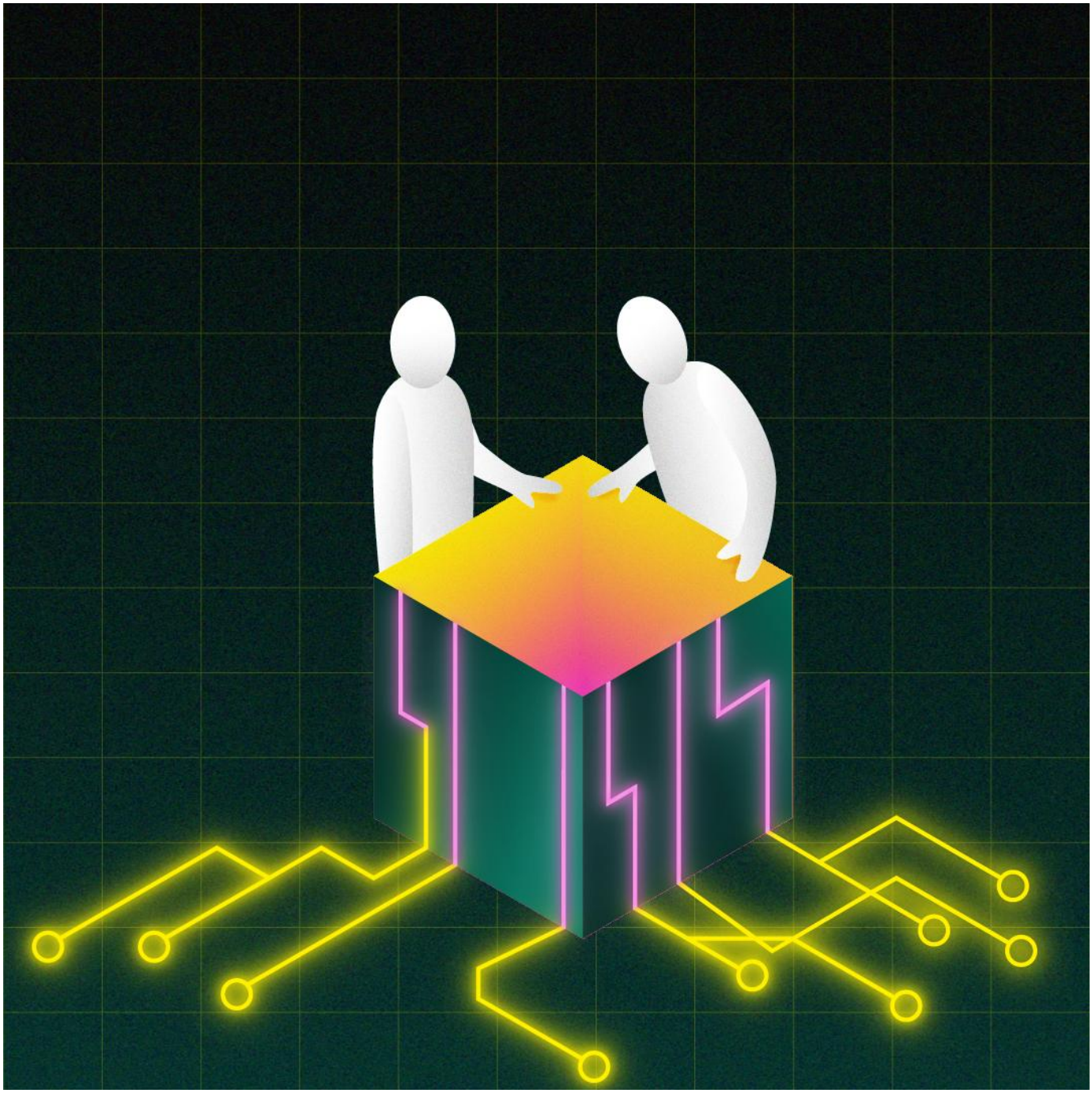


AUTOMATING SUSPICION

RISK PROFILING AS A SMOKESCREEN FOR STRUCTURAL DISCRIMINATION AND INEQUALITY

AMNESTY INTERNATIONAL

Amnesty International is a movement of 10 million people which mobilizes the humanity in everyone and campaigns for change so we can all enjoy our human rights. Our vision is of a world where those in power keep their promises, respect international law and are held to account. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and individual donations. We believe that acting in solidarity and compassion with people everywhere can change our societies for the better.

First published in 2026
by Amnesty International Ltd
Peter Benenson House, 1 Easton Street
London WC1X 0DW, UK

Index: POL 40/1096/2026
Original language: English

amnesty.org

Cover illustration: *© Almost Studio*

AMNESTY INTERNATIONAL

CONTENTS

1. GLOSSARY

WORD	DESCRIPTION
ACCURACY	In the field of artificial intelligence, accuracy measures are generally used to ascertain the number of "correct" outputs that a system produces, whether those outputs are predictions, identifications or simpler calculations (as a percentage of the number of total outputs made).
ALGORITHM	An algorithm is a procedure used to solve a problem or perform a computation. Algorithms act as an exact list of instructions that conduct specific actions step by step. Algorithms are used as specifications for performing calculations and data processing. Algorithmic systems are applications that perform one or more tasks, such as gathering, combining, cleaning, sorting, classifying and inferring data, as well as selection, prioritization, making recommendations and decision-making.
ALGORITHMIC DISCRIMINATION	Algorithmic discrimination occurs when automated systems contribute to unjustified differential treatment or outcomes that are unfavourable to people based on their race, colour, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, genetic information, or any other classification protected by law.
ALGORITHMIC RISK PROFILING	The semi- or fully automated processing of data for an assessment, evaluation or calculation (sometimes called "prediction") of the likelihood that individuals or groups will violate a law or rule. Based on the risk assessment and selection, a decision is then made to subject individuals or groups deemed "riskier" to additional enforcement measures such as more intensive scrutiny, fines or arrest. The decisions to investigate or punish individuals based on risk profiling thus target individuals based primarily on statistical, aggregate or group probabilities, not on actual individual behaviour that concretely indicates fraud or crime.
ARTIFICIAL INTELLIGENCE (AI)	Broadly speaking, AI is any technique or system that allows computers to mimic human reasoning. There is no widely accepted definition of the term "artificial intelligence" or "AI". But one definition describes AI as systems designed to carry out a specific task or process that "learn by doing" – whether through supervised learning (a system that is rewarded and corrected by a developer until it learns patterns over time) or newer methods of deep learning (systems programmed to learn in a more sophisticated way, modelled on processes in the human brain).
AUTOMATED DECISION-MAKING SYSTEM	An algorithmic decision-making system where no human is involved in the decision-making process. The decision is taken solely by the system.
FAIRNESS	There are numerous suggested methods, approaches and definitions for embedding fairness into AI systems to avoid algorithmic bias. They are all predicated on the idea of eliminating any prejudice, discrimination or preference

WORD	DESCRIPTION
	for certain individuals or groups based on a characteristic from the output of an AI system. Though fairness methods have become a routine element of ensuring AI systems are unbiased, Amnesty International generally considers them to be a limited tool, as discrimination and bias present within AI systems is not solely a technical issue.
FRAUD-DETECTION MODELS OR ALGORITHMS	ML algorithms used to identify recipients and claimants of social protection schemes who are at higher risk of committing fraud or an error in their application. The systems often use historical data on behaviours and characteristics that are considered to be commonly associated with fraud and error.
GROUND TRUTH	"Ground truth" refers to the information that is considered to be true and accurate based on direct observation or physical measurement. In various fields, especially in data science, remote sensing, and machine learning, it serves as the benchmark for validating the accuracy of models and predictions. The term implies a fundamental or baseline truth against which data or analyses can be compared.
MACHINE LEARNING (ML)	A sub-set of AI, ML is a technique to provide AI with the capacity to learn from data to perform a task (either specific or general) and, when deployed, ingest new data and change itself over time.
PREDICTIVE ALGORITHMS	The use of AI techniques to make future predictions about a person, event or any other outcome.
SEMI-AUTOMATED DECISION-MAKING SYSTEM	An algorithmic decision-making system where a human is involved in the decision-making process, or the algorithm is used to support the decision-making. Often, these systems are used to select cases for human review or to assist in the decision-making process by providing information and/or suggested outcomes.
SOCIAL PROTECTION	Social protection refers to a broad range of contributory programmes (those financed through contributions made by an individual or on their behalf) and non-contributory programmes (those funded through national tax systems). Social protection programmes can include (1) social insurance, such as pension insurance; (2) employment and labour programmes, including skills training, unemployment benefits and job search assistance; and (3) social assistance and cash benefits for people living in poverty or other categories of people entitled to them, such as parents for childcare subsidies.
SUSPECT GROUND	Non-discrimination provisions often explicitly enumerate certain grounds such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or "other status". These grounds are seen as "suspect" because they are immutable personal characteristics, irrelevant for performing in society, and/or go hand in hand with historical or social discrimination and stigmatization. When a ground is "suspect", there is a presumption that the difference in treatment cannot be justified.

ABBREVIATIONS

WORD	DESCRIPTION
AI	artificial intelligence
AI ACT	European Union Artificial Intelligence Act
CERD	UN Committee on the Elimination of Racial Discrimination
CESCR	UN Committee on Economic, Social and Cultural Rights
CJEU	Court of Justice of the European Union
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CRPD	Convention on the Rights of Persons with Disabilities
CSO	civil society organization
ECtHR	European Court of Human Rights
EU	European Union
FRA	Fundamental Rights Agency of the European Union
GDPR	European Union General Data Protection Regulation
HRC	UN Human Rights Committee
IACHR	Inter-American Court of Human Rights
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
ICESCR	International Covenant on Economic, Social and Cultural Rights
IHRL	international human rights law
ML	machine learning
OHCHR	Office of the UN High Commissioner for Human Rights
UK	United Kingdom
UN	United Nations
USA	United States of America

2. EXECUTIVE SUMMARY

This report builds on more than a decade of previous research by Amnesty International on the use of digital systems, artificial intelligence (AI), and risk profiling systems in the public sector of European countries, as well as research by other civil society organizations (CSOs) in other parts of the world. This body of research illustrates how the use of risk profiling systems across a wide range of contexts has consistently resulted in human rights abuses, particularly in high-stakes contexts, defined as the law enforcement, social security and migration policy domains. Amnesty International defines risk profiling as an assessment, evaluation or calculation (sometimes called "prediction") of the likelihood that individuals or groups will violate a law or rule.

This report provides an overview of these systemic issues, a framing and argumentation to support further investigative work on real-world cases, and provides rights holders, human rights advocates, civil servants, oversight authorities, lawyers and judges with trustworthy scientific and legal arguments to contest the use of risk profiling by states and other entities in high-stakes contexts.

A complete analysis and full understanding of the problems plaguing risk profiling algorithms that often result in human rights violations requires insights from several academic disciplines, as well as situating the technology in its historical and social context. Further, the report explores the often overlooked structural and intersectional effects of discriminatory risk profiling and illustrates how technology can perpetuate discrimination and inequality by providing states with a false veneer of objectivity. We further explore some of the most commonly proposed solutions to the issue of discriminatory profiling and highlight their limits. Finally, the report brings all the findings together in a legal analysis and bridges the gap between the academic literature and international human rights law (IHRL).

RESEARCH QUESTION

Governments are enthusiastic to implement risk profiling, despite numerous examples in which it has been shown to perpetuate discrimination, and without consideration of whether the use of risk profiling is necessary and compliant with IHRL. Partly, this is due to an assumption that technology, including risk profiling, is "objective" and "neutral". This façade of technological objectivity raises questions of accountability in case of harm, and makes it more difficult for affected individuals to seek redress and remedy.

Advocates of risk profiling claim that it allows organizations to take more objective and consistent decisions, to increase the effectiveness of decision-making, to streamline services and to improve cost-effectiveness by better allocating scarce resources in the fight against crime, fraud and in enforcing borders. These claims have become a dominant narrative to justify the use of risk profiling algorithms.

But despite these claims attributed to risk profiling systems in public policy, examples of their failure are abundant and there is remarkably little published evidence of their effectiveness in practice. This raises the question of whether risk profiling can ever be safely deployed in high-stakes contexts. To answer this question, Amnesty International surveyed a growing body of interdisciplinary academic literature on the impact of risk profiling, as well as on the suggested technical and policy solutions to its evidenced shortcomings.

WIDESPREAD HARMS AND HUMAN RIGHTS ABUSES

Risk profiling has been shown to cause discrimination on the grounds of, among others, race and ethnicity, gender, socio-economic status and disability. The discriminatory risks of risk profiling are especially evident

when analysed through the lens of intersectional discrimination. Amnesty International has documented various cases of harms in risk profiling across different countries, where people have been discriminated against based on one or multiple intersections of their race, ethnicity, national origin, gender, disability, age and social and economic status.

Additionally, the discriminatory impacts of risk profiling reach beyond biased outcomes and violation of individual rights, as they enable consolidation of more structural types of discrimination. Risk profiling systems have their roots in historical systems used for categorizing, cultivating and instrumentalizing personal data in order to create and maintain social and racial hierarchies and can best be viewed as an extension of these pre-existing systems of power.

In high-stakes domains, being subjected to risk profiling yields serious harms, which include both material and non-material effects. Victims suffer severe psychological distress, stigmatization and false accusations of fraud or crime, potentially leading to eviction or imprisonment. The financial fallout often leads to delayed or denied social benefits and crushing debt. People on the move, meanwhile, face the threat of unjust detention and deportation. Communities already experiencing marginalization suffer loss of autonomy and chilling effects from intrusive monitoring and, as a result, lose trust in institutions, which significantly affects state legitimacy and society as a whole. These abuses are worsened by a widespread lack of transparency, leaving individuals powerless to challenge these systems or defend their human rights.

Risk profiling poses risks to numerous other human rights, and these harms may also occur in a discriminatory manner, since the right to non-discrimination is a right unto itself, and also a cross-cutting right. Research by Amnesty International and other human rights organizations has consistently shown that risk profiling negatively affects the rights to a fair trial, remedy and redress, the presumption of innocence, the right to privacy and data protection, the rights to social security and an adequate standard of living and the full realization of human dignity. Because of the differential treatment, these harms are unevenly distributed across societal groups and likely to bring particular disadvantage to individuals belonging to marginalized groups.

KEY TAKEAWAYS FROM ACADEMIC LITERATURE

The goals and processes governments follow when building risk profiling systems resemble those of scientific research: governments attempt to gain knowledge by means of quantitative inference. In risk profiling, this knowledge takes the form of a prediction or assessment that a person will violate a rule, and the inference is based on the person's resemblance to past offenders. However, scientific research employs strict methods to prevent invalid practices and unreliable results. According to scientific literature surveyed by Amnesty International and conversations with leading academic experts, these strict methods and practices are not present for risk profiling, either in industry or governmental settings. This undermines the validity, robustness and reliability of risk profiling. A growing body of scholarship by prominent AI researchers qualifies risk profiling and other predictive practices as pseudoscience.

MULTIPLE AND INTERLOCKING TYPES OF BIAS ARE INHERENT TO RISK PROFILING

On the one hand, governments often use pre-existing and unfit-for-purpose administrative data to train predictive models (so-called convenience samples) rather than collecting new data that is specific or pertinent to the prediction of the behaviour labelled as "risky". This deviates from standard methodological practice applied in quantitative social science.

On the other hand, even with "flawless" data, it is impossible to design an "objective" or "neutral" risk profiling algorithm. Data about people is never "objective" – it is shaped by the social, historical and institutional context. When governments use past social data in order to predict who is going to commit a crime such as fraud, they inevitably target individuals who belong to historically oppressed or marginalized groups, thereby reproducing and compounding past injustices.

These data biases can never be completely eliminated by technical means or by adding more data, because they are caused by societal issues that underlie the data generation and collection processes. The misconception of thinking that all bias can be eliminated often results in flawed systems that reinforce systemic discrimination and inequality under the guise of technological neutrality. Risk profiling thereby enables the masking of implicit norms and existing inequalities and is highly likely to validate and amplify stereotypes.

Biases are not limited to the type of data and to its selection. When public authorities define a prediction goal – such as the risk of committing social security fraud – they embed institutional priorities and structural prejudices directly into the system's logic. These optimization goals dictate how the algorithm works,

meaning that even with "flawless" data, the system will still produce discriminatory outcomes if its core objective disproportionately scrutinizes marginalized populations.

RISK PROFILING IS METHODOLOGICALLY FLAWED

Literature on the scientific validity of prediction models raises a dire warning about their legitimacy.

Constructs like the individual risk of committing a crime or social security fraud are extremely complicated prediction goals for predictive systems, because they cannot be reliably operationalized and measured. Therefore, proxy constructs are used that are inaccurate and biased, such as instances of re-arrest as a proxy for re-offending, or involuntary mistakes in social security claims as a proxy for fraudulent claims. There is also a lack of reliable ground truth for profiling the risk of crime or fraud. As a result, the scientific validity of risk profiling models is fundamentally weak or absent, with potentially grave consequences for the people affected. Such models are prone to produce biased and erroneous predictions and expose targeted individuals to arbitrary outcomes.

Further, scientific literature describes "theory-free" predictive applications as scientific malpractice and potentially dangerous. The appeal of "data-driven" techniques, including risk profiling, resides partly in their ability to discover correlations directly from large datasets without relying on pre-specified theoretical assumptions. However, to guarantee that predictions generated by machine learning (ML) are well-founded, disciplinary methodological norms should be followed, such as articulating a causally plausible theory, grounding it in existing theoretical frameworks and then testing it on empirical data using gold standard methods of causal inference. The choice of which problem to solve, what to optimize for, which data to collect and how, its categorization, the choice and engineering of variables and the interpretation of model results are all informed by some sort of undeclared theoretical commitment. Making observations or taking measurements is inseparable from the guidance of background theory. If this theory is unspecified, it cannot be subjected to critical scrutiny. These rigorous methodological heuristics are glaringly absent from governmental risk profiling practices.

Another key practice of trained scientists is ruling out meaningless or spurious correlations and researching the causal mechanisms underlying an observed correlation. Since spuriousness is not ruled out, and because background theory is unspecified, risk profiling models are not robust and misrepresent the human reality in which they are deployed. There is no reliable way to know when, where or how the model's predictions will fail, leading to harm for targeted people. Arbitrary correlations will be misinterpreted as causal and treated as empirical truths, with the risk of masking implicit ideas and norms. Patterns in the social world reflect societal norms, conventions and social structures and nothing is inherently "natural" or "neutral" about individuals or social groups.

LIMITS OF PREDICTION IN COMPLEX SOCIAL SYSTEMS

Human beings and their intentional behaviour can be seen as complex adaptive phenomena, which makes them inherently indeterminable. This conclusion finds empirical support from large-scale computational social science studies. Accordingly, certain prediction tasks of human behaviour cannot be solved by ML. These are settings in which no specific AI developed for the task can ever possibly work. In some cases, there is no plausible connection between observable data and the proposed behaviour being predicted, such as between race or ethnicity and criminality - namely, racial profiling. In other cases, regardless of the amount of data, there is no data, or proxy data, that is good enough or objective enough to adequately model the underlying phenomenon. Such systems include risk profiling that attempts to predict criminality, life course or the propensity to commit social security fraud at the individual level or a specific location. These predictive systems have been debunked and decried as scientific malpractice. They have raised a distinctive and serious set of normative concerns that causes risk profiling systems to fail on their own terms because they do not deliver accurate predictions.

DOMINANT NARRATIVE IS UNSUBSTANTIATED

Risk profiling systems are often cited as a method by which states can streamline services, improve cost-effectiveness, prevent crimes, including fraud, and control migration. These claims, based on the premise of scarce resources, have been debunked as a policy rationale, because they are empirically unsupported and politically useful assumptions that turn poverty and other social issues from political problems into an "efficiency" problem to be solved through automated rationing and surveillance. Instead of the promised benefits, a more consistent outcome is the penalization of society's most marginalized groups for attempting to access their rights and/or essential services. Amnesty International and others have shown in numerous case studies that risk profiling systems disproportionately associate people who already experience one or multiple forms of discrimination or marginalization with higher criminal or financial "risk".

DISTURBING PARALLELS WITH EUGENICS AND SCIENTIFIC RACISM

Accepted norms in certain applied ML settings – such as government risk profiling – are strikingly similar to those of early eugenicists. Eugenicists promoted the objectivity of the numbers and methods with which they worked and advocated for "letting the numbers speak for themselves". Treating correlation as inherently predictive and avoiding recourse to theory and causality was a distinguishing practice of eugenics and scientific racism. By doing this, they explicitly obfuscated their racist ideas and presented numbers as devoid of political values. It is disturbing that such methods to score and rank individuals have returned to the policy scene, albeit under different stated aspirations: to select individuals for enforcement and scrutiny.

CONCLUSION FROM ACADEMIC LITERATURE

Therefore, constructing a risk profile for social security fraud or criminality is not a realistic technical undertaking, nor is it a credible exercise in evidence-based policy. It is an attempt to operationalize suspicion in the absence of a reliable ground truth, and it is bound to discriminate because of biases inherent to the data and the social phenomena that they quantify.

