? What could be these sanctions and rewards? Potential benefit would also additional data for better predictive models.

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly disagree	◯	◯	◯	◯	◯	Strongly agree

9. Which quantitative fairness measure is more important?

Märkige ainult üks ovaal.

◯ Equalized odds - where the true positive rates and false positive rates are equal across groups

◯ Equalized outcomes - where the difference in predicted outcomes between groups is less than the difference observed in the training data

10. Can anti-social attitudes be changed?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

11. Can improvement in anti-social attitudes increase responsivity to correctional treatments?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

12. Importance of support network (family, non-criminal friends) in desisting from crime?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

13. How important is offenders level of inner peace in terms of desistance from crime?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

14. Does offenders self-confidence and feeling of agency help in preventing recidivism?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

15. How important are offenders abilities to cope with stress in terms of overall recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

16. Is gang membership a good predictor of recidivism?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

17. How well does educational history predict recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

18. Please define if possibe: what are the warning signs you look for in educational history

19. How well does employment history predict recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

20. Please define if possible: what are the warning signs you look for in employment history?

21. The amount of divorces and/or alimonies correlates with recidivism risk

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

22. Social service and police contact count, frequency correlate with recidivism

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

23. Social service case severity correlate with recidivism

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

24. Creativity related treatments (art, music, etc) useful in reducing recidivism

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

25. Sports are important in reducing recidivism

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

26. How can leisure and recreational activities help in treating offenders?

27. How important is it, in terms of recidivism risk, where the offender has lived and currently lives (eg the neighbourhood, quality of accommodation)?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

28. When you examine an offenders criminal history, what do you look for in the first place?

29. Does age affect recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

30. Please elaborate further if needed on how does age affect recidivism risk

31. Does gender affect recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

32. Please elaborate further if needed on how does gender affect recidivism risk

33. Is offenders credit history important in terms of recidivism risk?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

34. Are outstanding debts, debts of partner, important?

Märkige ainult üks ovaal.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

35. Measurement: if you have one, please give an example of a prediction that would be useful for you as a user.

Thank you for yor answers so far! It would be very helpful for the research if you'd fine some minutes to answer some of following questions also:

36. What are the legal ramifications of different kinds of quantitative descriptions of risk (for example presumption of innocence)?

37. What should one keep in mind when giving risk, needs, or responsivity scores?

38. How to increase trust towards risk, need, responsivity assesments?

39. Transparency of assessments to offenders

What form should the feedback to offenders take?

40. How to engage offenders?

The willingness of offenders to engage in the process is negatively correlated to recidivism and thus crucial. What are your thoughts on this? What are the most pressing issues in engaging offenders?

41. Ethics of assessment

Predictions based on static risk factors are often as good as based on dynamic risk factors. On the one hand static risk factor information is easy to get, but these are something that offender cannot change. Gathering information on dynamic risk factors intrudes more into the private life of the offender, but at the same time, these are actionable. Considering that legal coercion is also being applied to the offender, where is the proper ethical balance? This is an important problem as the ethics of e-governance and especially of automation via algorithms/artificial intelligence/machine learning is being constantly regulated.

42. Transparency and interpretability

Please elaborate on the importance of transparency and interpretability of assessments

43. Involvement of offenders

If an offender can see their assessments in the hypothetical system along with why their risk scores are high and what they can do to lower them, would this be helpful in actually getting them more involved in the correctional process?

44. Merit based approach

Would it make sense to show to offenders what would be the rewards for improvement in scores and what would be the potential sanctions? What could be these sanctions and rewards? Potential benefit would also additional data for better predictive models.

45. Ethics of e-governance of offenders

When making predictions we are rendering uncertain futures knowable and actionable and by implication are ignoring subjectivity and the autonomy of the individual and thus also infringing on the presumption of innocence by acting pre-emptively.Thus we need understand what are our goals and what is acceptable risk. Where is the ethical line between justified e-governance of how offenders should behave and do, and where does it become unethical?

46. Parentalism

How to reduce the perception of parentalism by offenders and give them a sense of agency, but channeling that sense towards reintegration/rehabilitation goals.

47. Correctional program quality

How could the assessment of correctional program quality be improved?

48. Retention of professional discretion

How to ensure that professionals are firmly involved at crucial points in the correctional process? How to ensure that they do not become overly reliant on the assessment tools and the process itself? How to ensure that they retain the courage to apply their professional discretion, but also allow themselves be informed where needed? How to use the process to empower them to act on behalf of the victim, the offender, the state?