EVALUATION OF RISK PROFILING UNDER INTERNATIONAL HUMAN RIGHTS LAW

As well as a survey of the academic literature, this report provides a legal analysis of whether risk profiling causes differential treatment based on prohibited grounds and whether a reasonable and objective justification exists for this differential treatment.

DIFFERENTIAL TREATMENT BY DESIGN

Using past social data to predict crime or fraud inevitably results in the targeting of individuals belonging to historically oppressed or marginalized groups. This amounts to differential treatment based on prohibited grounds. If this differential treatment has no reasonable and objective justification, it is discriminatory under international human rights law (IHRL). Individuals are sorted and placed on socially constructed hierarchies at birth and assigned unequal opportunities, which ultimately result in different life outcomes, only to then come under surveillance from the institutions supposed to protect them. Instead of predicting future behaviour, risk profiling in high-stakes domains largely reproduces past injustices.

Common but inadequate responses to bias in algorithmic systems include removing prohibited characteristics from datasets or their individual proxies - such as postcode serving as a proxy for race. We find that these approaches fail to address the deeper, structural issues embedded in the data. Prohibited characteristics will be implicitly captured in other interconnected variables, such as one's spending patterns, living arrangements or no-shows in medical appointments. These variables can serve as indirect proxies because systemic inequalities and discrimination have produced meaningful disparities across entire domains of life. This reveals an uncomfortable truth: the prevalence of characteristics that may be relevant for risk profiling and are seemingly "neutral" are unequally distributed across groups. Such differences in base rates can reflect historical and ongoing discrimination at an institutional or systemic level.

CRIME AND FRAUD REDUCTION AS A COVER FOR INVASIVE SURVEILLANCE

While reducing fraud and other types of crime are legitimate aims for governments, there is a concrete risk of such aims being abused to form a cover for invasive, community-wide surveillance. In the absence of solid evidence related to levels of benefits fraud warranting intrusive surveillance and risk profiling, it is important to be critical of these stated aims. Risk profiling is most often deployed in contexts where it is likely to affect groups that are stigmatized, disenfranchised or otherwise already at the margins of society, while the most privileged members of society are largely exempt from narrowly targeted monitoring. This selective attention and use of invasive tools are often the result of pre-existing stereotypes and prejudices which posit marginalized groups in particular as inherently criminal or dangerous. These stereotypes are compounded by structural issues, including racial profiling, over-policing and higher conviction rates of racialized people.

IS RISK PROFILING NECESSARY AND EFFECTIVE?

To test any predictive system against the prohibition of discrimination, policymakers should clearly and transparently define outcome measures that map directly onto their stated legitimate aims; specify the causal pathways by which predictions are expected to produce those outcomes; and require evidence that interventions informed by predictions produce the intended societal benefits rather than merely altering downstream administrative statistics or market-driven notions of efficiency.

Instead, states are often caught up in narratives of tech inevitability and focus on mistakenly narrow notions of "effectiveness" – for example, by conflating effectiveness with the predictive performance of risk profiling algorithms in comparison to random selection. This is a shallow and unreliable measure of "effectiveness" in the context of a test against the prohibition of discrimination. Differentiation based on risk profiling is mostly based on statistical correlations and has no theoretical or empirical support. Unless spuriousness has been rigorously ruled out, the correlation should be assumed to be spurious. A simple correlation says nothing meaningful about the likelihood that an individual or group will violate a law or commit fraud. Therefore, statistical generalizations on increased detection rates cannot count as justifications for subjecting marginalized groups or individuals to differential treatment. This utilitarian calculus is incompatible with the values embodied by non-discrimination provisions in IHRL.

The literature and case studies surveyed by Amnesty International show that risk profiling systems are fundamentally inaccurate and produce an alarmingly high number of false positives – mostly distributed to racialized and marginalized people. Despite the empirical evidence, states fail to recognize these severe operational failures. Furthermore, even if risk profiling predictions were accurate, these predictions would not automatically lead to effective interventions. Indeed, there is a lack of an actual and verifiable reduction in offences following risk profiling predictions, which compromises the evaluation of the effectiveness of these systems under IHRL. Even assuming that risk profiling delivered accurate predictions, research shows that high predictive performance alone rarely meets the real-world goals for deployment of predictive models. Relating short-term measurable outcomes, such as risk predictions, to broader policy goals is extremely difficult. For example, research shows that predictive policing approaches have failed to deliver on broader policy goals, such as reducing or preventing crime. It is imperative that governments' stated aims be interrogated to demonstrate that both the chosen policy goals and tools meet a defined objective, and that states are not allowed to rely on overly broad generalizations to justify restrictions on rights.

In sum, the effectiveness of risk profiling as a policy intervention is, at best, subject to debate.

STEREOTYPING BY DESIGN AND PERFORMATIVE EFFECTS OF RISK PROFILING

Even assuming that a risk profile might have some, possibly unexplained, predictive performance on a group level, it will still be largely incorrect and therefore unfair at the individual level, and deny individuals the opportunity to defy an apparent norm. By selecting individuals for checks on the basis of a risk profiling prediction, governments *produce* its causal effects and inescapably treat an individual *as if* they had acted suspiciously or indeed violated the law. When people are subjected to additional scrutiny following risk profiling, they are being treated as de facto suspects. Correlations are treated as inherently predictive and therefore punish people for being part of a (statistical) group, which amounts to stereotyping.

This will inevitably result in observing a higher number of violations within these groups and lower rates in other groups. This phenomenon ultimately results in feedback loops and has been studied extensively. The data might then become the input for training future risk profiling models. The effect of risk profiling is therefore *performative:* individuals or groups are transformed from statistical, hypothetical suspects into actual suspects, solidifying pre-existing prejudices or generating new ones.

This is compounded by existing systemic discrimination. Since any measured trend or correlation is not inherently indicative of causality, and because constructs such as race or ethnicity do not causally relate to criminality, fraud or migration offences, these correlations actually uncover systemic discrimination rather than "risk".

INEFFECTIVE AND INSUFFICIENT SAFEGUARDS

Technical measures to prevent or repair discriminatory outcomes, such as algorithmic fairness, have proven inadequate to prevent discrimination in practice. The core problem lies in treating statistical methodology as a substitute for addressing underlying social and structural biases. These measures have consistently failed to address human rights concerns and have come under increasing criticism from scholars. Technical safeguards impede effective measures such as regulatory bans, ultimately eroding accountability. Meaningful human intervention prior to the final punitive decision remains an insufficient safeguard, first because the differential treatment has already materialized, second because of the deriving performative effects, and third because it fails to adequately mitigate automation bias, which remains a significant and unsolved issue.

CONCLUSION: RISK PROFILING IS INHERENTLY DISCRIMINATORY

In light of the grave harms inflicted by risk profiling on the individuals and communities affected, weighted against the significance of its aims and the overarching doubts concerning its effectiveness and necessity, the differential treatment inherent in risk profiling systems cannot be deemed proportionate and thus cannot

be reasonably and objectively justified. It follows that risk profiling in high-stakes contexts is incompatible with IHRL.

KEY RECOMMENDATION

Amnesty International believes that the use of data-driven or rule-based predictive, profiling and risk assessment systems, regardless of human involvement in the final decision, should be prohibited in high-stakes contexts, defined as law enforcement, social security and migration. States should develop or amend existing AI regulation to ensure this prohibition. Until such regulation is set in place and regardless of changes to regulation, public authorities must urgently discontinue the use of these systems. See the final section of this report for a full list of recommendations.

3. READING GUIDE AND METHODOLOGY

This report is part of Amnesty International's body of work on AI technologies and human rights, which seeks to uncover human rights-violating technology and routes to justice for those affected. This report serves to provide an overview of the systemic issues in the use of risk profiling in high-stakes contexts (see below), to provide helpful framing and argumentation to support further investigative work into real-world cases, and to provide rights holders, human rights advocates, civil servants, oversight authorities, lawyers and judges with trustworthy scientific and legal arguments to contest the use of risk profiling by states and other entities including in high-stakes environments. This is achieved by examining the scientific flaws that cause risk profiling algorithms to "fail on their own terms" – specifically, their fundamental inability to deliver the accurate predictions they promise – and demonstrating how this technical inadequacy renders them incompatible with IHRL. Further, the report will explore the often overlooked structural and intersectional impacts of discriminatory risk profiling and illustrate how technology can perpetuate discrimination and inequality by providing states with a false veneer of objectivity.

This report draws on several types of evidence and is structured accordingly:

- Chapter 4 defines important terminology used in this report.
- Chapter 5 offers a brief historical overview and synthesis of the current situation.
- Chapter 6 describes the applicable IHRL framework.
- Chapter 7 discusses previous research by Amnesty International and other CSOs that exposes the harms of risk profiling systems across multiple policy domains.
- Chapter 8 reviews relevant academic scholarship on the scientific flaws of risk profiling technology.
- Chapter 9 takes a critical perspective on the technical approaches that are currently dominating policy responses to the harms inflicted by risk profiling systems.
- Chapter 10 analyses the human rights harms inflicted by risk profiling systems and asks whether these systems comply with IHRL in light of the evidence presented.

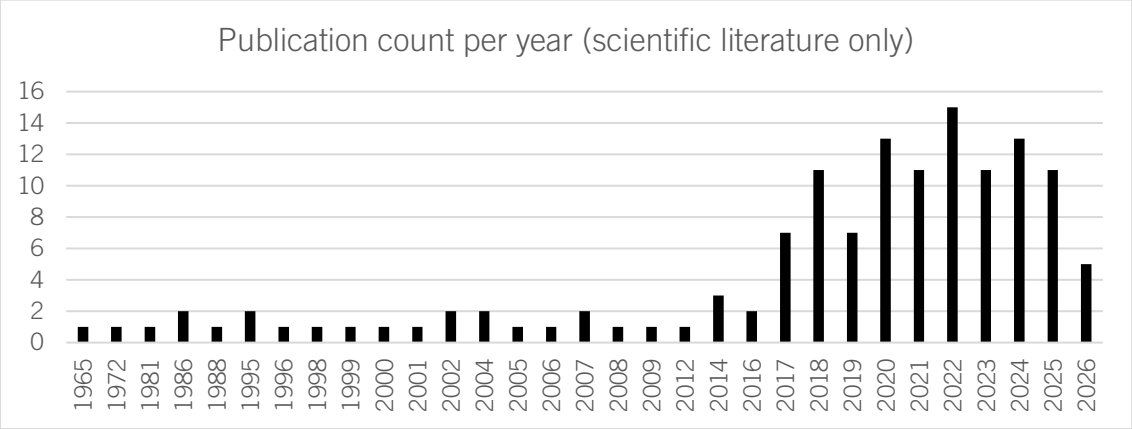
Chapter 7 builds on more than a decade of research by Amnesty International and other CSOs on the use of algorithmic risk profiling systems in the public sectors of various countries. Amnesty International has shown how the use of risk profiling systems across a wide range of contexts has consistently resulted in human rights violations. The risk of violations is especially high in contexts where decision-making by governmental authorities affects the rights and freedoms of individuals, and where it can have a major impact on society. Such contexts include seemingly dissimilar areas such as social protection, policing and border control, which are referred to as "high-stakes contexts" throughout this report. When algorithmic risk profiling systems are used in high-stakes contexts, they have an enormous impact on people's daily lives, particularly for individuals from marginalized and racialized groups. These groups have historically been subject to heightened scrutiny and profiling.

This raises the question of whether risk profiling can ever be safely deployed in high-stakes environments. To answer this question, Amnesty International turned to a second source of evidence and analysed the growing body of academic literature that highlights a pattern of similar shortcomings in applications of predictive AI

systems. We surveyed 133 peer-reviewed publications from 1965 to 2026 – including articles, systematic reviews and books – across more than a dozen scientific disciplines. Taken together, this literature shows an increasing interdisciplinary consensus that prediction of complex social behaviour is extremely difficult, unrealistic or practically unfeasible. This evidence is presented in Chapter 8. We also surveyed literature that was less adamant on this unfeasibility,[1] and actively sought out publications that reported more positive findings on prediction, as well as discussing the findings with academic experts from the Global North who hold more favourable positions on prediction.

A complete analysis and full understanding of the problems plaguing risk profiling algorithms that often result in human rights violations requires insights from several related disciplines, including statistics, data science, bias and fairness in ML, and computational social science, as well as situating the technology in its historical and social context, which is the focus of disciplines including sociology, science and technology studies and critical data studies. This report bridges the gap between the academic literature from these various fields and IHRL.

In this process, we reached out to nine Global North-based academic experts in the fields of computer science, AI and ML, data governance, AI ethics and algorithmic fairness, statistics, biostatistics, and philosophy of science. The final text was reviewed by seven internationally renowned experts from the Global North in the disciplines of computer science, AI and ML, statistics, AI accountability, cognitive science, data governance, and philosophy of science and technology – with some overlap with the consulted experts mentioned above.



We also surveyed reports and recommendations from other CSOs including Human Rights Watch, European Digital Rights, Fundación Karisma, Derechos Digitales and Algorithm Watch; reports and recommendations from UN bodies including the Human Rights Committee (HRC), Special Procedures, the Office of the UN High Commissioner for Human Rights (OHCHR) and various reporting committees on the implementation of treaties; reports from regional bodies including the Council of Europe, European Data Protection Board, European network of legal experts in gender equality and non-discrimination, the EU Fundamental Rights Agency, and the Inter-American Commission on Human Rights; case law from regional and international jurisdictions; publications from national scientific and research bodies; scholarly legal and human rights literature; and investigative journalism and other media reporting.

Amnesty International did not investigate individual cases of people affected by risk profiling during research for this report. While individual cases are crucial to understanding the lived experiences of those affected by risk profiling systems, these cases are included in the referenced reports and case studies rather than being included directly in this report.

[1] Jiani Yan and Charles Rahal, "On the unknowable limits to prediction", March 2025, Nature Computational Science, Volume 5, Issue 3, https://www.nature.com/articles/s43588-025-00776-y

4. DEFINITIONS

4.1 RISK PROFILING

Amnesty International defines risk profiling as an assessment, evaluation or calculation (sometimes called "prediction") of the likelihood that individuals or groups will violate a law or rule.

Risk profiling is usually based on one or more characteristics (criteria, indicators) on which an estimation is made of the risk of an offence. According to the organization that uses the risk profile, those characteristics are purportedly associated with a higher risk of violating a norm, rule or law. Risk profiling can involve undocumented assessment by individual officials, such as police agents on the street or at the border, or more structured forms of assessment, including with the support of algorithms. For example, a risk profiling algorithm can score social security claimants for risk of fraud and automatically take an enforcement decision – known as automated risk profiling. Alternatively, the selection resulting from the risk assessment may be checked by an officer before taking a final decision on enforcement measures. This second variant is called semi-automated risk profiling.

Therefore, whether or not fully automated, a selection decision is made based on the risk assessment and individuals or groups deemed "riskier" are subjected to additional enforcement measures such as more intensive scrutiny, fines or arrest. Decisions to investigate or punish individuals based on risk profiling thus target individuals based primarily on statistical, aggregate or group probabilities, not on actual individual behaviour that concretely indicates fraud or crime. A defining feature of risk profiling is therefore that it is used proactively by governments without a concrete, individualized suspicion that a specific person has committed an offence. Descriptions of a suspect used in criminal justice, therefore, do not qualify as risk profiling because they are based on a concrete and individualized suspicion that a person has committed a crime.

In this report, the automation of risk attribution is implied whenever we use the term "risk profiling". However, from an IHRL perspective, the same issues and risks associated with profiling apply whether or not the process is automated.

4.2 HOW RISK PROFILING WORKS

Automated profiling algorithms work by processing large amounts of data and determining statistical patterns that associate the inputs (personal characteristics) to an output – typically, the behaviour that is considered worthy of additional scrutiny by authorities. Because these statistical patterns are group-based, they apply to populations rather than to individual people. This ultimately leads to treating individuals as members of a (statistical) group.[2] The assumption underlying risk profiling is therefore that people with "similar" characteristics will behave similarly in the future. A central problem in risk profiling is that similarity between people is not defined by behaviour that is in any way "suspect" or indicative of wrongdoing. Instead, similarity is often defined mainly by recorded identity markers and unobservable traits, which are themselves socially constructed and imperfect proxies for a person's values and preferences. These markers and how they are measured can lead to discrimination for the people affected.

[2] Lorna McGregor and others, "International human rights law as a framework for algorithmic accountability", April 2019, International and Comparative Law Quarterly, Volume 68, Issue 2, https://www.cambridge.org/core/product/identifier/S0020589319000046/type/journal_article

4.3 RULE-BASED RISK PROFILING

A rule-based algorithm is a system that is based on pre-determined rules to automatically execute operations. These rules can be defined by humans.[3]

4.4 HIGH-STAKES CONTEXTS

The risk of human rights violations is especially high in contexts where decision-making by governmental authorities affects the rights and freedoms of individuals, increases their precarity or otherwise affects their life trajectory, and where it can have a major impact on society by solidifying or increasing existing power imbalances. Such contexts include seemingly dissimilar areas such as social protection,[4] policing[5] or border control,[6] and are referred to as "high-stakes contexts" throughout this report. When algorithmic risk profiling systems are used in high-stakes contexts, they have an enormous impact on the daily lives of people, particularly those from racialized and other marginalized groups. These groups have historically been subject to heightened scrutiny, profiling and criminalization by authorities as a result of systemic racism and other forms of discrimination. In this report, our analysis and recommendations are mostly specific to risk profiling deployed in these high-stakes contexts.

OTHER PROFILING DEFINITIONS

The Council of Europe has defined profiling in recommendations that concern the applicability of the Convention for the protection of individuals with regard to the processing of personal data (Convention 108+). Profiling "refers to any form of automated processing of personal data, including use of ML systems, consisting in the use of data to evaluate certain personal aspects relating to an individual, in particular to analyse or predict aspects concerning that person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."[7] Under these definitions, risk profiling by states in high-stakes contexts, as defined by Amnesty International, amounts to "high risk profiling".[8]

Another definition of profiling is given by the Fundamental Rights Agency of the EU (FRA), which defines profiling as an activity that involves categorizing individuals according to personal characteristics, thereby "using existing data to make assumptions about an individual outcome or behaviour".[9] Yet another definition can be found in Article 4(4) of the General Data Protection Regulation of the EU (GDPR), where profiling is defined as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements". In its *guidelines on automated individual decision making and profiling*, the European Data Protection Board further specifies that: "Profiling is a procedure which may involve a series of statistical deductions. It is often used to make predictions about people, using data from various sources

[3] Academic reviewers pointed out that people increasingly rely on large language models to define such rules.
[4] Amnesty International, *Social Protection in the Digital Age* (Index: POL 40/7771/2024), 6 March 2024, https://www.amnesty.org/en/documents/pol40/7771/2024/en/
[5] Amnesty International, *Netherlands: We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands* (Index Number: EUR 35/2971/2020), 2020, https://www.amnesty.org/en/documents/eur35/2971/2020/en/; Amnesty International, *UK: Automated Racism - how police data and algorithms code discrimination into policing*, 2025, https://www.amnesty.org.uk/files/2025-02/Automated%20Racism%20Report%20-%20Amnesty%20International%20UK%20-%202025.pdf?VersionId=JqCcTODw37yAXyINmAY6uAzrKEWucFF7; Amnesty International, *Trapped in the Matrix: secrecy, stigma and bias in the Met's gang database*, 2018, https://www.amnesty.org.uk/files/reports/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf
[6] Amnesty International, *The Digital Border: Migration, Technology and Inequality* (Index: POL 40/7772/2024), 2024, https://www.amnesty.org/en/documents/pol40/7772/2024/en/
[7] Council of Europe, "Recommendation of the Committee of Ministers to member States on the protection of individuals with regard to automatic processing of personal data in the context of profiling (CM/Rec(2021)8)", 2021, https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680a46147%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}
[8] See "High risk profiling" under the appendix to the Recommendation, para. 1(j) points 1-4.
[9] FRA, Preventing Unlawful Profiling Today and in the Future: A Guide, 2018, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-preventing-unlawful-profiling-guide_en.pdf

to infer something about an individual, based on the qualities of others who appear statistically similar."[10]

4.5 PAST VERSUS FUTURE EVENTS, FRAUD DETECTION, ANOMALY DETECTION AND PROFILING

IS TARGETING PAST OUTCOMES STILL RISK PROFILING?

One important difference between predictive tasks and fraud detection is that fraud detection focuses on past events or events taking place at the current time, whereas prediction is an attempt to forecast future events based on patterns from the past. A relevant question is therefore whether social security fraud detection qualifies as risk profiling. Given the above definition of risk profiling, one could conclude that risk profiling only encompasses the prediction of future events. However, a close look at the underlying logic implies that social security fraud detection is still a form of risk profiling.

To explain this further, statistics inherently compress temporal processes into static snapshots.[11] In the context of state risk profiling, however, this temporal compression does not alter the underlying conceptual problem. Authorities attempt to determine the likelihood of an uncertain event – such as an offence – by grouping people, estimating group-level risk and then mapping that risk onto individuals. This holds true regardless of whether the suspected offence occurred in the past or may occur in the future. Ultimately, if "detection" relies on the statistical likelihood of an event based on people sharing personal characteristics rather than concrete individual behaviour, it still constitutes risk profiling. This is best exemplified by comparing the use of fraud detection techniques that have been described as successful in detecting credit card fraud, but that transpose poorly to social security fraud.

ANOMALY DETECTION OR RISK PROFILING?

Although in several cases researched by Amnesty International, governments use ML or rule-based risk profiling, governments do sometimes also apply anomaly detection techniques to detect fraud in social security payments allocation.[12]

Anomaly detection methods in credit card fraud detection look for prominent and unusual patterns of behaviour relative to an individual cardholder's normal activity, for example, geolocation, time and frequency of transactions. An anomaly is taken as a signal of a discrete and actionable event, such as identity fraud or the card being stolen.[13] Therefore, in credit card fraud detection, an anomaly is assumed to map well to an offender's behaviour or *modus operandi*.

In social security, anomalies are compared to a vaguely defined administrative "norm" across a population, such as household composition and living arrangements.[14] Deviations from this norm are treated as indicators of potential fraud. Therefore, "anomalies" in social security fraud are often actually proxies for demographic differences. This means that anomaly detection in this domain effectively amounts to risk profiling based on socio-demographic factors and not simply identification of a past instance of fraud.

For the reasons outlined in this section, social security fraud detection, including when it is achieved through anomaly detection, can be characterized as risk profiling for the purpose of this report. Section 8.2 further details validity issues in social security fraud detection.

[10] Article 29 Data Protection Working Party, Guidelines on Automated individual Decision Making and Profiling for the purposes of Regulation 2016/679, 2018.
[11] Judea Pearl, Causality, 2009.
[12] See, for example, Amnesty International, *Denmark: Coded Injustice: Surveillance and Discrimination in Denmark's Automated Welfare State* (Index: EUR 18/8709/2024), 2024, https://www.amnesty.org/en/documents/eur18/8709/2024/en/
[13] Waleed Hilal and others, "Financial fraud: a review of anomaly detection techniques and recent advances", May 2022, Expert Systems with Applications, Volume 193, https://www.sciencedirect.com/science/article/pii/S0957417421017164, 116429, p. 6.
[14] Amnesty International, *Denmark: Coded Injustice* (previously cited).

4.6 FACTUAL INDIVIDUAL INDICATORS VERSUS RISK PROFILING

Governmental organizations that conduct tasks such as allocating and distributing social security payments or registering tax filings are also tasked with ensuring that individuals have followed administrative procedures correctly and are entitled to the allowances for which they apply. This is a legitimate aim for these organizations. Similar to risk profiling, governmental organizations can carry out automated checks to ensure that, for example, social security claimants do not violate a law or rule. Checks or corrections can be based on factual and individual observations that an error has been made in a benefits application or income tax declaration. Such observations do not qualify as risk profiling, because they indicate a real, concrete and individualized behavioural observation, as opposed to a likelihood calculated from how other 'similar' people have behaved on aggregate.

For example, factual discrepancies between the employer and filer around income from work are a legitimate reason for a tax authority to apply a correction or to ask for additional documentation. However, if such rules disproportionately flag marginalized groups, leading to differential treatment, they might be indirectly discriminatory and must be interrogated and tested accordingly. Moreover, governments should apply penalties with caution in the context of social security, and the burden of proof for all aspects of accusations of abuse or fraud must lie firmly with the state authorities, not with those facing investigation.[15] The combination of automating fraud detection with harsh and punitive enforcement policies is a serious risk to the rights of the people who rely on social protection schemes, with particular risks for people subject to structural and systemic racial discrimination, inequality and marginalization.

[15] Amnesty International, *Profiled Without Protection: Students in The Netherlands Hit by Discriminatory Fraud Detection System* (Index: POL 40/7108/2023), 2024, https://www.amnesty.org/en/documents/pol40/7108/2023/en/

5. BACKGROUND

The claims made about the benefits of risk profiling algorithms have become a dominant narrative to justify their use. Advocates of risk profiling claim that it allows organizations to make more objective and consistent decisions, to increase the efficiency of decision-making and to better allocate scarce resources. While efforts to reduce administrative costs in delivering welfare payments may be a legitimate goal, the introduction of digital technologies is taking place in the context of highly politicized narratives that welfare benefit fraud is out of control.[16] Predicting who is going to misuse social security payments, overstay their visa or commit a crime gives governments a defence to focus enforcement efforts on individuals who are not otherwise individually suspected of wrongdoing.

5.1 POLICING, LAW ENFORCEMENT AND ADMINISTRATION OF JUSTICE

In criminal law enforcement, profiling has produced many problematic precedents, most notably racial profiling. This discriminatory practice was found in 2023 to be "prevalent in all places in which racially marginalized persons live" by the UN Human Rights Council Advisory Committee.[17] Despite widespread recognition that racial profiling is both incompatible with IHRL and an ineffective law enforcement tool,[18] there are little signs that the practice is decreasing.[19] Scholars recently called for a European convention against racial profiling.[20]

With the development of new technology that facilitates machine prediction, risk profiling has been automated and applied to policing. Predictive policing fits within a wider trend in which governments increasingly adopt more measures that they think will exclude or reduce future risks and dangers.[21] Police forces use these systems to attempt to predict where an alleged crime will occur and to predict and profile who will commit a crime in the future or who is at "risk" of committing a crime or other criminalized behaviour. Police use these so-called predictions, profiles and risk assessments to target specific locations,

[16] Gabriel Geiger, "How Denmark's welfare state became a surveillance nightmare", 2023, https://www.wired.com/story/algorithms-welfare-state-politics/; "Knowledge of social security fraud and wrong payments", February 2014; Annika Lindberg, "The production of precarity in Denmark's asylum regime", 2020, Zeitschrift für Sozialreform, Volume 66, Issue 4.
[17] UN Human Rights Council Advisory Committee, Advancing Racial Justice and Equality by Uprooting Systemic Racism: Report, 8 August 2023, UN Doc. A/HRC/54/70.
[18] See, for example, ECtHR, *Seydi and Others v. France*, Application 35844/17, Judgment, 26 June 2025; Committee on the Elimination of Racial Discrimination (CERD), General Recommendation 36: Preventing and Combating Racial Profiling By Law Enforcement Officials, 17 December 2020, UN Doc. CERD/C/GC/36; Inter-American Court of Human Rights, *Acosta Martínez et al. v. Argentina*, Judgment, 31 August 2020; Human Rights Committee (HRC), *Williams Lecraft v. Spain*, 17 August 2009, UN Doc. CCPR/C/96/D/1493/2006.
[19] International Independent Expert Mechanism to Advance Racial Justice and Equality in Law Enforcement, Systemic Racism Against Africans and People of African Descent in the Criminal Justice System: Report, 9 September 2025, UN Doc. A/HRC/60/75; OHCHR, Promotion and Protection of the Human Rights and Fundamental Freedoms of Africans and of People of African Descent Against Excessive Use of Force and Other Human Rights Violations by Law Enforcement Officers, 1 June 2021, UN Doc. A/HRC/47/53; Inter-American Commission on Human Rights, African Americans, Police Use of Force, and Human Rights in the United States, 26 November 2018.
[20] Karin de Vries, "Is it time for a European Convention against Racial Profiling?", 2024, Netherlands Quarterly of Human Rights, Volume 42, Issue 3, https://doi.org/10.1177/09240519241274846
[21] David Garland, The Culture of Control: Crime and Social Order in Contemporary Society, 2001; Amnesty International, *Netherlands: We Sense Trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands* (Index: EUR 35/2971/2020), September 2020, https://www.amnesty.org/en/documents/eur35/2971/2020/en/

and people and groups in those locations, with increased policing. The aim is to target certain individuals and intervene before the predicted behaviour has occurred.[22]

5.2 USE IN OTHER DOMAINS

The emphasis on "pre-crime" or proactive, pre-emptive forms of enforcement is increasingly being introduced in other domains where human rights risks are high, such as fraud detection among social security claimants[23] and migration.[24] Governments justify the use of risk profiling technology in the prediction or detection of social security fraud with promises of increased effectiveness and efficiency of enforcement policies. In the migration domain, states around the world are increasingly deploying invasive border technology, including risk profiling tools used to process or determine migration or asylum status.[25] For example, disproportionate and unlawful surveillance and other measures increasingly used for racial profiling and policing create and sustain human rights violations, and are also increasingly adopted for use against asylum seekers, refugees and migrants.[26] The human rights risks are essentially the same across the social protection and migration domains: people should be treated as individuals based on their behaviour and not their actual or perceived race, ethnicity, gender, sexuality, religion, disability, or national or social origin.

Digital technologies have their roots in historical systems used for categorizing, cultivating and instrumentalizing personal data. Digital systems are used to create and maintain social and racial hierarchies and can best be viewed as an extension of these pre-existing systems of power.[27] Although technology in the public sector is often presented as objective and unbiased, it is virtually impossible to create a value-neutral technology or database that is free from bias.[28] Technology is not built and introduced in a vacuum, but into existing societies with specific social and political dynamics, where it can have potentially unpredictable and unintended consequences for individuals. Such effects can also vary widely depending on whether those individuals are already subject to systemic and intersectional forms of discrimination and marginalization. To identify and mitigate any potential bias, discrimination and other human rights harms, governments and policymakers should fully understand both the context in which these systems are deployed and the existing power imbalances and inequalities that underpin such systems.[29] As the UN Special Rapporteur on contemporary forms of racism notes, states "must address not only explicit racism and intolerance in the use and design of emerging digital technologies" but also "and just as seriously, indirect and structural forms of racial discrimination that result from the design and use of such technologies".[30]

---

[22] Amnesty International, *UK: Automated Racism - How Police Data and Algorithms Code Discrimination into Policing*, 1 February 2025, https://media.amnesty.org.uk/documents/Automated20Racism20Report20-20Amnesty20International20UK20-202025.pdf

[23] Karin de Vries, "Is it time for a European Convention against Racial Profiling?" (previously cited).
Amnesty International, *Denmark: Coded Injustice* (previously cited); Amnesty International, *Netherlands: Xenophobic Machines: Discrimination Through Unregulated Use of Algorithms in the Dutch Childcare Benefits Scandal* (Index: EUR 35/4686/2021), 2021, https://www.amnesty.org/en/documents/eur35/4686/2021/en/

[24] Amnesty International, *The Digital Border: Migration, Technology and Inequality* (Index: POL 40/7772/2024), 21 May 2024, https://www.amnesty.org/en/documents/pol40/7772/2024/en/

[25] Refugee Studies Centre, Automating Immigration and Asylum: The Uses of New Technologies in Migration and Asylum Governance in Europe, 23 January 2023, https://www.rsc.ox.ac.uk/files/files-1/automating-immigration-and-asylum_afar_9-1-23.pdf, pp. 20-21.

[26] Access Now, "Civil society joint statement: Europe's (digital) borders must fall," 4 December 2023, https://www.accessnow.org/press-release/joint-statement-eurodac-europes-digital-borders-must-fall/

[27] Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code, 2020; Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, 2018.

[28] Amnesty International, *Digitally Divided: Technology, Inequality and Human Rights* (Index: POL 40/7108/2023), 2023, https://www.amnesty.org/en/documents/pol40/7108/2023/en/

[29] Amnesty International, *Digitally Divided* (previously cited).

[30] UN Special Rapporteur on extreme poverty and human rights, Brief by the UN Special Rapporteur on Extreme Poverty and Human Rights as Amicus Curiae Before the District Court of the Hague on the Case of NJCM c.s./De Staat der Nederlanden (SyRI), Case C/09/550982/ HA ZA 18/388, 2019, https://www.ohchr.org/sites/default/files/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf

6. LEGAL FRAMEWORK FOR NON-DISCRIMINATION

The principles of equality and non-discrimination are among the cornerstones of IHRL. Almost all core human rights instruments contain non-discrimination provisions[31] while other specialized human rights treaties prohibit specific forms of discrimination.[32] States parties are obliged to respect, protect and promote the rights to equality and non-discrimination. Discrimination is prohibited, and it undermines the fulfilment of other human rights.[33]

The HRC defines discrimination for the purposes of Articles 2 and 26 of the International Covenant on Civil and Political Rights (ICCPR) as: "any distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms".[34]

The prohibition on racial discrimination is also a peremptory norm of customary international law (*jus cogens*), which means that it is applicable to all states regardless of whether they are parties to the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) and other international treaties that prohibit racial discrimination.[35] The ICERD defines racial discrimination as "any distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life".[36]

Article 26 of the ICCPR is an open norm in two respects. First, the list of prohibited grounds is not exhaustive (hence "or other status"), which means that cases of discrimination on grounds other than those explicitly enumerated can be brought under its scope,[37] including cases of discrimination on the grounds of social

[31] See, for example, Universal Declaration of Human Rights (UDHR), Article 7; International Covenant on Civil and Political Rights (ICCPR), Article 26; European Convention on Human Rights, Article 14; Protocol 12 to the European Convention for the Protection of Human Rights and Fundamental Freedoms, Article 1; Charter of Fundamental Rights of the European Union, Article 21; American Convention on Human Rights, Article 1; African Charter on Human and Peoples' Rights, Article 2.
[32] For example, International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), Article 1; Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), Article 1; Convention on the Rights of the Child (CRC), Article 2(1); Convention on the Rights of Persons with Disabilities (CRPD), Article 5; Inter-American Convention against All Forms of Discrimination and Intolerance; European Social Charter and Revised European Social Charter (ESC); Framework Convention for the Protection of National Minorities, Article 4.
[33] CESCR, General Comment 20: Non-discrimination in Economic, Social and Cultural Rights, 2 July 2009, UN Doc. E/C.12/GC/20, para. 2.
[34] HRC, General Comment 18: Non-Discrimination (1989), in Compilation of General Comments and General Recommendations adopted by Human Rights Treaty Bodies, 8 May 2006, UN Doc. HRI/GEN/1/Rev.8, p. 185; See also International Covenant on Economic, Social and Cultural Rights (ICESCR), Article 2; CEDAW, Article 2; Convention Relating to the Status of Refugees (Refugee Convention), Article 3; CRPD, Articles 4 and 5.
[35] See, for example, ILC, "Draft conclusions on identification and legal consequences of peremptory norms of general international law", 2022, Conclusion 23 & Annex; Barcelona Traction, Light and Power Company, Limited, Judgment, ICJ Reports 1970, p. 3.
[36] ICERD, Article 1(1).
[37] Janneke Gerards, Fundamental Rights: The European and International Dimension, 2023.

and economic status.[38] Also, cases related to social and economic issues can be addressed under the scope of Article 26 of ICCPR.[39] Such cases are also protected by Article 2(2) of the International Covenant on Economic, Social and Cultural Rights (ICESCR). The term "other status" has generally been given a wide, but not unlimited, meaning and its interpretation has not been limited to characteristics which are personal in the sense that they are innate or inherent.[40]

Secondly, Article 26 of the ICCPR is an open norm because it does not contain specific exemption clauses and instead adopts an open test of justification,[41] whereas non-discrimination provisions in EU law state that direct discrimination is forbidden with specific exemptions explicitly listed in the clauses of the directives. Therefore, both the HRC and national courts addressing discrimination under Article 26 can apply an open test of justification, which is generally similar to the one applied by the Court of Justice of the European Union (CJEU) for indirect discrimination.[42] Some elements of the right to non-discrimination are non-derogable,[43] and in relation to racial discrimination specifically, the prohibition on racial discrimination is a *jus cogens* norm from which no derogation is permitted.

6.1 DIRECT AND INDIRECT DISCRIMINATION

Direct discrimination occurs when an explicit distinction, exclusion, restriction or preference that is reasonably unjustifiable occurs in law, policy, or in the treatment between groups of people, resulting in some groups or persons experiencing less favourable or detrimental treatment based on a prohibited ground, such as race, ethnicity, colour, or national or ethnic origin.[44] A classic example is that of racial profiling, when "stop and frisk" police practices focus on individuals selected on the grounds of their race, skin colour, ethnicity or religion, or when proxies for race are used as selection criteria in a risk profiling algorithm.[45]

Indirect discrimination occurs when a law, policy or practice that appears neutral on its face nevertheless disadvantages or has a disproportionate effect on a group or individual defined by a prohibited ground, unless it is objectively justified by a legitimate aim and the means of achieving that aim are appropriate, necessary and proportionate. The definition of racial discrimination in Article 1 of the ICERD includes any distinction, exclusion, restriction or preference "which has the purpose *or effect* of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights" (emphasis added). The Committee on the Elimination of All Forms of Racial Discrimination (CERD) has clarified that the prohibition extends to indirect racial discrimination: "In seeking to determine whether an action has an effect contrary to the Convention, it will look to see whether that action has an unjustifiable disparate impact upon a group distinguished by race, colour, descent, or national or ethnic origin".[46]

Indirect discrimination is especially relevant when evaluating algorithmic risk profiling under discrimination prohibitions, because risk factors often correlate with prohibited grounds such as race, ethnicity or national origin – even if those characteristics are not used explicitly.[47] These correlations do not need to be strong or apparent, and can regard the risk profile as a whole instead of single risk criteria – such as classic examples involving strong correlations between race and postal code, so-called *proxy* variables. Individual variables can each be found not to correlate strongly with prohibited grounds, but the risk profile taken in its entirety

[38] See, for example, Amnesty International, *Profiled Without Protection* (previously cited).
[39] For example, HRC, *Broek v. The Netherlands*, 9 April 1987, UN Doc. CCPR/C/29/D/172/1984.
[40] In its General Comment 20 on non-discrimination, the CESCR recognizes that "[a] flexible approach to the ground of 'other status' is thus needed in order to capture other forms of differential treatment" and goes on to list the various other prohibited grounds which have been recognized, including "economic and social situation". See also, for example, ECtHR, *Carson and Others v. the United Kingdom*, Application 42184/05, Grand Chamber, 2010, para. 70; ECtHR, *Kiyutin v. Russia*, Application 2700/10, 2011, para. 56; ECtHR, *Clift v. the United Kingdom*, Application 7205/07, 2010, para. 56. See also Council of Europe, "Guide on Article 14 of the ECHR and on Article 1 of Protocol No. 12 to the Convention", updated on 31 August 2025, https://ks.echr.coe.int/documents/d/echr-ks/guide_art_14_art_1_protocol_12_eng
[41] Article 26 of the ICCPR does not contain such an explicit clause on the allowable restrictions to the guarantee of non-discrimination.
[42] Janneke Gerards, Fundamental Rights (previously cited).
[43] See, for example, HRC, General Comment 29: Article 4: Derogations During a State of Emergency, 31 August 2001, UN Doc. CCPR/C/21/Rev.1/Add.11, para. 8.
[44] See, for example, CESCR, General Comment 20: Article 2(2): Non-discrimination in Economic, Social and Cultural Rights, 2 July 2009, UN Doc. E/C.12/GC/20, para. 10(a).
[45] See, for example, Amnesty International, Netherlands: Xenophobic Machines (previously cited).
[46] CERD, General Recommendation 14: Definition of Racial Discrimination, 1993, U.N. Doc. A/48/18, para. 2. In the context of criminal justice, see: CERD, General Recommendation 31: Prevention of Racial Discrimination in the Administration and Functioning of the Criminal Justice System, 2005, UN Doc. A/60/18, pp. 98-108, paras 4-5. In the context of algorithmic risk profiling by law enforcement, see: CERD, General Recommendation 36: Preventing and Combating Racial Profiling by Law Enforcement Officials, 17 December 2020, UN Doc. CERD/C/GC/36, para. 32.
[47] Philipp Hacker, "Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law", August 2018, Common Market Law Review, Volume 55, Issue 4, https://kluwerlawonline.com/journalarticle/Common+Market+Law+Review/55.4/COLA2018095

can still have indirect effects on marginalized groups. Such a difference in treatment is presumptively considered discriminatory unless authorities provide a reasonable and objective justification.

6.2 REASONABLE AND OBJECTIVE JUSTIFICATION

International human rights bodies have clearly established that not all cases of differential treatment based on a prohibited ground constitute discrimination. Once it is established that there has been a difference in treatment of two similar cases, this difference in treatment presumptively constitutes discrimination unless it can be shown that the treatment has reasonable and objective justification.

The HRC, in its General Comment on non-discrimination, has stressed that, for the purposes of interpreting the ICCPR, "not every differentiation of treatment will constitute discrimination, if the criteria for such differentiation are reasonable and objective and if the aim is to achieve a purpose which is legitimate under the Covenant".[48]

The criteria for establishing whether there exists a reasonable and objective justification for the difference in treatment have been most clearly articulated in the jurisprudence of the ECtHR.[49] According to the ECtHR, a difference in treatment is discriminatory "if it has no objective and reasonable justification, that is, if it does not pursue a legitimate aim or if there is not a reasonable relationship of proportionality between the means employed and the aim sought to be achieved".[50] The test of the reasonable and objective justification is at least two-fold. The first part concerns the existence of a legitimate aim, such as ensuring the observance of the law by residents. The aim pursued must have a sound legal basis and it must be legitimate; that is, the aim itself may not be unlawful or discriminatory. For the second part of the test, it must be shown that the restriction of the right must be (i) **appropriate**, that is, the policy is a suitable and effective means of achieving the intended aim; (ii) **necessary** for achieving the aim, meaning that there are no other, less discriminatory policies or practices that could achieve the same aim; and (iii) **proportionate** to the aim pursued, such that the significance of the aim pursued outweighs the disadvantage suffered by the targets of discrimination and their wider community (so-called proportionality *stricto sensu*). Therefore, the test is generally referred to as three-fold (legitimate aim, necessity and proportionality).