49. What are the main obstacles in changing offenders anti-social attitudes?

50. In prison an offenders network of anti-social peers will increase. How to mitigate its harmful effects?

Source: Composed by author, screenshots from Google Sheets

Appendix 2. Question to concept mapping

Table 3. Factor questionnaire questions along with the used terms

Question	Terms
Which quantitative fairness measure is more important?	measurement, instrument, users, offender, fairness
Can anti-social attitudes be changed?	risk, dynamic, treatment, responsivity, offender
Can improvement in anti-social attitudes increase responsivity to correctional treatments?	risk, dynamic, treatment, responsivity, offender
Creativity related treatments (art, music, etc) useful in reducing recidivism	protective, treatment, offender
Does age affect recidivism risk?	protective, non-crim-need, treatment, offender
Does gender affect recidivism risk?	protective, treatment, offender
Does offender's self-confidence and feeling of agency help in preventing recidivism?	protective, offender
How can leisure and recreational activities help in treating offenders?	risk, dynamic, offender
How important are offenders' abilities to cope with stress in terms of overall recidivism risk?	risk, dynamic, offender
How important is it, in terms of recidivism risk, where the offender has lived and currently lives (e.g. the neighbourhood, quality of accommodation)?	risk, static, offender
How important is the offender's level of inner peace in terms of desistance from crime?	risk, dynamic, offender
Are outstanding debts, debts of a partner, important?	risk, dynamic, offender
How well does educational history predict	risk, static, offender

recidivism risk?	
How well does employment history predict recidivism risk?	risk, static, offender
Importance of support network (family, non-criminal friends) in desisting from crime?	risk, dynamic, offender
In prison an offender's network of anti-social peers will increase. How to mitigate its harmful effects?	risk, dynamic, offender
Is gang membership a good predictor of recidivism?	protective, non-crim-need, treatment, offender
Is offenders credit history important in terms of recidivism risk?	protective, non-crim-need, treatment, offender
Please define if possible: what are the warning signs you look for in educational history	risk, static, offender
Please define if possible: what are the warning signs you look for in employment history?	risk, static, offender
Please elaborate further if needed on how does age affect recidivism risk	risk, static, offender
Please elaborate further if needed on how does gender affect recidivism risk	risk, static, offender
Social service and police contact count, frequency correlate with recidivism	risk, static, offender
Social service case severity correlate with recidivism	risk, static, offender
Sports are important in reducing recidivism	protective, non-crim-need, treatment, offender
The amount of divorces and/or alimonies correlates with recidivism risk	risk, static, offender
What are the main obstacles in changing	risk, dynamic, offender

offenders' antisocial attitudes?	
When you examine an offender's criminal history, what do you look for in the first place?	risk, static, offender

Source: Created for this study by the Author by breaking each question of appendix 1 into concepts

Table 4. Service design questionnaire questions along with the used terms

Question	Terms
Correctional program quality	measurement
Ethics of assessment	trust
Ethics of e-governance of offenders	trust
How detailed should scoring be?	measurement, trust, users, instrument
How important is it to describe the precision of an assessment to the user?	measurement, instrument, users
How important is it to give feedback to offenders on the risk, need, and responsivity assessments?	non-criminogenic need, trust, instrument, feedback, responsivity
How to engage offenders?	feedback, responsivity, offender
How to increase trust towards risk, need, responsivity assessments?	trust, users
Involvement of offenders	feedback, treatment, responsivity, offender, instrument
Is it morally more correct to assess risk via dynamic risk factors than via static risk factors?	dynamic factor, static factor, instrument
Measurement: if you have one, please give an example of a prediction that would be useful	measurement, instrument, users

for you as a user.	
Merit based approach	feedback, treatment, motivation, responsivity, offender, fairness
Paternalism	trust
Retention of professional discretion	instrument, users
Transparency and interpretability	trust, instrument, feedback
Transparency of assessments to offenders	trust, instrument, offender
What are the legal ramifications of different kinds of quantitative descriptions of risk (for example presumption of innocence)?	measurement, instrument
What should one keep in mind when giving risk, needs, or responsivity scores?	measurement
Which kind of risk measure do you prefer?	measurement, instrument, users
Which quantitative fairness measure is more important?	measurement, instrument, users, offender, fairness
Possibility to collaborate: request professional clinical judgement	users, treatment, responsivity
Possibility to collaborate: request additional assessments	users, instrument, treatment, responsivity
Possibility to collaborate: give feedback on the quantitative assessments	users, feedback, measure, responsivity
Possibility to collaborate: scheduler for managing work and interacting with offender or	users, responsivity

victim	
Possibility to collaborate: automated treatment recommendation	instrument, treatment, responsivity

Source: Created for this study by the Author by breaking each question from appendix 1 into constituent concepts.

Appendix 3. Non-exclusive licence

A non-exclusive licence for reproduction and publication of a graduation thesis[11]

I, Iren Irbe (*author's name*)

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Risk Prediction and Governance Based on Machine Learning: A Criminal Justice Case Study",
(*title of the graduation thesis*)

supervised by Associate Professor Anu Masso,
(*supervisor's name*)

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

_____ (date)

[1] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.*