Most international human rights bodies have implicitly or explicitly adopted the test formulated by the ECtHR.[51] For example, the Inter-American Court of Human Rights (IACHR) refers to the need for "an objective and reasonable justification"[52], and has a practice of referencing the ECtHR with regard to the recognition of indirect discrimination.[53] The African Court on Human and People's Rights cited the ECtHR and IACHR extensively when establishing its approach to the justification analysis,[54] including the formula "objective and reasonable justification"[55] that the ECtHR developed in the *Belgian Linguistics* case.[56]

While the assessment of the reasonable and objective justification is dependent on the specific details of the case at hand, international and regional human rights bodies and courts have consistently rejected certain categories of reasons as insufficient to justify differential treatment on a prohibited ground,[57] such as mere

[48] HRC, General Comment 18 (previously cited).
[49] Daniel Moeckli, "Equality and non-discrimination", in Stephanie Farrior (editor) Equality and Non-Discrimination under International Law, The Library of Essays on International Human Rights, Volume 2, 2015, p. 63.
[50] See, for example, ECtHR, *Biao v. Denmark*, Application 38590/10, 24 May 2016, para. 90.
[51] ECtHR, *Biao v. Denmark* (previously cited), para. 90.
[52] For example, *YATAMA v. Nicaragua*, Series C No. 127, (n 25) para. 185; *Norín Catrimán and others v. Chile*, Serie C No. 279, (n 35); *Duque v. Colombia*, Series C No. 310, (n 13) para. 124; *Flor Freire v. Ecuador*, Series C No. 315, (n 13) para. 125; *López Soto and others v. Venezuela*, Series C No. 362, judgment of 26 September 2018, para. 231; *Jenkins v. Argentina*, Series C No. 397, judgment of 26 November 2019, para. 91; *Almeida v. Argentina*, Series C No. 416, judgment of 17 November 2020, para. 185. In *Norín Catrimán and others v. Chile*, the Court explained that the lack of objective and reasonable justification refers to the lack of a legitimate purpose and a reasonable relationship of proportionality. See, for a discussion, Tainá Garcia Maia, "Inter-American Court of Human Rights", in Equality's Guardians, Niels Petersen (editor), 2025, https://doi.org/10.1093/9780198961109.003.0010
[53] For example, *Artavia Murillo et al. ("In vitro fertilization") v. Costa Rica* (n 15), para. 286; Advisory Opinion, "Right to freedom of association, right to collective bargaining and right to strike, and their relation to other rights, with a gender perspective", 5 May 2021, OC-27/21, para. 171. See, for a discussion, Tainá Garcia Maia, "Inter-American Court of Human Rights" (previously cited).
[54] *Tanganyika Law Society and the Legal and Human Rights Centre v. Tanzania* (n 4) paras 106.2–107.1.
[55] *African Commission on Human and Peoples' Rights v. Kenya* (n 7) para. 139.
[56] ECtHR, Case "Relating to certain aspects of the laws on the use of languages in education in Belgium", Applications 1474/62, 1677/62, 1691/62, 1769/63, 1994/63, 2126/64, 1968, https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-57525%22]}
[57] ECtHR, Case "Relating to certain aspects of the laws on the use of languages in education in Belgium" (previously cited), p. 64.

administrative inconvenience,[58] existence of a long-standing tradition,[59] prevailing views in society,[60] stereotypes,[61] or convictions of the local population.[62]

6.3 SUSPECT GROUNDS AND THE 'VERY WEIGHTY REASONS' TEST

The strictness and rigour applied by a court in judging whether a difference in treatment had a reasonable and objective justification depends on a number of case-dependent factors, the most important one being the grounds of the differentiation. Discriminatory grounds can include any personal quality, characteristic or circumstances upon which arbitrary distinctions are made. Non-discrimination provisions often explicitly enumerate certain grounds such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or "other status". These grounds are seen as "suspect" because they are immutable personal characteristics, irrelevant for performing in society, and/or go hand in hand with historical or social discrimination and stigmatization.[63] For a differential treatment, whether direct or indirect, based on a "suspect ground", the ECtHR has developed an additional test: there can be an objective and reasonable justification only if there are "very weighty reasons" for the differential treatment.[64]

When a ground is "suspect", there is a presumption that the difference in treatment cannot be justified. In these cases, the court will carry out an intensive evaluation and set very high standards for the justification of the interference. When the ground of differentiation is race or ethnicity, the ECtHR has also established that the "very weighty reasons" test must be interpreted "as strictly as possible".[65] In practice, these requirements will not easily be met. The test is thus "*strict in theory, fatal in practice*": when seemingly neutral criteria result in a differential impact on groups based on race or ethnicity, this almost automatically results in a finding of discrimination.[66]

Other regional human rights bodies have employed the European Court's "very weighty reasons" test,[67] and apply an additional level of scrutiny when judging on differential treatment based on suspect grounds. The Inter-American Commission on Human Rights has recommended ensuring that racial profiling and other explicit or implicit discrimination on the basis of race, ethnicity, colour, national origin and other prohibited grounds is explicitly prohibited and punished.[68] Until now, when the IACHR has tested differential treatment based on suspect grounds, it has always resulted in the finding of a violation.[69] Likewise, an analysis of jurisprudence of the African Court on Human and People's Rights revealed that 100% of cases of discrimination based on suspect grounds brought before the Court have resulted in the finding of a violation.[70]

---

[58] HRC, *Gueye et al. v. France*, 1985, UN Doc. CCPR/C/35/D/196/1985, https://juris.ohchr.org/casedetails/676/en-US
[59] HRC, *Muller and Engelhard v. Namibia*, 26 March 2002, UN Doc. CCPR/C/74/D/919/2000, para. 6.8.
[60] HRC, *Broeks v. The Netherlands*, 9 April 1987, UN Doc. CCPR/C/29/D/172/1984.
[61] ECtHR, *Konstantin Markin v. Russia*, Application 30078/06, 2012.
[62] ECtHR, *Inze v. Austria*, Application 8695/7, 1988.
[63] Janneke Gerards, "The margin of appreciation doctrine, the very weighty reasons test and grounds of discrimination", 2017, SSRN Electronic Journal, https://www.ssrn.com/abstract=2875230
[64] ECtHR, *Biao v. Denmark* (previously cited), para. 114.
[65] ECtHR, *D.H. and Others v. the Czech Republic*, Application 57325/00, 13 November 2007, para. 196; see also ECtHR, *Bakirdzi and E.C. v. Hungary*, Applications 49636/14, 65678/14, 10 November 2022, para. 50.
[66] Janneke Gerards, "The margin of appreciation doctrine, the very weighty reasons test and grounds of discrimination" (previously cited); Janneke Gerards, General Principles of the European Convention on Human Rights, 2023; Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005.
[67] *Maria Eugenia Morales de Sierra v. Guatemala*, Case 11.625, IACommHR Report No 4/01, 19 January 2001; see also *Dagmar Schiek and others*, Cases, Materials and Text on National, Supranational and International Non-Discrimination Law: Ius Commune Casebooks for the Common Law of Europe, 2007.
[68] Inter-American Commission on Human Rights, Police Violence Against Afro-Descendants in the United States, recommendation 9, 2018, https://www.oas.org/en/iachr/reports/pdfs/PoliceUseOfForceAfrosUSA.pdf
[69] See, for example, *I.V. v. Bolivia*, Serie C No. 329, (n 18) para. 241; *Atala Riffo and Daughters v. Chile*, Series C No. 239, (n 13) para. 124. The enhanced level of scrutiny means that the impugned measure must (i) pursue an imperative social need; (ii) respond to real and proven risks; (iii) be proportionate. See, for a discussion, Tainá Garcia Maia, "Inter-American Court of Human Rights" (previously cited), pp. 145-146.
[70] Helina Stiphanos Teka, "African Court on Human and Peoples' Rights", 2025, Equality's Guardians, Niels Petersen (editor), 2025, https://doi.org/10.1093/9780198961109.003.0010, p. 243.

6.4 INTERSECTIONAL DISCRIMINATION

Different forms of discrimination can overlap and interact with each other to produce a unique and compounded experience of marginalization or oppression for an individual.[71] This is often referred to as intersectional discrimination. International law obliges states to take measures to eliminate all forms of discrimination, including the intersectional discriminatory effects of technology. Amnesty International has documented various cases of such harms across different countries, where people were discriminated against based on one or multiple intersections of their race, ethnicity, national origin, gender, disability, age and social and economic status.[72] Automated decision-making and the introduction of technology in social protection systems have been shown to create barriers for rights holders in claiming their rights or when trying to appeal decisions.[73]

THE USE OF NATIONALITY AS A RISK CRITERION MAKES A DE FACTO DISTINCTION ON THE GROUNDS OF RACE

In the Netherlands, the Ministry of Foreign Affairs assesses Schengen short-stay visa applications with a risk profiling algorithm in order to assess the "risk" of overstaying.[74] All visa applications are scored by an automated risk profile containing seven characteristics, including nationality. The risk score determines whether a more intensive assessment by a human operator is needed. Based on their nationality, among other criteria, people who are automatically attributed a higher risk score and are more likely to be selected by the Ministry for a more intensive assessment of their application. This includes – although not exclusively – people from countries with a prevalent population of people of African descent, such as Suriname, a former colony of the Netherlands.[75] This is in violation of the prohibition of discrimination as laid down in Article 26 of the ICCPR and Article 1 of the ICERD.[76] Information about nationality in the sense of citizenship can be relevant in the granting process for statutory schemes, for example to check whether someone meets the requirements for a visa or social benefits. But once states single out specific nationalities or groups of nationalities, or use a person's nationality to calculate risk in the context of migration control or other forms of law enforcement, nationality no longer refers to an eligibility criterion but rather is used as a risk criterion. The information about nationality used in the risk profile is not useful to make a distinction on the grounds of nationality intended as citizenship. The Ministry of Foreign Affairs apparently assumes a suspected or proven (statistical) connection between certain nationalities and improper visa applications or a higher likelihood of abuse. The factual distinction therefore does not refer to a distinction on the grounds of their citizenship, but to the identification of a group of persons who share certain characteristics derived from the information on their nationality. Therefore, the risk profile used by the Ministry of Foreign Affairs makes a de facto distinction on the grounds of race.

---

[71] Amnesty International, Intersectionality from a Racial Justice Perspective: Submission to the UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance (Index: IOR 40/9141/2025), 21 May 2025, https://www.amnesty.org/en/documents/ior40/9141/2025/en/
OHCHR, Promotion and Protection of the Human Rights and Fundamental Freedoms of Africans and of People of African Descent Against Excessive Use of Force and Other Human Rights Violations by Law Enforcement Officers Through Transformative Change for Racial Justice and Equality: Report, 15 July 2024, UN Doc. A/HRC/57/67, para. 3; Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance (UN Special Rapporteur on racism), Ashwini K.P, Intersectionality from a Racial Justice Perspective: Report, 29 May 2026, UN Doc. A/HRC/59/62.
[72] Amnesty International, *Netherlands: Xenophobic Machines* (previously cited); Amnesty International, *Denmark: Coded Injustice* (previously cited); Amnesty International, *Profiled Without Protection* (previously cited).
[73] Amnesty International, *Trapped by Automation: Poverty and Discrimination in Serbia's Welfare State* (Index: EUR 70/7443/2023), 2023, https://www.amnesty.org/en/latest/research/2023/12/trapped-by-automation-poverty-and-discrimination-in-serbias-welfare-state/
Amnesty International, *Denmark: Coded Injustice* (previously cited).
[74] Amnesty International, Crossings and Journeys: People of African Descent in Global Migration and the Enduring Architecture of Racialization: Submission to the 37th Session of the UN Working Group of Experts on People of African Descent (Index: IOR 40/0469/2025), 2025, https://www.amnesty.org/en/documents/ior40/0469/2025/en/
Amnesty International Netherlands, Buitenlandse Zaken gaat willens en wetens door met discrimineren [The Ministry of Foreign Affairs Knowingly Continues to Discriminate], 2 May 2024, https://www.amnesty.nl/actueel/buitenlandse-zaken-gaat-door-met-etnisch-profileren (in Dutch); Amnesty International Netherlands, Etnisch Profileren is overheidsbreed probleem [Ethnic profiling is a government-wide problem], 21 March 2024, https://www.amnesty.nl/actueel/het-kabinet-moet-burgers-beschermen-tegen-etnisch-profileren (in Dutch).
[75] NRC, Autoriteit Persoonsgegevens roept minister Hoekstra op het matje om algoritme [Data Protection Authority calls Minister Hoekstra in over algorithm], https://www.nrc.nl/nieuws/2023/05/01/minister-moet-uitleg-geven-over-algoritme-voor-visa-a4163510 (in Dutch).
[76] ICERD, Article 1(3): "Nothing in this Convention may be interpreted as affecting in any way the legal provisions of States Parties concerning nationality, citizenship or naturalization, provided that such provisions do not discriminate against any particular nationality."

6.5 POSITIVE OBLIGATIONS AND THE DUTY TO COMBAT AND PREVENT DISCRIMINATION

Under international law, the state is obligated not only to respect the right to non-discrimination but also to protect and fulfil it.[77] This is both a positive and a negative obligation. The prohibition of discrimination requires the state to ensure that laws and regulations, circulars, guidelines or policies do not permit or result in discrimination. There should also be no discrimination in the policies and practices of public institutions and civil servants when implementing legislation, providing services or fulfilling other functions.

6.5.1 THE OBLIGATION TO COMBAT RACIAL PROFILING

The obligation to take measures against racial profiling follows, among other sources, from the ICERD, which has been in force since 1969. Every state party to that convention undertakes the duty to "pursue by all appropriate means and without delay a policy of eliminating racial discrimination in all its forms" and "to ensure that all public authorities and public institutions, national and local, shall act in conformity with this obligation".[78] The obligation also includes that each state adopt effective measures to review government policies and to amend, repeal or nullify laws and regulations that may lead to or perpetuate racial discrimination, including racial profiling.[79] This also encompasses the duty to counteract the potentially indirect discriminatory effects of legislation.[80] ICERD further requires the provision of effective protection and remedies against racial discrimination[81] – including when racial profiling is a consequence of (automated) risk profiling. The CERD also recommends that states ensure that all instances of algorithmic bias are duly investigated and sanctioned.[82]

When addressing racial profiling, regional human rights courts such as the ECtHR have repeatedly emphasized that racial discrimination requires states to exercise "special vigilance and vigorous reaction" because of its perilous consequences.[83] The IACHR has addressed racial profiling on multiple occasions.[84] In a 2020 case involving the arbitrary arrest and death of José Delf Acosta Martinez, the police officers justified his arrest on the basis of his alleged state of drunkenness, obscuring the use of racial profiling as the main reason for the detention.[85] In 2008, when it decided a case concerning killings by Brazilian police, the IACHR recommended that Brazil take measures to avoid racial discrimination in law enforcement and criminal justice proceedings. It referenced jurisprudence of the ECtHR when reiterating positive obligations,[86] including "by establishing, for such purposes, distinctions based on de facto inequities for the protection of those who must be protected".[87] The IACHR defined racial profiling as a

"repressive tactic… adopted for supposed reasons of public safety and protection… motivated by stereotypes based on race, colour, ethnicity, language, descent, religion, nationality, place of birth, or a combination of these factors, rather than on objective suspicions, and it tends to single out individuals or groups in a discriminatory way based on the erroneous assumption that people with such characteristics are prone to engage in specific types of crime".[88]

---

[77] Amnesty International, *Dealing with Difference: A Framework to Combat Discrimination in Europe* (Index: EUR 01/003/2009), 1 July 2009, pp. 31-52.
[78] ICERD, Article 2(1).
[79] ICERD, Articles 2 and 1(c); CERD, General Recommendation 36 (previously cited), para. 23.
[80] CERD, General Recommendation 31 (previously cited), para. 4(b)
[81] ICERD, Article 6; CERD, General Recommendation 36 (previously cited), para. 24.
[82] CERD, General Recommendation 36 (previously cited), para. 65.
[83] See ECtHR, *Timishev v. Russia*, Applications 55762/00, 55974/00, 13 December 2005, para. 56. See also ECtHR, *D.H. and others v. Czech Republic* (previously cited), para. 176; *ECtHR, Basu v. Duitsland*, Application 215/19, 18 October 2022, para. 24; ECtHR, *Wa Baile v. Switzerland*, Applications 43868/18, 25883/21, para. 90.
[84] For example, Inter-American Court of Human Rights, *Acosta Martínez et al. v. Argentina* (previously cited); Inter-American Court of Human Rights, *Fernández Prieto and Tumbeiro v. Argentina*, 2020, https://corteidh.or.cr/docs/casos/articulos/seriec_411_ing.pdf
IACHR, Police Violence Against Afro-Descendants in the United States (previously cited).
OAS, "IACHR calls on the states of the region to eliminate all forms of racial discrimination, promote cultural change and adopt comprehensive reparation measures for people of afro-descendant", 12 September 2020, https://www.oas.org/en/iachr/media_center/PReleases/2020/216.asp
[85] Inter-American Court of Human Rights, *Acosta Martínez et al. v. Argentina* (previously cited), para. 100.
[86] In particular, *Nachova et al v. Bulgaria*, Applications 43577/98, 43579/98, 6 July 2005, in "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits", paras 139-140, https://cidh.oas.org/annualrep/2009eng/Brazil12440eng.htm
[87] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 145.
[88] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 143.

The Court also referred to the often disadvantaged socio-economic position of racialized people in Brazil,[89] and admonished the state for having "no respect for the special situation of belonging to a group that is considered vulnerable (of African descent, poor, living in a favela)".[90]

In February 2024, the ECtHR found Switzerland guilty of the racial profiling of Mohamed Wa Baile, a Black Swiss citizen who was stopped for an identity check by police while on his way to work.[91] The Court reiterated that authorities must use all available means to combat racism and thereby reinforce the understanding that diversity is not a threat but an enrichment of democratic society.[92] The ECtHR notes that the state is obliged to take measures, in particular to secure the rights and freedoms of minorities.[93] Legal and administrative frameworks must be adequate to fulfil treaty obligations.[94] Specifically, national legislation regulating police action must provide a system of adequate and effective safeguards against arbitrariness and abuse.[95] Police officers must not be left "in a vacuum" when performing their duties; therefore, the state must "establish clear legal and administrative frameworks with limiting conditions".[96] The ECtHR further referred to CERD Recommendation 36 and reiterated the obligation to take positive measures to eliminate racial profiling.[97] It also referred to the recommendation of the European Commission against Racism and Intolerance to introduce the standard of reasonable suspicion in order to avoid discriminatory identity checks.[98]

6.5.2 POSITIVE OBLIGATIONS AND DIGITAL TECHNOLOGY

The ICESCR obligates states to guarantee non-discrimination in the exercise of each of the economic, social and cultural rights enshrined in the Covenant.[99] The CESCR draws specific attention to systemic discrimination, as discrimination against some groups is often "pervasive and persistent and deeply entrenched in social behaviour and organization, often involving unchallenged or indirect discrimination".[100] The UN Special Rapporteur on contemporary forms of racism notes that states "must address not only explicit racism and intolerance in the use and design of emerging digital technologies" but also, "just as seriously, indirect and structural forms of racial discrimination that result from the design and use of such technologies".[101] The UN Special Rapporteur on extreme poverty and human rights, in the context of the "digital welfare state", has stated that:

"Predictive analytics, algorithms and other forms of AI are highly likely to reproduce and exacerbate biases reflected in existing data and policies. In-built forms of discrimination can fatally undermine the right to social protection for key groups and individuals. There therefore needs to be a concerted effort to identify and counteract such biases in designing the digital welfare state. This, in turn, requires transparency, and broad-based inputs into policy-making processes. The public, and especially those directly affected by the welfare system, need to be able to understand and evaluate the policies that are buried deep within the algorithms".[102]

[89] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 63.
[90] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 150.
[91] ECtHR, *Wa Baile v. Switzerland* (previously cited); see also Amnesty International, Switzerland: Mohamed Wa Baile Wins Ethnic Profiling Case as the European Court of Human Rights Unanimously Condemns Racial Discrimination (Index: EUR 43/7774/2024), 2024, https://www.amnesty.org/en/documents/eur43/7774/2024/en/
K. de Vries, "The ECtHR advances the battle against racial profiling in Wa Baile c. Suisse", Verfassungsblog, 6 March 2024.
[92] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 90.
[93] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 124.
[94] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 125.
[95] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 126.
[96] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 126.
[97] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 127.
[98] ECtHR, *Wa Baile v. Switzerland* (previously cited), para. 129.
[99] ICESCR, Article 2(2); CESCR, General Comment 20 (previously cited), para. 7.
[100] CESCR, General Comment 20 (previously cited), para. 12.
[101] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis", 18 June 2020, UN Doc. A/HRC/44/57, para. 48.
[102] UN Special Rapporteur on extreme poverty and human rights, Report, 11 October 2019, UN Doc. A/74/493, para. 82.

7. THE HUMAN TOLL OF RISK PROFILING: A DECADE OF HARMFUL STATE PRACTICES

This Chapter lists more than a decade of previous research by Amnesty International and other CSOs exposing the harms of risk profiling systems across multiple policy domains, in the public sectors of various countries. This body of research illustrates how the use of risk profiling systems across a wide range of contexts has consistently resulted in human rights violations. These cases were selected because their prediction objectives and methodologies closely match the definition of risk profiling used in this report.

7.1 AUSTRALIA

This case concerns Australia's Robodebt scheme, an automated debt-recovery system that used data from Centrelink's Online Services to trigger rule-based validations of recipients' reported employment income. Centrelink's Online Services refer to online interfaces that let Australians view and manage their governmental accounts to claim and track payments, report income, upload documents, update personal details, and communicate with public services.

In Australia, a royal commission of inquiry found in 2023 that about 400,000 social security claimants were wrongly accused by an automated fraud detection algorithm of having misreported their income and were punished with fines.[103] The design of this system was informed by "an enduring assumption that all persons on welfare or pension payments are potential or actual cheats."[104] Stigma surrounding social security payments could be "so deep-seated that it discourages eligible people from seeking support, even in the face of severe economic and personal hardship".[105] Fraud cases made up a minuscule proportion of payments, approximately 0.1% of the total.[106] The people targeted reported financial and psychological harm including distress, trauma, anxiety and mental ill-health, with at least two documented cases of death by suicide.[107]

While the scheme did not use AI, it did use automation. It involved a system of "extremely rigid" business rules with no ability to move outside of specific and defined action on the basis of the data received.[108] If the

[103] Seb Starcevic, "Australian Robodebt scandal shows the risk of rule by algorithm", 2022, Context by Thomson Reuters Foundation, https://www.context.news/surveillance/australian-robodebt-scandal-shows-the-risk-of-rule-by-algorithm (accessed on 6 March 2026).
[104] Royal Commission into the Robodebt Scheme, Report, 7 July 2023, https://robodebt.royalcommission.gov.au/publications/report, p. 330.
[105] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 330.
[106] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 330.
[107] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 184; British Broadcasting Corporation, "Robodebt: Illegal Australian welfare hunt drove people to despair", 7 July 2023, https://www.bbc.com/news/world-australia-66130105
[108] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 472.

recipient updated their employment data in Centrelink's Online Services, the system applied risk rules to validate the updated data.[109] Automated validation risk rules were applied to any new information provided by the recipient. There was no intervention by a compliance officer.[110]

A human rights analysis by an academic scholar found that Robodebt "raised multiple human rights concerns including in relation to the rights to privacy, freedom of movement, social welfare and an effective remedy. Robodebt demonstrated a disregard for the dignity of welfare recipients and indiscriminately issued erroneous debts largely to individuals who were vulnerable due to their disadvantaged socioeconomic status."[111]

7.2 CHILE

In Chile, the digital rights CSO Derechos Digitales investigated in 2022 the "Child Alert System", a computer system developed for Chile's Ministry of Social Development and Family.[112] Its objective is to estimate and predict the level of risk to children and adolescents of suffering a violation of their rights in the future. In practice, the system generates a "risk index" score for each child/adolescent, allowing cases to be prioritized by the Municipal Offices for Children. In addition, the system has been set up as a platform for registering, managing and monitoring the cases of children and adolescents identified as being at greatest risk. The research highlighted a lack of attention toward the possible performative role of the system when the state contacts families who have not requested government assistance and who could experience the contact as highly invasive of their privacy. A lack of public documentation on the design and performance of the predictive model is described as "worrying", as well as the absence of processes for citizen participation or consultation. Finally, the report evidences the risk of socio-economic discrimination because of the composition of the targeted population.

7.3 COLOMBIA

In Colombia, the government rolled out an algorithm into a system called "Identification System of Potential Beneficiaries of Social Programmes (SISBEN)" that individually rates the Colombian population in terms of prosperity in order to find individuals living in poverty to assist them with social security.[113] The fourth version, launched in 2021, first integrated interoperability elements, with data exchange and cross-checking of information between public and private databases. Algorithms were introduced as far back as 2002. Specialized software is used to generate the individual score, and the resulting score is used to establish cut-off points to determine which person can request a specific social security payment. In its fourth version, the government began to use data analytics technologies "to search for inconsistencies in the database, to punish people who have allegedly lied, and to reduce the number of people who could have access to benefits."[114] This includes risk profiles.[115] These automated fraud prediction mechanisms "may inadvertently stigmatize individuals living in poverty, deny them due process, and impose undue administrative burdens or disproportionate sanctions".[116] Human rights researchers warned about possible issues with differential treatment, discrimination and due process.[117]

Research by Fundación Karisma on Families in Action, a cash transfer system for people living in poverty, has highlighted that, as a result of the system's design, "an intricate network of monitoring and data utilization ultimately curtails the independence of disenfranchised women by enforcing gender stereotypes,

[109] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 473.
[110] Royal Commission into the Robodebt Scheme, Report (previously cited), p. 475.
[111] Saabiq Chowdhury, "Technology is never neutral: Robodebt and a human rights analysis of automated decision-making on welfare recipients", January 2024, Australian Journal of Human Rights, Volume 30, Issue 1, https://doi.org/10.1080/1323238X.2024.2409620, p. 39.
[112] Derechos Digitales América Latina and Matías Valderrama, "Chile: The Child Alert System and predicting the risk of violations of children's rights", 2022, https://www.derechosdigitales.org/wp-content/uploads/02_Informe-Chile-EN_180222.pdf
[113] Joan Lopez-Solano and Juan Diago Castañeda, "Automation, digital technologies and social justice: experimenting with poverty in Colombia", in Maximiliano Campos Ríos (editor), Artificial intelligence in Latin America and the Caribbean: Ethics, governance and policies, 2020.
[114] Joan Lopez-Solano and Juan Diago Castañeda, "Automation, digital technologies and social justice: experimenting with poverty in Colombia" (previously cited), p. 1; Victoria Adelmant, "Global perspectives on automated welfare: comparative considerations for assessing impacts", 2025, SSRN Electronic Journal, https://papers.ssrn.com/abstract=5205451 (accessed 5 March 2026).
[115] Joan Lopez-Solano and Juan Diago Castañeda, "Automation, digital technologies and social justice: experimenting with poverty in Colombia" (previously cited), p. 246.
[116] Sebastian Smart, Algorithmic Discrimination in Latin American Welfare States, 2024, p. 16.
[117] Sebastian Smart, Algorithmic Discrimination in Latin American Welfare States, 2024, p. 17.

along with the independence of those women who are living in poverty and reliant on social security payments".[118]

7.4 DENMARK

In Denmark, Amnesty International has been conducting research into, and monitoring the use of, profiling systems by law enforcement and social security agencies since 2019. In 2023, Amnesty International published *Coded Injustice*, a long-form research investigation into the use of technology within Denmark's social benefits system, administered by the public authority Udbetaling Danmark (UDK, or Pay Out Denmark) and the company Arbejdsmarkedets Tillægspension (ATP).[119]

In 2012, the Danish government established UDK to centralize the payment of social security payments overseen by municipalities, including child allowances, pension benefits, housing benefits, unemployment benefits, maternity pay and sick pay benefits under the social security state system. UDK/ATP established a Joint Data Unit tasked with developing data-driven fraud detection algorithms with the purported aim of identifying fraudulent social benefit applications for further investigations, in collaboration with private companies. The Joint Data Unit links or merges the personal data of millions of Danish residents from registers (public databases) that contain information about benefit recipients and their family members or other household members. This information includes, but is not limited to, their residency and residency moves, citizenship, place of birth, family relationships and circumstances, housing arrangements and building conditions, employment, income, tax, health and education (see also box in section 8.3). UDK uses risk-profiling models (which analyse aspects of an individual's personality, behaviours, interests and habits to make predictions or decisions about them) to analyse this data in order to identify persons who supposedly have an increased risk of receiving benefits fraudulently, and flag them for further investigations. Amnesty International's research found UDK and ATP's use of algorithms to detect fraud in the distribution of social benefits negatively affects the human rights of social security benefits recipients, including their rights to privacy, equality and non-discrimination, dignity, social security and remedy.

7.5 FRANCE

In October 2024, Amnesty International and 14 other coalition partners led by La Quadrature du Net (LQDN) submitted a complaint to the Council of State, the highest administrative court in France, demanding that a risk-profiling algorithmic system used by the French Social Security Agency be stopped.[120] Ten additional organizations joined the complaint in 2025.[121]

The risk-scoring algorithm used by the French Social Security Agency's National Family Allowance Fund (CNAF), which is used to detect overpayments and errors regarding benefit payments, was discriminating against marginalized groups, violating people's right to equality and non-discrimination. In 2023, LQDN gained access to versions of the algorithm's source code – a set of instructions written by programmers to create a piece of software – and exposed the discriminatory nature of the system. Since 2010, CNAF has used a risk-scoring algorithm to identify people who are potentially committing benefits fraud by receiving overpayments. The algorithm assigns a risk score between zero and one to all recipients of family and housing benefits. The closer the score is to one, the higher the probability of being flagged for investigation. Overall, there are 32 million people in France living in households that receive a benefit from CNAF. Their sensitive personal data, as well as that of their family, is processed periodically, and a risk score is assigned. The criteria that increase one's risk score include parameters which discriminate against households of marginalized people, including being on a low income, being unemployed, living in an underserved neighbourhood, spending a significant portion of income on rent, and working while having a disability. The details of those who are flagged due to having a high-risk score are compiled into a list that is investigated further by a fraud investigator.

---

[118] Fundación Karisma, Vigilando a las Buenas Madres [Watching the Good Mothers], 2021, https://blog.karisma.org.co/vigilando-a-las-buenas-madres/ (in Spanish).
See also Sebastian Smart, Algorithmic Discrimination in Latin American Welfare States, 2024.
[119] Amnesty International, *Denmark: Coded Injustice* (previously cited).
[120] Amnesty International, *France: CNAF State Council Complaint* (Index: EUR 21/8795/2024), 27 November 2024, https://www.amnesty.org/en/documents/eur21/8795/2024/en/
[121] LQDN, "CNAF's discriminatory scoring algorithm: 10 new organisations join the case before the Conseil d'État in France", 20 January 2026, https://www.laquadrature.net/en/2026/01/20/cnafs-discriminatory-scoring-algorithm-10-new-organisations-join-the-case-before-the-conseil-detat-in-france/

On 15 January 2026, CNAF released the source code of its current algorithm. The coalition welcomed the efforts towards transparency but stressed that this was insufficient.[122] A 2025 internal CNAF study obtained by the claimants recognized the algorithm's discriminatory effects. This study was included in a new brief sent to the court in December 2025.[123]

7.6 NETHERLANDS

In the Netherlands, Amnesty International has been researching and monitoring the use of risk profiling by authorities in policing and in the detection of fraud in social protection since 2013. The first report focused on the use of ethnic profiling in the use of stop and search powers. However, in the years that followed, multiple scandals about algorithmic risk profiling in the context of social protection, policing and border control by a range of government agencies came to light in the Netherlands. The use of automated or algorithmic risk profiling systems emerged as a new harmful practice. This led Amnesty International to conclude in 2024 that discriminatory profiling is a structural and government-wide problem in the Netherlands.[124]

RACIAL PROFILING BY STREET-LEVEL POLICE OFFICERS

In October 2013, the Dutch section of Amnesty International published a report on ethnic profiling in the Netherlands: *Stop and Search Powers Pose a Risk to Human Rights: Acknowledging and Tackling Ethnic Profiling in the Netherlands*.[125] The report focused mainly on ethnic profiling in the use of stop and search powers: that is, the exercise of powers arising from the police's general supervisory powers, in which members of the public may be stopped or checked without being suspected of any criminal offence. This report showed that the practice of ethnic profiling in the Netherlands went beyond the level of isolated incidents.

WE SENSE TROUBLE: DISCRIMINATORY MASS SURVEILLANCE

In 2020, Amnesty International published the report *We Sense Trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands*.[126] This report documented the dangers of emerging "predictive policing" projects that were being rolled out by law enforcement agencies across the Netherlands. The projects – branded "living labs" by Dutch police – use mathematical models to assess the risk that a crime will be committed by a certain person or at a certain location, with law enforcement efforts then directed towards those individuals or locations deemed to be "high risk". Amnesty International investigated one such project in the city of Roermond, called the Sensing Project. This policing experiment treated people in Roermond as "guinea pigs" under mass surveillance and discriminated against people with Eastern European nationalities.

XENOPHOBIC MACHINES: CHILDCARE BENEFITS SCANDAL

During the 2010s, concerns about widespread fraud in the claiming of childcare benefits committed by parents and caregivers led the Dutch government to introduce increasingly harsh and highly automated enforcement policies and practices. What followed was a "tough on crime" approach with harsh treatment for many parents and caregivers – even in cases where they had done nothing wrong, had only made minor errors, or had merely made an administrative omission. In 2021, Amnesty International published the report *Xenophobic Machines* about the algorithmic decision-making system for fraud detection that was introduced in 2013 by the tax authorities.[127] Because the risk profiling system included the criterion "non-Dutch nationality", people from racialized groups had an increased chance of being categorized as possibly committing fraud and, therefore, of being selected for an investigation. The system also had a self-learning element that caused it to focus on lower-income households. The report concluded that the risk classification model was a form of direct discrimination on the grounds of race. Amnesty International's analysis also showed that single parents and caregivers – and their families – in lower-income households were particularly affected. The scandal led to the fall of the Dutch government and the resignation of the

[122] LQDN, "CNAF's discriminatory scoring algorithm" (previously cited).
[123] LQDN and others, Mémoire en Réplique No. 498440 [Reply No. 498440], December 2025 https://www.laquadrature.net/wp-content/uploads/sites/8/2026/01/06_LQDN_Cnaf_Replique.pdf (in French).
[124] Amnesty International Netherlands, *Etnisch profileren is een overheidsbreed probleem* (previously cited).
[125] Amnesty International Netherlands, *Stop and Search Powers Pose a Risk to Human Rights: Acknowledging and Tackling Ethnic Profiling in the Netherlands*, October 2013, https://www.amnesty.nl/content/uploads/2016/11/amnesty_stopandsearchpowersposearisktohumanrights.pdf?x45368
[126] Amnesty International, *Netherlands: We Sense Trouble* (previously cited).
[127] Amnesty International, *Netherlands: Xenophobic Machines* (previously cited).

entire Cabinet of Ministers in January 2021, as well as widespread harms experienced by individuals who were wrongfully targeted for investigation and forced to erroneously pay back funds to the state. (See further, box in section 8.2)

STUDENTS DISCRIMINATED AGAINST BY DUO

In June 2023, a new scandal of a discriminatory risk profiling practice emerged in the Netherlands. A collective of journalists investigated possible discriminatory effects in the way the Netherlands' Education Executive Agency (Dienst Uitvoering Onderwijs, DUO), under the responsibility of the Ministry of Education, Culture and Science, detected and handled students who were possibly abusing the out-of-home student grant.[128] A report published by Amnesty International showed how DUO used a discriminatory risk-profiling algorithm.[129] Some of the criteria for selection were on the grounds of race and socio-economic status, leading to discriminatory outcomes. (See also Chapter 4 on rule-based risk profiling.)

MIGRATION DOMAIN: DISCRIMINATORY VISA ASSESSMENTS

Currently, the Dutch Ministry of Foreign Affairs assesses Schengen short-stay visa applications with a risk profiling algorithm in order to assess the risk of overstaying.[130] This system appears to be in violation of the prohibition of discrimination, as it makes a de facto distinction on the grounds of race. (See further, box in Chapter 6.) It also affects the right to private life of both applicants and Dutch nationals whose relatives face obstacles when trying to attend family events such as weddings or funerals.

7.7 SPAIN

In Spain, Algorithm Watch and La Fede Cat researched VioGén, a comprehensive monitoring system for gender violence cases, used since 2007 to assess and predict recidivism risk in gender-based violence incidents through a 35-item questionnaire scored by algorithms. This system was shown to have "reasonable performance", with an AUC (Area Under the Curve) of between 0.65 and 0.8.[131] In theory, Spanish agents can increase the score manually if they believe there is a higher risk. But a 2014 study found that, in 95% of the cases, agents did not alter the automatic outcome.[132] An independent audit on 800,000 cases revealed significant flaws, including limited reliability, arbitrary correlations in risk factors, and insufficient human oversight. Only 3% of women were classified as being medium or high risk of being victims of repeat violence, and more than 80% reported issues with the system.[133] A leaked document from the General Council of the Judiciary from 2014 showed that 14 out of 15 women killed that year, having previously reported their aggressor had been given a score of low or non-specific risk. These deficiencies raise concerns about transparency, accountability, and the adequacy of protective measures, risking the safety of victims and perpetuating systemic biases that may also unfairly affect those accused, due to potentially flawed or discriminatory risk assessments.

7.8 SWEDEN

In Sweden, Amnesty International supported an investigation led by Lighthouse Reports and Svenska Dagbladet (SvD) of a risk-profiling algorithmic system deployed by Sweden's Social Insurance Agency.[134] The ML system has been used by the Swedish Social Insurance Agency since at least 2013. The system assigns risk scores, calculated by an algorithm, to social security applicants to detect social benefits fraud. The investigation exposed that the system disproportionately flagged certain groups for further investigation regarding social benefits fraud, including women, individuals born overseas or whose parents were born overseas, low-income earners, and individuals without university degrees. IMY, the Swedish Data Protection

[128] The research was carried out by Hoger Onderwijs Persbureau, Platform Investico, De Groene Amsterdammer, NOS op 3 and Trouw, and resulted in various publications, for example, https://www.platform-investico.nl/onderzoeken/de-discriminerende-fraudecontroles-van-duo
[129] Amnesty International, *Netherlands: We Sense Trouble* (previously cited); Amnesty International, *Netherlands: Xenophobic Machines* (previously cited); Amnesty International, *Etnisch profileren is overheidsbreed probleem* (previously cited); Amnesty International, *Profiled Without Protection* (previously cited).
[130] Amnesty International, *Crossings and Journeys* (previously cited); Amnesty International Netherlands, *Buitenlandse Zaken gaat willens en wetens door met discrimineren* (previously cited); Amnesty International Netherlands, *Etnisch Profileren is overheidsbreed probleem* (previously cited).
[131] Algorithm Watch, Automating Society Report, 2020, p. 227, https://automatingsociety.algorithmwatch.org/
[132] Algorithm Watch, Automating Society Report (previously cited), p. 225.
[133] Eticas Foundation, Can AI Solve Gender Violence? Auditing the Use of AI to Assess Risk: The Case of Viogén, 2022, https://www.eticasfoundation.org/the-case-of-viogen-can-ai-solve-gender-violence/
[134] Amnesty International, "Sweden: Authorities must discontinue discriminatory AI systems used by welfare agency", 27 November 2024, https://www.amnesty.org/en/latest/news/2024/11/sweden-authorities-must-discontinue-discriminatory-ai-systems-used-by-welfare-agency/

Authority, opened an investigation into the Swedish Social Insurance Agency's use of algorithmic systems after the publication of Lighthouse Reports and SvD's research exposed discrimination within an algorithmic system, which resulted in the system being shut down.

7.9 UNITED KINGDOM

TRAPPED IN THE MATRIX

In 2018, Amnesty International published a report on the Metropolitan Police's use of the so-called Gangs Matrix, a risk management tool for preventing serious violence.[135] The concepts of "gang" and "gang member" were vague and ill-defined, and the process for adding people's names to or removing them from the matrix lacked clear parameters, thresholds and criteria. This led to over-broad and arbitrary identification of people as gang members. Many indicators used by the Metropolitan Police conflated elements of urban youth culture with violent offending and were heavily racialized. There were no clear processes for reviewing the matrix, or for correcting or deleting outdated information. There was no formal process to notify individuals that they were included in the matrix and no official system through which they could challenge their inclusion or have their names removed. Data sharing between the police, housing associations, schools, job centres, the criminal justice system and the Home Office lacked safeguards. This created a risk that these services would discriminate against already marginalized young people.

The Gangs Matrix was an excessive interference with the right to privacy that affected the rights of Black boys and young men disproportionately. The weak data governance and lack of safeguards that characterized the database show that it was designed and put to use without sufficient regard for the rights of those whose names were listed on it. Amnesty International believes that the Gangs Matrix was unfit for purpose: it put rights at risk, and appeared both ineffective and counter-productive.

AUTOMATED RACISM

In 2025, Amnesty International released a report showing that almost three-quarters of United Kingdom (UK) police forces are using data-based and data-driven systems to attempt to predict, profile and assess the risk of crime or criminalized behaviour occurring in the future.[136] The use of such approaches is influencing decisions in policing and the criminal justice system.

The use of these so-called predictive policing tools in policing and the criminal justice system violates people's rights, including the right to a fair trial and the presumption of innocence, the right to privacy, the rights to freedom of assembly and association, and the right to equality and non-discrimination. These systems are, in effect, a modern method of racial profiling, reinforcing systemic racism and discrimination in policing. They also risk violating people's economic, social and cultural rights, such as the right to social security.

Police forces use these systems to attempt to predict and profile who will commit crime in the future or who is at "risk" of committing crime or other criminalized behaviour. Police use these so-called predictions, profiles and risk assessments to target specific people and groups with increased policing. The aim is to target certain individuals and to intervene before the predicted behaviour occurs.

Police forces also share these predictions, profiles, risk assessments and related data with other criminal justice system authorities including the Crown Prosecution Service, prison and probation services; essential public service providers such as councils, local authorities and the Department for Work and Pensions (DWP); and with unspecified third-party agencies and organizations. Predictive policing systems are contributing to racist and discriminatory policing and criminalization of locations, groups and individuals, perpetuating institutional racism in policing and society. Their use is leading to the repeated targeting of individuals from Black, racialized and other marginalized groups.

The use of risk profiling targets individuals or communities exercising their rights to freedom of association and peaceful assembly. In the UK, people seek to avoid areas known to be targeted by the police, resulting in a chilling effect. In this example, predictive policing amounts to mass surveillance, which decreases people's ability and willingness to exercise their right to freedom of association.[137]

---

[135] Amnesty International UK, *Trapped in the Matrix: Secrecy, Stigma, and Bias in the Met Gangs Database*, May 2018, https://media.amnesty.org.uk/documents/Trapped20in20the20Matrix20Amnesty20report.pdf
[136] Amnesty International, *UK: Automated Racism* (previously cited).
[137] Amnesty International, *UK: Automated Racism* (previously cited).

TOO MUCH TECHNOLOGY, NOT ENOUGH EMPATHY

In 2025, Amnesty International found that the introduction of digital technologies into the UK's flawed and inadequate social security system has, in many cases, led to further hardship for social security claimants. This has negatively affected the realization of claimants' human rights, including their rights to social security and an adequate standard of living.[138] Concerns have been raised over the use of AI, ML tools and algorithmic decision-making by a number of UK parliamentary committees, CSOs, UN Special Rapporteurs such as the UN Special Rapporteur on extreme poverty and human rights, as well as by several individuals who were interviewed by Amnesty International as part of this research. The DWP has committed millions of pounds for projects that use AI, including for advanced analytics in fraud and error detection. The operations of these systems are almost entirely hidden from public scrutiny. The very limited information available is often obtained through Freedom of Information requests. This allows for minimal oversight and analysis of the potentially discriminatory or otherwise human rights-harming impacts of these systems, as well as of the types of data that underpin these systems to establish whether they comply with data protection and human rights standards.

The DWP's Personal Information Charter states that it uses profiling in a range of ways including "to detect and prevent fraud and error".[139] Furthermore, the DWP describes profiling as involving "the use of personal data to evaluate certain personal aspects such as a person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."[140] The DWP does not justify why it needs to gather information on personal preferences, interests, "reliability" or other metrics to meet its obligations to social security claimants.

[138] Amnesty International, *"Too Much Technology, Not Enough Empathy": How the UK's Push to Digitalize Social Security Harms Human Rights* (Index: EUR 45/9478/2025), 2025, https://www.amnesty.org/en/documents/eur45/9478/2025/en/
[139] DWP, "Personal Information Charter", https://www.gov.uk/government/organisations/department-for-work-pensions/about/personal-information-charter, accessed on 22/05/2026.
[140] DWP, "Personal Information Charter", cited previously.

8. THE FINE LINE BETWEEN RISK PROFILING AND PSEUDOSCIENCE

> **"There is evidence for considerable overoptimism in scientific claims that are based on machine learning model performance, and this probably arises from a poor understanding of the limits of machine prediction in fields beyond computer science"**
>
> Lisa Messeri (Yale University) and M.J. Crockett (Princeton University)[141]

In certain clearly delimited cases, human behaviour can be predictable to a certain extent, but there are limits. Specifically, complex social behaviour has been described as extremely hard or impossible to predict. This chapter will review recent scientific literature on what these limits are, and the reasons behind them. When these limits are ignored, data scientists fail to comply with minimum safeguards that are routine in quantitative science. Information produced in this manner conflicts with sound scientific practice and is therefore unreliable.

An in-depth review of general scientific methodology is beyond the scope of this report, but we will address a few points relevant to protecting human rights from algorithmic harm.[142] These methods include appropriate use of relevant theoretical concepts, subject-matter-appropriate research standards, and guardrails to enforce them. These are drawn from related fields such as sociology and other social sciences, science and technology studies, cognitive science and philosophy of science.

[141] Lisa Messeri and M. J. Crockett, "Artificial intelligence and illusions of understanding in scientific research", March 2024, Nature, Volume 627, Issue 8002, https://www.nature.com/articles/s41586-024-07146-0
[142] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions", September 2024, Patterns, Volume 5, Issue 9, https://www.sciencedirect.com/science/article/pii/S2666389924001600

Understanding how risk profiling leads to human rights harms requires an analysis of the following:

1. How the assumption of objectivity and the omission of rigorous social scientific methods produce harm (section 8.1).
2. Issues of scientific validity, such as understanding what can be measured and predicted (section 8.2).
3. The role of theory in quantitative social science (section 8.3).
4. Making sure that correlations are not spurious and the role of causality in quantitative scientific research (section 8.4), as well as the historical precedents of ignoring causation (section 8.5).
5. What large-scale empirical studies reveal about the limits of prediction of complex social systems (section 8.6).

8.1 ASSUMPTION OF OBJECTIVITY AND OMITTING SOCIAL SCIENTIFIC METHODS PRODUCES HARM

8.1.1 RISK PROFILING IS A TYPE OF SCIENTIFIC INFERENCE

> **"Any attempt at inductive inference about some system in the world on the basis of data… is a kind of scientific inference or, at least, mirrors the fundamental structure of scientific inference and can be thought of as a science-adjacent task"**
>
> Mel Andrews (Princeton University)[143]

The process followed by public authorities deploying risk profiling resembles the process of scientific discovery. Governments and other deployers developing risk profiles engage in a form of scientific inference, with the goal of obtaining some form of knowledge. In risk profiling, this knowledge takes the form of a prediction or assessment that an individual will violate a law or rule, based on the person's shared characteristics with an inferred group of offenders.

In order to produce valid scientific inferences and obtain reliable knowledge, scientists are "trained exhaustively in the protocols of a particular scientific paradigm" and engage in the slow and painstaking scientific process of weighing alternative explanations and carefully evaluating theoretical assumptions.[144] These scientific practices serve as "gatekeeping methods… that typically prevent pseudoscientific research practices from getting through".[145] According to scientific research surveyed by Amnesty International and conversations with leading academic experts, these scientific gatekeeping practices are not present for applied ML in either industry or governmental risk profiling settings.[146] At its core, risk profiling doesn't focus on "how" or "why" a prediction is made, but rather on the predictive accuracy of the outcomes. However, predictions without regard for theory, explanations and causality are unreliable (see further following

[143] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[144] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 10.
[145] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 10.
[146] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited); M. Z. Naser, "On the philosophical naivety of engineers in the age of machine learning", November 2025, Topoi, https://doi.org/10.1007/s11245-025-10304-2
Inioluwa Deborah Raji and others, "The fallacy of AI functionality", 2022, 2022 ACM Conference on Fairness, Accountability and Transparency, https://dl.acm.org/doi/10.1145/3531146.3533158 (accessed 18 June 2024).

sections). As a result, an increasing number of recent studies note that certain ML-powered predictive applications, including risk profiling, "have a pseudoscience problem".[147]

Pseudoscience in the context of this report refers to an activity resembling science but based on fallacious assumptions. The resemblance of risk profiling to science brings an assumption of (increased) objectivity that depoliticizes its discriminatory harms. Risk profiling predictions that disproportionately target racialized and marginalized groups are seen as "objective". This carries an inherent risk for human rights and is becoming a major focus in data science, "as algorithm after algorithm is revealed to be sexist, racist, or otherwise flawed".[148]

8.1.2 FALSE VENEER OF OBJECTIVITY LAUNDERS BIAS AND ERODES ACCOUNTABILITY

> **"These data products seem objective only because the perspectives of those who produce them – elite, white men and the institutions they control – pass for the default."**
>
> Catherine D'Ignazio (Massachusetts Institute of Technology) and Lauren Klein (Emory University, Georgia Tech)[149]

A growing body of scholarship by prominent researchers of AI discusses how the politics, goals, biases and preconceptions of groups that dominate the production of data technologies can be "obscured"[150] and "laundered"[151] by the "imagined objectivity" of these techniques,[152] while reproducing and reinforcing existing inequalities.[153] This is consistent with sociological and psychological research showing that "quantitative measures tend to attract more attention and command greater authority in organizational contexts, and they may thus drive resource allocation and decision-making… regardless of its relative importance to stakeholder values or decision quality".[154] Objectivity suggests an air of neutrality that Donna Haraway and Thomas Nagel famously criticized as "the god trick" and "the view from nowhere" – in other words, an all-seeing and perspective-free stance.[155]

UN Special Procedures have issued several warnings about this issue. For example, when describing the high approval of algorithmic decisions by judges in a court of law, the Special Rapporteur on racism noted that "this high approval rate may well result from a presumption of technological objectivity and neutrality".[156] When sentencing decisions are based on risk assessment instruments that include variables

[147] Jérémie Sublime, "The return of pseudosciences in artificial intelligence: have machine learning and deep learning forgotten lessons from statistics and history?", 2024, http://arxiv.org/abs/2411.18656 (accessed 26 February 2025); Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited); Anita Say Chan, Predatory Data: Eugenics in Big Tech and Our Fight for an Independent Future, 2025; EDRi, EDRi: Beyond Debiasing, 2021, https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/
Mel Andrews, "The immortal science of ML: machine learning and the theory-free ideal", September 2025, Erkenntnis, https://doi.org/10.1007/s10670-025-01010-x
[148] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023.
[149] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023.
[150] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023.
[151] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[152] Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (previously cited); Andrew Guthrie Ferguson, The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement, 2017.
[153] Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, 2016.
[154] Lydia T. Liu and others, "Bridging prediction and intervention problems in social systems", arXiv, 7 July 2025, p.38, https://doi.org/10.48550/arXiv.2507.05216; Theodore M. Porter, Trust in Numbers: The Pursuit of Objectivity in Science and Public Life, 1995; Linda W. Chang and others, "Does counting change what counts? Quantification fixation biases decision-making", 12 November 2024, Proceedings of the National Academy of Sciences, Volume 121, Issue 46, https://doi.org/10.1073/pnas.2400215121; Wendy Nelson Espeland and Mitchell L. Stevens, "A sociology of quantification", December 2008, European Journal of Sociology / Archives Européennes de Sociologie, Volume 49, Issue 3, https://doi.org/10.1017/S0003975609000150; James Chu, "Cameras of merit or engines of inequality? College ranking systems and the enrollment of disadvantaged students", May 2021, American Journal of Sociology, Volume 126, Issue 6, https://doi.org/10.1086/714916; Jerry Muller, The Tyranny of Metrics, 2018.
[155] Thomas Nagel, The View from Nowhere (1986), https://personal.lse.ac.uk/ROBERT49/teaching/ph103/pdf/nagel1986.pdf; Donna Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," Feminist Studies 14, no. 3 (1988): 575–99, https://doi.org/10.2307/3178066; Dasha Pruss, Carceral Machines: Algorithmic Risk Assessment and the Reshaping of Crime and Punishment, 2023.
[156] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis" (previously cited).

on income and gender, this amounts to discrimination "based on demographics and socioeconomic status".[157]

In 2024, four years after the first report of the Special Rapporteur on racism, the new mandate holder reported that the assumption that technology is objective and neutral "remains pervasive and drives a race to integrate AI into society", despite the "multitude of ways in which they are already perpetuating racial discrimination across societal domains", and "without due consideration of whether it is necessary".[158] Andrew Guthrie Ferguson described how police departments hoping to distance themselves from claims of racial bias have eagerly adopted these technologies;[159] resulting in what Ruha Benjamin, author of *Race After Technology,* dubbed a "New Jim Code".[160] This façade of technological objectivity raises questions of accountability in case of harm, and makes it more difficult for affected individuals to seek redress and remedy. The Special Rapporteur on racism called for a rejection of a "colour-blind approach" and urged states to regulate these technologies in a way that recognizes and understands structural racism.[161]

8.1.3 RISK PROFILING IS PLAGUED BY MULTIPLE, INTERLOCKING AND UNAVOIDABLE BIASES

> **"Any predictive models we make are inherently value laden and never purely objective"**
>
> Abeba Birhane (AI Accountability Lab, Trinity College Dublin)[162]

Multiple and interlocking types of bias plague risk profiling. These biases are inherent to the types of data and collection methods that are used to build risk profiling systems and cannot be eliminated by technical means or by adding more data, because they are caused by issues that precede the data generation and collection processes. Besides, all historical data encodes societal inequalities. Bias in risk profiling is not limited to data inputs; it also extends to the fundamental design choices of the algorithm, particularly what the system is configured to optimize in the first place. When public authorities define a goal to be predicted – such as the risk of social security fraud – they embed institutional priorities and structural prejudices directly into the system's logic. These optimization goals dictate how the algorithm works, meaning that even with "flawless" data, the system will still produce discriminatory outcomes if its core objective disproportionately scrutinizes marginalized populations.

Historical and societal bias occur because all historical data reflects historical inequalities. The choice of which behaviours to criminalize, which crimes to focus on and which types of crimes have higher chances of leading to arrests all influence the data distribution and, consequently, the predictions. There are countless other examples of societal bias, and they are not confined to crime data. For example, projects such as Data2X have advanced the idea of a systematic "gender data gap" due to the fact that most research data in scientific studies is based around men's bodies. As a result, drugs can be less safe or effective for women. This has a negative impact on women's health, safety and daily lives. Systemic discrimination can be intersectional and give rise to intersectional data bias. In the United States of America (USA), women of colour are significantly more likely than white women to die from pregnancy- or childbirth-related causes.[163] These issues were decried by women-led reproductive justice groups for decades, but a national system for collecting data on maternal mortality still did not exist in 2018.[164] This shows how what gets measured is a reflection of who and what is prioritized.

---

[157] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination", 2014, Stanford Law Review, Volume 66; Dasha Pruss and others, "Prediction and punishment: critical report on carceral AI", 2024, SSRN Electronic Journal, https://papers.ssrn.com/abstract=5017321 (accessed 26 November 2024).
[158] UN Special Rapporteur on racism, Report, 3 June 2024, UN Doc. A/HRC/56/68.
[159] Andrew Guthrie Ferguson, The Rise of Big Data Policing (previously cited).
[160] Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (previously cited).
[161] UN Special Rapporteur on racism, Report, 3 June 2024, UN Doc. A/HRC/56/68, para. 66.
[162] Abeba Birhane, Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence, PhD thesis, 2022, http://arxiv.org/abs/2206.04179 (accessed 4 June 2024).
[163] US Center for Disease Control and Prevention, cited in D'Ignazio and Klein, Data Feminism, 2023, p. 23.
[164] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023, p. 23; ProPublica, "Nothing protects Black women from dying in pregnancy and childbirth", 7 December 2017, https://www.propublica.org/article/nothing-protects-black-women-from-dying-in-pregnancy-and-childbirth

The proportion of women and men differs in low-paying versus high-paying jobs. Hiring candidates trained on past job data will therefore reproduce gender discrimination. Likewise, a superficial analysis of this data could lead to the conclusion that women are worse candidates for a certain job type, rather than a reflection of systemic gender bias. It merely illustrates that societal bias – the fact that men more often occupy positions of power– is reflected in the generation and collection of data.

> **"If you want to do research on women, you have to embrace qualitative data. There's no two ways about it, because the reality of women's lives is simply not captured in quantitative statistics. Absolutely not."**
>
> Valerie Hudson (Texas A&M University)[165]

Therefore, rather than merely asking whether a risk profile accurately predicts social security fraud, deployers should reflect on the systemic effects of inserting such an algorithm into a public programme that is supposed to be assistive to disadvantaged and marginalized people. In healthcare settings, no-shows in medical appointments "likely correlate with socio-economic status mediated by the patient's inability to cover for transportation or childcare costs" and likely correlate with race because of correlations between socio-economic status and race, resulting from structural racism. Moreover, race correlates with jobs that are paid an hourly wage, and thus with not having flexibility in working hours. Predicting no-shows in order to schedule patients into overbooked appointments is therefore likely to discriminate. (See footnote for an example of how using no-shows in the prediction of "vulnerable children" led to biased outcomes).[166]

During the measurement phase, decisions about what to measure – such as which indicators to collect, what constitutes a good proxy for these abstract concepts, how they are defined, their possible values, and data collection methods – present opportunities for racism and other types of discrimination to occur via inherent biases and stereotypes (see section 8.2 on measurement bias and construct validity). This shows that AI technology, including the predictive models commonly employed in risk profiling algorithms, "is never neutral – it reflects the values and interests of those who influence its design and use, and is fundamentally shaped by the same structures of inequality that operate in society".[167] This explains why "ethical" fraud detection initiatives funded by well-intended public authorities, where significant effort and time was spent obtaining "clean" data, limited proxies, and having open and transparent deliberations with the public, have nevertheless resulted in unpredictable and problematic algorithms (see box in Chapter 9).

Belief in the objectivity of data technologies has real-world implications. Predictive policing systems based on AI "operate in ways that reinforce racial discrimination despite the perceived 'objectivity' of these technologies".[168] In a 2022 report on algorithmic bias, the European Union Agency for Fundamental Rights (FRA) described how police authorities that use "seemingly objective means" to pursue "seemingly objective goals" can cause indirect discrimination if certain marginalized groups are over-represented in these neighbourhoods. For example, systemic racism leads to higher arrest rates in neighbourhoods with a majority of racialized people. If this data is then fed into a geographic predictive policing system, the system will flag such neighbourhoods as "riskier".

[165] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023, p. 170.
[166] In Gladsaxe, Denmark, missed medical appointments were one of several risk indicators in a profiling system used for the early detection of vulnerable children. This system was discontinued after critique from academics, journalists and the Data Protection Authority. In 2025, the Danish Parliamentary Ombudsman launched an inquiry into systems like the Gladsaxe model to examine discrimination concerns and the right of citizens to be notified. This was partly due to increased scrutiny of algorithmic systems after Amnesty International's 2024 publication *Coded Injustice* (described in section 7.4). Algorithm Watch, Automating Society Report (previously cited), p. 52; Jeremy Werner, "Denmark's automated welfare system under fire for surveillance and discrimination", 18 November 2024, BABL AI, https://babl.ai/denmarks-automated-welfare-system-under-fire-for-surveillance-and-discrimination/
[167] UN Special Rapporteur on racism, Report, 3 June 2024, UN Doc. A/HRC/56/68, p. 7.
[168] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis" (previously cited).

8.1.4 NOT JUST DATA BIAS: THE EXAMPLE OF DISABILITY

> **"Including concepts from disability studies can help us refocus and refine our approach to AI bias, moving away from an emphasis on the technology alone, and toward an approach that accounts for the context in which such technology is produced and situated, the politics of classification, and the ways in which fluid identities are (mis)reflected and calcified through such technology."**
>
> Meredith Whittaker (AI Now Institute)[169]

The issue of categorizing disability helps to illustrate the tension between AI systems' reliance on fixed, quantitative data as the primary means of representing the world, and the fluidity of identity and lived experience.[170] The category of "disability" resists classification, because disability encompasses a "vast and fluid" number of physical, mental, intellectual and sensory impairments which can come and go throughout a person's lifetime.[171] Moreover, it is not the impairments on their own, but their interaction with attitudinal and environmental barriers which hinders the full and effective participation of persons with disabilities in society on an equal basis with others.[172] The boundaries of this category have shifted continuously in relation to unstable and culturally specific notions of "ability" that have been constructed according to the needs of industrial capitalism and the shifting nature of work.[173] The way in which disability resists a neat categorization "points to bigger questions about how other identity categories are mistreated as essential, fixed classifications in the logics of AI systems and in much of the research examining AI and bias.[174] (See footnotes for further discussion of the same issues on race and ethnicity.)[175]

These issues are not solved by adding more data or including more categories in order to diversify datasets – two frequently suggested workarounds for data bias. Adding more data on people living with disabilities "simply reinforces the normative model at the core of a given system's calculations, meaning that those who fall outside of this norm become increasingly remote 'outliers'".[176] Given that risk profiling systems are deployed in high-stakes contexts, being labelled as an outlier can have significant life consequences and can contribute to enforcing fixed categories that further marginalize those who do not "fit".[177]

DENMARK: ATTRIBUTION OF THE TERM 'UNUSUAL' REVEALS UNDERLYING NORMS

One of the main principles behind UDK/ATP's fraud detection models is to identify unusual or atypical living patterns or arrangements, relationship patterns and residency patterns as an indicator of fraud.

During an in-person interview with UDK/ATP officials in January 2024, officials stated that a residential address recorded in the database with "far too many residents" in relation to its size would be regarded

[169] AI Now Institute and Meredith Whittaker, Disability, Bias, and AI, 20 November 2019, https://ainowinstitute.org/publications/disabilitybiasai-2019, p. 11.
[170] AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 10.
[171] Shari Trewin, "AI fairness for people with disabilities: point of view", 2018, http://arxiv.org/abs/1811.10670 (accessed 11 March 2026); AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 10.
[172] CRPD, preamble (e).
[173] Sarah F. Rose, No Right to Be Idle: The Invention of Disability, 1840s–1930s, 2017, p. 2; AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 10.
[174] AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 10.
[175] Sebastian Benthall and Bruce D. Haynes, "Racial categories in machine learning", 2018, https://arxiv.org/abs/1811.11668
Amina A. Abdu and others, "An empirical analysis of racial categories in the algorithmic fairness literature", 2023, https://dl.acm.org/doi/10.1145/3593013.3594083 (accessed 19 June 2023); Alex Hanna and others, "Towards a critical race methodology in algorithmic fairness", 2020, https://arxiv.org/abs/1912.03593 (accessed 27 November 2023); Gerwin van Schie, "The datafication of race-ethnicity: an investigation into technologically mediated racialization in Dutch governmental data systems and infrastructures", 2022, https://dspace.library.uu.nl/items/de710224-becb-4e5b-b6e6-8133f7502085
[176] AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 10.
[177] AI Now Institute and Meredith Whittaker, Disability, Bias, and AI (previously cited), p. 3.

as an indicator of fraud. UDK does not clearly define within the law what constitutes "unusual" or "atypical arrangements" in regard to households, leaving the door open to arbitrary decision-making. This broad definition of cohabitation creates a risk that people or groups living in the same household in what are regarded as "unusual" living or family arrangements are more likely to be flagged for fraud. This risks disproportionately targeting and surveilling people with "non-traditional" living arrangements, such as people living with disabilities, older people, low-income groups and migrants. This is because, for example, due to cultural preferences, households with people from migrant backgrounds tend to be composed of multigenerational families, unlike "traditional" Danish households. Therefore, the models are embedded with social norms that reflect the view of dominant groups in Denmark about what a household or a family is.

The use of what appear to be neutral policies and variables in UDK/ATP's models to identify "unusual" patterns in behaviour as an indicator of benefits fraud can lead to stress and anxiety, particularly for marginalized communities, when faced with needing to prove that they are not cohabiting. This evidence also shows that the use of these policies and variables in UDK/ATP's models can lead to indirect discrimination against groups based on their migration status, race, class, disability, age and marital or relationship status, which can subsequently lead to denial of the right to social security if they are forced to repay the benefits they have received, or if benefits are unduly delayed.

8.1.5 SOCIAL DATA IS CONSTRUCTED AND IMBUED WITH SOCIAL MEANING

> **"Data has power"**
>
> danah boyd (Data & Society Institute)[178]

In the social sciences, unlike in the natural sciences, the measurements that serve as data for scientific debate are objects of debate themselves.[179] This is due to the latent nature of the concepts that are being measured – a methodological challenge so profound that an entire field of research is dedicated to addressing it: measurement science. Social scientists are typically concerned with social constructs (for example, the construct of "intelligence"; see also section 8.2 and section 8.5) or quantifications of internal personal states, beliefs or attitudes, such as the propensity to commit crime or fraud, that are not directly observable or measurable. This is different from the natural sciences that use technical instruments to measure physical properties such as weight, distance or speed. It is unsurprising, therefore, that the legitimacy of data used for training ML and AI algorithms to make predictions about human behaviour has come under increased scrutiny.

> **"Data tell stories. And they can be used to sell lies"**
>
> danah boyd (Data & Society Research Institute)[180]

[178] danah boyd, "Questioning the legitimacy of data", August 2020, Information Services and Use, Volume 40, Issue 3, https://doi.org/10.3233/ISU-200098
[179] Alain Desrosières, The Politics of Large Numbers: A History of Statistical Reasoning, 1998.
[180] danah boyd, "Questioning the legitimacy of data" (previously cited).

Societal bias is not merely a "data problem", but a reflection of living social processes.[181] Social science theories can help situate data, AI and risk profiling within larger historical and social forces and systems of oppression such as white supremacy,[182] patriarchy,[183] heteronormativity[184] and ableism.[185]

The relevant sociological frameworks, research principles and methodologies remain notably absent from the educational background of data scientists and AI developers. Computer scientist Ben Green and sociologist Mike Zajko explain that part of the tension between data science and social science arises from the attempt by data scientists to "remain neutral, objective and fair" and avoid "being overtly political".[186] Social data is used as if it represented "natural properties" of individuals rather than reflections of historical social processes.[187] Ignoring the context of social data therefore reproduces an unequal past shaped by systemic discrimination into a present marked by unequal opportunities.

8.1.6 THE ROLE OF PREDICTION IN SOCIAL SCIENCE

Human behaviour is non-deterministic, meaning it cannot be reliably predicted from any given initial set of conditions. This has consequences for the reliability and effectiveness of technology designed to predict human behaviour.

Statistical prediction has been a pillar of scientific inquiry since the 18th century: scientific theories are not only evaluated based on their descriptive accuracy, but also on their ability to make testable predictions about future observations.[188] This is particularly the case for the physical, engineering and biological sciences, where certain objects of research behave in ways that can be described as relatively deterministic if compared to the complexity of human beings. In other words, the physical sciences often deal with *relatively* unambiguous systems with rule-like mechanisms, which makes theories and models of prediction *relatively* straightforward.

Early social scientists considered the field of sociology to be similar to the natural sciences, but they soon realized that the goal of predicting human behaviour was too ambitious. "Individual-level prediction has not historically been a goal" for many sociologists and social scientists.[189] The natural and social worlds are inherently different. Human society is characterized by *culture, social structures, complex interactions, relationships* and processes of *socialization*. In other words, human beings, unlike physical or simpler biological systems, are non-deterministic. Their actions cannot unambiguously be inferred by some pre-determined set of inputs. This problem is not confined to sociology but applies to other social sciences including economics and psychology.

Computational social scientists Duncan Watts and Jake Hoffman affirm that "there is no single answer" to the question of how predictable human behaviour is, though noting that behaviour resembles "the outcome

[181] Mike Zajko, "Conservative AI and social inequality: conceptualizing alternatives to bias through social theory", September 2021, AI and Society, Volume 36, Issue 3, http://arxiv.org/abs/2007.08666
[182] Charles W. Mills defines global white supremacy as a political system that "encompasses de facto as well as de jure white privilege and refers more broadly to the European domination of the planet that has left us with the racialized distributions of economic, political and cultural power that we have today." See Charles W. Mills, "Revisionist Ontologies: Theorizing White Supremacy" in Blackness Visible: Essays on Philosophy and Race, 1998, p. 98.
[183] "Patriarchy represents a system entrenched in beliefs, behaviours, and practices that uphold the subservient status of women while elevating men to superior positions. This system revolves around the perception of the male figure as the central figure within the family, entrusted with making all pivotal decisions in social spheres, while relegating women to domestic roles, expected to prioritize caregiving duties, including tending to children, managing household chores, and caring for elderly or ill family members." See Veronica Beechey, "On patriarchy", 1970, Feminist Review, Volume 3, Issue 1, https://doi.org/10.1057/fr.1979.21
[184] "Heteronormativity is a hegemonic system of norms, discourses, and practices that constructs heterosexuality as natural and superior to all other expressions of sexuality. Queer theorist Michael Warner (1991) coined the term heteronormativity to illuminate the privileging of heterosexuality in social relations, which relegates sexual minorities to a marginal status position. Heteronormativity legitimates homophobia – the irrational fear of gay and lesbian people – and heterosexism – the discrimination of sexual minorities within social relations and structures." Brandon Andrew Robinson, "Heteronormativity and homonormativity", in Angela Wong and others (editors), The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies, 1st edition, 2016, https://doi.org/10.1002/9781118663219.wbegss013
Michael Warner, "Introduction: fear of a queer planet," 1991, Social Text, Volume 29, https://www.jstor.org/stable/466295
[185] "Ableism against disabled people reflects a preference for species-typical normative abilities leading to the discrimination against them as 'less able' and/or as 'impaired' disabled people. This type of ableism is supported by the medical, deficiency, impairment categorization of disabled people (medical model). It rejects the 'variation of being', biodiversity notion and categorization of disabled people (social model). It leads to the focus on 'fixing' the person or preventing more of such people being born and ignores the acceptance and accommodation of such people in their variation of being. Ableism has also long been used to justify hierarchies of rights and discrimination between other social groups, and to exclude people not classified as 'disabled people'." Gregor Wolbring, "The politics of ableism", 2008, Development, Volume 51, Issue 2, https://doi.org/10.1057/dev.2008.17
[186] Mike Zajko, "Conservative AI and social inequality" (previously cited).
[187] Raphaële Xenidis and Linda Senden, "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination", in Ulf Bernitz and others (editors), General Principles of EU law and the EU Digital Order, 2020.
[188] Jake M. Hofman and others, "Prediction and explanation in social systems", February 2017, Science, Volume 355, Issue 6324.
[189] Ian Lundberg and others, "The origins of unpredictability in life outcome prediction tasks", June 2024, Proceedings of the National Academy of Sciences, Volume 121, Issue 24, https://pnas.org/doi/10.1073/pnas.2322973121

of a die roll" and that it ranges from "regular to wildly unpredictable".[190] Correspondingly, the potential for predicting individual behaviour and the outcomes of complex social systems is highly constrained, according to Hoffman and colleagues.[191] This explains why social scientists "have generally de-emphasized the importance of prediction relative to explanation, which is often understood to mean the identification of interpretable causal mechanisms".[192] This reflects the intrinsic complexity of human social systems.[193]

The challenges involved in building technologies that successfully predict human behaviour are ultimately not related to the technology itself, but to the very nature of social processes.[194] The limitations of predictive technology such as automated risk profiling algorithms are therefore unlikely to be solved by technological advances or quick technical fixes:

> **"accurately predicting people's social behaviour is not a solvable technology problem"**
>
> Arvind Narayanan (Princeton University)[195]

8.2 MEASUREMENT AND VALIDITY ISSUES ARE INHERENT TO RISK PROFILING

> **"In many cases where algorithms prove unsuitable for real-world use, the problem originates in the initial problem formulation stages"**
>
> Amanda Coston (Carnegie Mellon University)[196]

Literature on the scientific validity of prediction models raises a dire warning on their legitimacy, especially when they are powered by ML. Many recently published ML applications claiming to predict "unobservable character traits" such as criminality, "misrepresent the human reality" in which they are being deployed. This is because they do not observe the scientific safeguards routinely applied in social science to correctly infer unobservable characteristics.[197] When these predictive technologies are employed in high-stakes contexts, they should be expected to harm marginalized groups.[198] Procedural safeguards "widely adopted in social sciences" include criteria to test the validity of a system or method. Validity considerations help to detect algorithmic harms *before* deployment and help to decide whether algorithms "are suitable in the first place for… high-stakes decision-making tasks".[199]

[190] Jake M. Hofman and others, "Prediction and explanation in social systems" (previously cited).
[191] Jake M. Hofman and others, "Prediction and explanation in social systems" (previously cited).
[192] Jake M. Hofman and others, "Prediction and explanation in social systems" (previously cited), p. 1.
[193] Jake M. Hofman and others, "Prediction and explanation in social systems" (previously cited).
[194] Arvind Narayanan and Sayash Kapoor, AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference, 2024.
[195] Arvind Narayanan and Sayash Kapoor, AI Snake Oil (previously cited).
[196] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms", 2022, https://arxiv.org/abs/2206.14983 (accessed 18 June 2024).
[197] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 2.
[198] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[199] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms" (previously cited).

8.2.1 RISK IS NEITHER OBSERVABLE NOR MEASURABLE

First, for a predictive model to be considered valid, it needs to accurately predict a specific construct or quantity (*construct validity*).[200] The term "construct" indicates that the object of inquiry is abstract, socially constructed and shaped by the social context in which it exists. For example, "the concept of criminal propensity is a human construct, and the data, to the extent that they encode this concept, do so only by virtue of having been shaped by human judgement of criminality. There is no objective signal of criminality that can be discovered independently of human judgement".[201] Early positivist criminologists like Cesare Lombroso believed that criminality or criminal types existed in an "objective", biological sense, and believed that they could therefore be measured and quantified. These theories are not supported by empirical evidence (see section 8.5 on eugenics and positivist criminology).[202] Therefore, criminality, or the individual propension to commit a crime or fraud, are unmeasurable and thus fundamentally unrealistic goals for predictive systems.[203]

Attempting to predict the individual propension to commit crime or fraud via risk profiling requires a rigorous measurement process that involves agreeing on a valid and reliable definition of "criminal propensity" and identifying valid and reliable observable indicators of "criminal propensity" in order to quantify it. To Amnesty International's best knowledge, and based on conversations with academic experts, this rigorous measurement process is absent in risk profiling by governments. Doubts about the scientific validity of risk profiling used in law enforcement, social security or migration are therefore legitimate: validity concerns are "pervasive".[204]

When measurements of the construct being predicted are unavailable, a readily available proxy is chosen, which can encode historical biases.[205] There are many concrete examples of predictive models using readily available proxies because of the impossibility of measuring the intended construct. For example, it is currently not possible to predict the rate of re-offending, compared to the rate of re-arrest. This is concerning given racial disparities in arrest rates.[206] "A model that appears accurate with respect to re-arrests may be quite inaccurate with respect to actual crime".[207] The desired object of prediction (criminal activity) is not directly observable, so a proxy that is observable is chosen (re-arrests). This difference is crucial: not only is this an entirely different quantity, but some racialized groups are often subjected to higher surveillance and arrest rates without necessarily committing more crimes because of structural and institutional discrimination. Crimes committed by more privileged demographic groups, in contrast, often go undetected.[208] Moreover, the chosen proxy (re-arrests) is just as likely to indicate increased police activity and institutional racism rather than increased criminality. If the rate of observation or sampling is increased and the number of arrests increases as a result, this cannot be differentiated from a true base rate discrepancy (so-called "ground truth", discussed below).

Because input and prediction goals are chosen early during risk profiling development, threats to validity can therefore arise as early as during the formulation of the problem. All validity criteria are also influenced by the choice of the inputs or personal attributes used in building a risk profile. If there is no plausible causal relationship between a personal attribute and the behaviour being predicted, this means any correlation between them is purely spurious. Including such an attribute in a prediction model immediately undermines the model's *internal* and *external* validity.[209] A spurious correlation is one in which two or more variables are associated, but not causally related, by sheer coincidence or through the presence of other unseen factors. (Spurious correlations are further discussed in section 8.4).

---

[200] Abigail Z. Jacobs, "Measurement as governance in and for responsible AI", 2021, http://arxiv.org/abs/2109.05658 (accessed 18 November 2024); Abigail Z. Jacobs and Hanna Wallach, "Measurement and fairness", 2021, Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency, https://dl.acm.org/doi/10.1145/3442188.3445901 (accessed 22 February 2026).
[201] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[202] Eamon Carrabine and others, *Criminology: A Sociological Introduction*, 4th ed. (Routledge, 2020), https://doi.org/10.4324/9781315123509
[203] Lydia T. Liu and others, "Bridging Prediction and Intervention Problems in Social Systems" (previously cited), p. 27.
[204] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms" (previously cited).
[205] Amanda Coston, "Falsifying predictive algorithms", 2026, http://arxiv.org/abs/2601.17146 (accessed 30 January 2026).
[206] Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (previously cited); Lydia T. Liu and others, "Bridging prediction and intervention problems in social systems" (previously cited).
[207] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms" (previously cited).
[208] Jeffrey Reiman and Paul Leighton, The Rich Get Richer and the Poor Get Prison: Thinking Critically About Class and Criminal Justice, 2020.
[209] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms" (previously cited).

8.2.2 SHAKY GROUND TRUTH: RISK OF CRIME OR SOCIAL SECURITY FRAUD ARE NOT RELIABLE MEASURES FOR PREDICTION TASKS

"Ground truth" refers to the information that is considered to be true and accurate based on direct observation or physical measurement. In various fields, especially in data science, remote sensing, and machine learning, it serves as the benchmark for validating the accuracy of models and predictions. The term implies a fundamental or baseline truth against which data or analyses can be compared.

In fraud detection, the idea of "ground truth" takes on a particular importance, because automated systems are trained and evaluated against the labels that organizations treat as true instances of fraud. Automated fraud detection has become one of the most established applications for industry and government data mining.[210] In certain domains, such as credit card fraud, automated fraud detection is described by review articles as "effective".[211]

Nevertheless, even in credit card fraud detection, accurately defining and measuring a behavioural goal requires a reliable ground truth: "it is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction".[212] Statistical analysis alone is insufficient to ascertain with certainty that fraud has been perpetrated.[213] The best a statistical approach can deliver is an observation that is "anomalous or more likely to be fraudulent than others, so that it can be investigated in more detail".[214] One can think of this observation as a "suspicion score",[215] which draws parallels to Amnesty International's definition of risk profiling.

"One can think of the objective of the statistical analysis as being to return a suspicion score (where we will regard a higher score as more suspicious than a lower one). The higher the score is, then the more unusual is the observation or the more like previously fraudulent values it is."[216]

This is why banks typically reach out to credit card holders to inquire whether a flagged transaction is legitimate. A credit card transaction is a single event that can reliably and quickly be labelled as authentic or fraudulent, and banks typically contact customers to verify transactions.[217] Therefore, credit card fraud has a relatively reliable ground truth, which makes it possible to clearly specify what the system should detect when identifying past fraud.

By contrast, social security fraud typically unfolds over much longer periods of time and is much harder to verify. Crucially, in social security fraud detection, distinguishing between honest mistakes and deliberate fraud is "hard and messy".[218] Potentially, such a process could take years and require judicial procedures in courts of law. In applications of ML for detection of social security fraud, researchers have highlighted that fraud "can be hard to ascertain because it requires knowledge of the intent".[219] Circling back to the concept of validity, using involuntary mistakes as a predictive goal in the detection of fraud in social security payments measures a different quantity than actual social security fraud, which immediately undermines the validity of the predictions and has potentially grave consequences for the people affected (see box below).

Therefore, even though detecting "fraud" involves investigating a past event and is not a strictly predictive task, it is much more realistic to use these techniques to identify credit card fraud than social security fraud. The availability of a reliable ground truth is very different for these two tasks, and as a result it is extremely difficult to define a clear indicator for detecting fraud among social security claimants. To solve this problem, authorities often treat mistakes in social security applications as a proxy for fraudulent applications. However,

---

[210] Clifton Phua and others, "A comprehensive survey of data mining-based fraud detection research", May 2012, Computers in Human Behaviour, Volume 28, Issue 3, http://arxiv.org/abs/1009.6119

[211] Although "an evident lack of publicly available datasets, labelled or not, was identified as a significant limitation in this field", see Rejwan Bin Sulaiman and others, "Review of machine learning approach on credit card fraud detection", June 2022, Human-Centric Intelligent Systems, Volume 2, Issue 1, https://doi.org/10.1007/s44230-022-00004-0
Similarly, Amnesty International's conversations with academic experts highlighted a difficulty in finding scientific studies on the effectiveness of credit card fraud detection models that are actually deployed in practice. Interview with Hilde Weerts, Eindhoven University of Technology, https://hildeweerts.nl/

[212] Clifton Phua and others, "A comprehensive survey of data mining-based fraud detection research" (previously cited), p. 1.

[213] Richard J. Bolton and David J. Hand, "Statistical fraud detection: a review", August 2002, Statistical Science, Volume 17, Issue 3, https://projecteuclid.org/journals/statistical-science/volume-17/issue-3/Statistical-Fraud-Detection-A-Review/10.1214/ss/1042727940.full, 235–255, p. 2.

[214] Richard J. Bolton and David J. Hand, "Statistical fraud detection" (previously cited), p. 2.

[215] Richard J. Bolton and David J. Hand, "Statistical fraud detection" (previously cited), p. 2.

[216] Richard J. Bolton and David J. Hand, "Statistical fraud detection" (previously cited), p. 2.

[217] Andrea Dal Pozzolo and others, "Credit card fraud detection: a realistic modeling and a novel learning strategy", August 2018, IEEE Transactions on Neural Networks and Learning Systems, Volume 29, Issue 8, https://ieeexplore.ieee.org/document/8038008/

[218] Gabriel Geiger, Suspicion Machines, Lighthouse Reports, 2023, https://www.lighthousereports.com/investigation/suspicion-machines/

[219] Amanda Coston, "Falsifying predictive algorithms" (previously cited), p. 2.

this doesn't solve the underlying problem: the true instances of fraud in the total population are unknown. Any meaningful estimation will need to grapple with the social, historical, and cultural complexity as well as measurement validity of defining, operationalizing and quantifying the concept "social security fraud". As things stand, social security fraud detection is therefore a predictive exercise based on mostly socio-demographic factors, rather than the identification of past instances of fraud.

Structural discrimination further aggravates this problem. The observation of fraud among social security claimants is likely to be biased towards marginalized groups, since they are over-represented in social security as a result of systemic racism (among others), and are likely to have been subjected to more checks because of institutional racism. Risk profiling models will learn discriminatory patterns resulting from past enforcement rather than true fraud signals. Moreover, variables used in social security fraud detection – such as household size and living arrangements - are much more likely to act as proxies for demographic and socio-economic characteristics. If techniques such as anomaly detection are applied to these variables, this results in risk profiling in practice and disproportionally affects marginalized groups who deviate from what is seen as the societal norm. Social security data includes socio-demographic variables and personal traits that correlate with disability, age, socio-economic status, race and ethnicity, among others. This makes anomaly detection in fraud detection among social security claimants very likely to have discriminatory outcomes.

For example, healthcare costs are a poor proxy for healthcare needs, because Black patients with similar health needs to white patients incur lower costs due to structural barriers and inequalities in access to healthcare.[220] Such misalignment happens considerably often in practice.[221] In child social security, screening algorithms have been found to be biased against Black families and families with disabilities, raising the question of whether algorithms are simply flagging families who already interact more with social services.[222]

NETHERLANDS: INACCURACIES IN SOCIAL SECURITY CLAIMS AS PROXY FOR FRAUD

In 2021, Amnesty International published the report *Xenophobic Machines* about the algorithmic decision-making system for fraud detection that was introduced in 2013 by the Dutch tax authorities.

The practice of detecting inaccuracies in applications was used as an early warning sign, or proxy, for "potential fraud". This is because actual social security fraud cannot be quantitatively measured or operationalized for a predictive task. The risk classification model was developed by comparing historical examples of correct and incorrect applications. In practice, this meant that the more an incoming application resembled an application that had previously been classified as inaccurate, the higher the risk score assigned to that incoming application – in other words, the higher the supposed risk of fraud.

Because the risk profiling system included the criterion "non-Dutch nationality", people from racialized groups had an increased chance of being categorized as possibly committing fraud and, therefore, being selected for an investigation compared to Dutch people. People from a non-Dutch background were seen as more prone to commit fraud, indicative of the tax authorities' perception that a link exists between race, ethnicity and crime, and an acceptance of the practice of applying generalizations to the behaviour of individuals who share the same race or ethnicity. This practice not only reflected the tax authorities' attitude towards certain nationalities and ethnic minorities who were negatively stereotyped as deviant or fraudulent, but also drove further stigmatization of these groups, which is incompatible with the prohibition of racial discrimination, as is outlined in Chapter 6. The use of other criteria such as low income in combination with a high benefit claim resulted in a disproportionate effect on single caregivers, usually women and/or people from low-income households.

In sum, the scientific validity of risk profiling models is fundamentally weak or absent. Constructs like the individual risk of committing crime or social security fraud are complex social and cultural phenomena that are ambiguous, dynamic and intertwined with numerous other factors. Therefore, they are extremely difficult to operationalize and measure and, as such, they are not realistic goals for predictive systems. As a result,

[220] UN Special Rapporteur on the right of everyone to the enjoyment of the highest attainable standard of physical and mental health, Tlaleng Mofokeng, Racism and the Right to Health: Report, 2022, UN Doc. A/77/197; CERD, General Recommendation 37 on Equality and Freedom from Racial Discrimination in the Enjoyment of the Right to Health, 2025, UN Doc. CERD/C/GC/37.
[221] Amanda Coston and others, "A validity perspective on evaluating the justified use of data-driven decision-making algorithms" (previously cited).
[222] Sally Ho and Garance Burke, "Oregon dropping AI tool used in child abuse cases", 2 June 2022, Associated Press, https://apnews.com/article/politics-technology-pennsylvania-child-abuse-1ea160dc5c2c203fdab456e3c2d97930

constructing a risk profile for social security fraud or criminality is not a realistic technical undertaking, nor is it a credible exercise in evidence-based policy. It is an attempt to operationalize suspicion in the absence of a reliable ground truth, and it is bound to discriminate because of biases inherent to the data and the social phenomena that they quantify. Such models are prone to produce biased and erroneous predictions and expose targeted individuals to arbitrary and abusive outcomes.

8.3 RISKS OF THEORY-FREE PREDICTION

"No use of Machine Learning in science is 'theory-free'"

Mel Andrews (Princeton University)[223]

Risk profiling developers, and data scientists more broadly, often develop their systems without explicitly stating the underlying theory or theories used. ML is believed to be "theory-agnostic", in that "there are no a priori assumptions concerning the mechanism of the target phenomenon".[224] An application of ML that would bypass the need for theory "has been decried as a scientific malpractice, its results at best uninformative, at worst, dangerously misleading".[225] If decisions are taken based on predictions that are not rigorously validated, they should be expected to cause social harm, since "wrong theories generate wrong interventions. Wrong interventions cause harm".[226]

To guarantee that predictions generated by ML are well-founded, disciplinary methodological norms should be followed, such as articulating a causally plausible theory, grounding it in existing theoretical frameworks and then testing it on empirical data using gold standard methods of causal inference.[227] A popular belief among proponents of ML is that mining vast amounts of data and using correlations in the data to make predictions bypasses the need for articulating a theory or causal relationship behind the prediction.[228] In this approach, data and the prediction they generate are considered objective, and correlation is treated as inherently predictive. The appeal of "letting the numbers speak for themselves" and of assuming that this will generate a more value-free or objective form of knowledge is ever-present in contemporary discourse about (big) data analysis, as demonstrated by a widely referenced article in 2008 that claimed big data would make the scientific method obsolete and determine the "end of theory".[229] AI and other digital technologies have been presented to possess revolutionary potential despite very thin evidence for the claims made.[230] AI hype might be responsible for the belief in theory-free science,[231] but this belief is actually making science "less innovative and more vulnerable to errors".[232]

As discussed in previous sections, in the social sciences, any data is understood as an abstract, formalized representation, an "indicator" that is being chosen to measure a given latent construct. In the context of risk profiling, where indicators relate to individuals and their situation in society, any choice of indicators is necessarily "theory-laden". The indicators do not hold any meaning beyond the one attributed to them by the developer or user; they are not objective "in some strong, ontological sense".[233] The choice of which problem to solve, what to optimize for, which data to collect and how, its categorization, the choice and engineering of variables and the interpretation of model results are all informed by some sort of undeclared theoretical commitment.[234] Making observations or taking measurements is inseparable from the guidance of

[223] Mel Andrews, "The immortal science of ML" (previously cited).
[224] Sanja Srećković and others, "The automated Laplacean demon: how ML challenges our views on prediction and explanation", March 2022, Minds and Machines, Volume 32, Issue 1, https://doi.org/10.1007/s11023-021-09575-6, 159–183, p. 165.
[225] Mel Andrews, "The immortal science of ML" (previously cited), p. 14.
[226] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 2.
[227] Mel Andrews, "The immortal science of ML" (previously cited); Megan T. Stevenson, "Cause, effect, and the structure of the social world", May 2023, SSRN Electronic Journals, https://papers.ssrn.com/abstract=4445710
[228] Cristian S. Calude and Giuseppe Longo, "The deluge of spurious correlations in big data", September 2017, Foundations of Science, Volume 22, Issue 3, https://doi.org/10.1007/s10699-016-9489-4
[229] Chris Anderson, "The end of theory: the data deluge makes the scientific method obsolete", Wired, https://www.wired.com/2008/06/pb-theory/
Cristian S. Calude and Giuseppe Longo, "The deluge of spurious correlations in big data" (previously cited).
[230] Emily M. Bender and Alex Hanna, The AI Con: How to Fight Big Tech's Hype and Create the Future We Want, 2025.
[231] Jens Ulrik Hansen and Paula Quinon, "The importance of expert knowledge in big data and machine learning", January 2023, Synthese, Volume 201, Issue 2, https://doi.org/10.1007/s11229-023-04041-5, p. 35.
[232] Lisa Messeri and M. J. Crockett, "Artificial intelligence and illusions of understanding in scientific research" (previously cited).
[233] Mel Andrews, "The immortal science of ML" (previously cited), p. 13.
[234] Jens Ulrik Hansen and Paula Quinon, "The importance of expert knowledge in big data and machine learning" (previously cited); Mel Andrews, "The immortal science of ML" (previously cited).

background theory.[235] Philosophers commonly object that the prospect of arriving at empirical knowledge by a fully predictive science is naïve and ends up taking the status quo as an assumption. Most of all, it ignores one of the primary aims of science, namely, explanation.[236]

Research carried out by Amnesty International does not suggest that governments are grounding risk profiling in rigorously validated theories. Typically, governments act on a perceived necessity to make policy more effective, efficient or otherwise cost-effective, and are influenced by corporate narratives and over-optimistic promises about technology. Rather than collecting data that is specific or pertinent to the prediction of the behaviour labelled as "risky", governments use pre-existing and unfit-for-purpose administrative data to train predictive models (so-called convenience samples).[237] To guarantee methodological or epistemic validity, researchers would need to follow discipline-specific research norms that are appropriate to the domain of application. The fact that constructs such as propensity to crime are fundamentally unfeasible to operationalize, or that the "objective" type of data required to perform prediction does not and will never exist, would emerge in early stages of this process. This would likely prevent the proliferation of ill-founded and unreliable prediction models.

DENMARK: CHOICE OF MODEL VARIABLES REVEALS THE UNDERLYING CAUSAL THEORY

In 2023, Amnesty International published *Coded Injustice*, a long-form research investigation into the use of technology within Denmark's social benefits system, administered by the public authority Udbetaling Danmark (UDK, or Pay Out Denmark) and the company Arbejdsmarkedets Tillægspension (ATP).[238]

As part of its targeting of people with "foreign affiliations" for fraud investigations, our research has found that UDK uses data on residents' foreign residence, entry and exit abroad, marital status, number of children, real estate or vehicles abroad and social benefits received to be used within its algorithm to flag people for further investigation for social benefits fraud.

This choice of variables to be included in the predictive model reveals the underlying theory: these indicators are presumed to contribute to the likelihood that an individual will commit fraud. However, this choice of indicators is not corroborated by a rigorously validated theory. For example, it lacks systematic empirical support demonstrating robust causal relationships and clear theoretical mechanisms connecting the selected indicators to the outcomes they are intended to explain.

To sum up, if the underlying causal relation is unspecified, it cannot be subjected to critical scrutiny. More importantly, the model cannot be robust, because it is unknown whether the patterns the model is picking up on are causal or spurious. There can be no reliable way to know where, when and how the model and its predictions will fail. Therefore, even if such a model were to achieve high predictive performance, its predictions can be expected to cause social harm, because the model will arguably misrepresent the human reality in which it is deployed, facilitate the misinterpretation of correlations as causal, and mask implicit ideas and norms into the output predictions. These predictions will then be mistreated as objective empirical truths.[239]

[235] Jules Desai and others, "The epistemological foundations of data science: a critical review", November 2022, Synthese, Volume 200, Issue 6, https://doi.org/10.1007/s11229-022-03933-2, p. 469.
[236] Mel Andrews, The Immortal Science of ML (previously cited), p. 13; M. Z. Naser, On the Philosophical Naivety of Engineers in the Age of Machine Learning (previously cited).
[237] Lydia T. Liu and others, "Bridging prediction and intervention problems in social systems" (previously cited).
[238] Amnesty International, *Denmark: Coded Injustice* (previously cited).
[239] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 6.

8.4 IGNORING CAUSALITY DOES NOT BELONG IN (DATA) SCIENCE

> **"[C]ausality cuts through multiple correlations in order to find the things that really matter".[240]**

In contrast to statistical models focused on explaining observations based on theory, most ML models do not supplement the correlations found in data with a potential causal interpretation.[241] This is certainly true for risk profiling models. The very appeal of "data-driven" techniques resides in their ability to discover correlations directly from large datasets without relying on pre-specified theoretical assumptions. However, ruling out meaningless or spurious correlations by researching the causal mechanisms underlying an observed correlation is a key practice of trained scientists.[242]

A spurious correlation is one in which two or more variables are associated but not causally related, such as the number of ice cream sales and the number of deaths by drowning. These quantities tend to vary together – they are correlated. However, this can be due to coincidence or the presence of other unseen factors, called "confounding variables". In this example, such a confounder is the variable "temperature": people tend to eat more ice cream and swim more often on hot days and less often on cold days, meaning that while the two phenomena (ice-cream eating and drowning) occur together, one is not causing the other.

Sometimes, statistical significance tests are applied to validate the selection of variables in risk profiling (see section 9.2). However, finding that a correlation is statistically significant in a hypothesis test is not indicative of causality. Statistical significance merely indicates that such a correlation would be highly unlikely to occur by pure chance. But it does not imply a genuine or causal effect by the variable of interest,[243] nor does it measure the size or practical importance of the effect. For example, if the correlation between ice cream sales and deaths by drowning is statistically significant, this means that this correlation is unlikely to have happened due to pure chance and is therefore likely to be observed again by repeating the measurement on a different day. If we repeat the measurement on another hot day, we are likely to again observe both high ice cream sales and deaths by drowning. However, this says nothing about the causal relation between these two variables. Causality depends on a bare minimum of three conditions: "(1) correlation; (2) the cause preceding the effect; and (3) the absence of a third variable that could explain the correlation".[244]

Ruling out confounders is hard and time-consuming work, because the number of possible confounding variables is open-ended.[245] Austin Bradford Hill delineated nine criteria for distinguishing correlation and causality, including temporality, consistency and theoretical plausibility.[246] Hill emphasized that statistical correlation alone can never substantiate causal claims and that researchers must rigorously evaluate a range of multiple factors and conditions before accepting the validity of correlations.

> **"Machine learning, at its core, is a tool that predicts. It reveals statistical correlations but with no understanding of causal mechanisms."**
>
> Abeba Birhane (AI Accountability Lab, Trinity College Dublin)[247]

---

[240] Wendy Hui Kyong Chun and Alex Barnett, Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition, 2021, p. 57.
[241] Sanja Srećković and others, "The automated Laplacean demon" (previously cited), p. 160.
[242] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[243] Joshua D. Angrist and Jörn-Steffen Pischke, Mastering 'Metrics': The Path from Cause to Effect, 2014; Megan T. Stevenson, "Cause, effect, and the structure of the social world" (previously cited).
[244] Wendy Hui Kyong Chun and Alex Barnett, Discriminating Data (previously cited), p. 57.
[245] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited); Megan T. Stevenson, "Cause, effect, and the structure of the social world" (previously cited).
[246] Austin Bradford Hill, "The environment and disease: association or causation?", May 1965, Proceedings of the Royal Society of Medicine Volume 58, Issue 5, https://doi.org/10.1177/003591576505800503
[247] Abeba Birhane, Automating Ambiguity (previously cited).

When government officials subject someone to an investigation following a high score in a risk profile, they give de facto causal interpretation to the prediction or risk score.[248] Government officials dealing with risk profiling commonly treat correlation as inherently predictive, which effectively amounts to stereotyping. Therefore, people are punished for being part of a statistical group; something which "should be expected" to harm marginalized groups in particular.[249] "Potential harms flowing from the misdiagnosis of a causal hypothesis… include false detainment and imprisonment… exclusion from educational opportunities, and exclusion from employment opportunities", among others.[250]

Amnesty International is not implying that statistical correlations are meaningless. The world is structured and contains regularity: robust patterns in data are to be expected.[251] But it is well-known that, given sufficiently large amounts of data, the number of observed correlations increases, both meaningful and spurious. It has been demonstrated mathematically that very large datasets are riddled with spurious correlations.[252] In other words, ML applied to very large datasets has the potential to make everything appear relevant.[253] This phenomenon is so notorious that it has several idiomatic expressions, such as the "multiple comparison problem", "data dredging" and "fishing expeditions".[254]

Amnesty International is also not implying that causation should be demanded as the lowest threshold for justifying any and all policies, as true causation is hard to prove. But policy should be based on reliable evidence, and mere correlations cannot be accepted as justification for high-impact, rights-restricting policies. Patterns in the social world reflect societal norms, conventions and social structures and nothing inherently "natural" or "neutral" about individuals or social groups.

8.5 A DARK PAST: THE DISTURBING PARALLEL WITH EUGENICS AND SCIENTIFIC RACISM

> **"The needs of eugenics in large part determined the content of Galton's statistical theory"**
>
> Donald A. MacKenzie (Edinburgh University)[255]

Eugenics and scientific racism refer to the scientifically false ideologies that racial differences and hierarchies are biologically determined and fixed.[256] These ideologies have been used historically to justify slavery, colonialism and imperialism; and to support genocide, racial domination and discrimination, including racial segregation, enforced sterilization and anti-miscegenation laws.[257]

Eugenics aimed to purportedly improve the genetic quality of a population by inhibiting the reproduction of humans considered racially "inferior" – including people with mental and physical disabilities. Disturbingly,

[248] Mel Andrews and others (previously cited) point out that "the use of Machine Learning or its products in the design of real-world interventions or the architecture of public facing technologies should, in general, be understood as cementing causal interpretations of the outputs of such models. Acting on model outputs is de facto causal interpretation."
[249] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited); Janneke Gerards and others, Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-discrimination Law, 2021, p. 40.
[250] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[251] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[252] Cristian S. Calude and Giuseppe Longo, "The deluge of spurious correlations in big data" (previously cited).
[253] Sandra Wachter, "The theory of artificial immutability: protecting algorithmic groups under anti-discrimination law", 2022, http://arxiv.org/abs/2205.01166 (accessed 24 June 2024).
[254] Wendy Hui Kyong Chun and Alex Barnett, Discriminating Data (previously cited); danah boyd, "Questioning the legitimacy of data" (previously cited); Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023.
[255] Donald A. MacKenzie, Statistics in Britain, 1865-1930: The Social Construction of Scientific Knowledge, 1981.
[256] This idea has been firmly rejected in international human rights law as scientifically false, socially unjust and dangerous. See, for example, ICERD, preamble; UNESCO, "Four statements on the race question", 1969, https://unesdoc.unesco.org/ark:/48223/pf0000122962
and Durban Declaration and Programme of Action, 2001, UN Doc. A/CONF.189/12, Chapter I, para. 7.
[257] Report of the Working Group of Experts on People of African Descent, 2 August 2019, UN Doc. A/74/274; David Olusoga & Casper W. Erichsen, The Kaiser's Holocaust: Germany's Forgotten Genocide, 2011; American Psychological Association, "Apology to people of color for APA's role in promoting, perpetuating, and failing to challenge racism, racial discrimination, and human hierarchy in U.S.", 2021, https://www.apa.org/about/policy/racism-apology (accessed on 9 December 2025); Cummings Center for the History of Psychology, "Historical Chronology: examining psychology's contribution to the belief in racial hierarchy and perpetuation of inequality for people of color in U.S.", 2021, https://www.apa.org/about/apa/addressing-racism/historical-chronology.pdf

the goal of making predictions about humans and their behaviour without departing from theory has been a key motivating factor for founding figures in the field of statistics such as Francis Galton, Ronald Fischer and Karl Pearson. They are mainly known for laying the ground for modern statistics by formalizing concepts like statistical correlation and the correlation coefficient. The mathematical instruments they developed were explicitly designed with the goal of advancing eugenics.[258]

> **"[T]he eugenics movement slowly constructed a national bureaucratic and juridical infrastructure to cleanse America of its 'unfit'. Specious intelligence tests, colloquially known as IQ tests, were invented to justify incarceration of a group labelled 'feebleminded'. Although much of the persecution was simply racism, ethnic hatred and academic elitism, eugenics wore the mantle of respectable science to mask its true character."**
>
> Edwin Black[259]

Ironically, these racist scientists promoted the objectivity of the numbers and methods they worked with. They avoided the use of theory and advocated for "letting the numbers speak for themselves", much in the same fashion as some speak of data and technology today. By doing this, they obfuscated their racist ideas and presented numbers as devoid of political values. Pearson, famously known for his formalization of the correlation coefficient, published an article where he attempted to measure the intelligence of Jewish refugees and migrants in the UK. In this article, he delineated an "objective" path to eugenics,[260] under "the cold light of statistical inquiry".[261] Even though they explicitly and outspokenly supported racist politics and developed mathematical tools for the specific purpose of advancing these politics, Galton, Pearson and other eugenicists claimed objectivity and believed themselves to have "no political, no religious and no social prejudices".[262] Eugenic theories failed to stand against empirical and conceptual scrutiny.

Much like eugenicists, the field of ML sometimes operates under a blanket of assumed objectivity, without recourse to theory, and by relying only on correlation while ignoring causal mechanisms.[263] It would be cynical to attribute these ML practices to bad faith. Instead, academic literature on this topic suggests that accepted norms in ML include believing that "the algorithmic discovery of correlations in increasingly large datasets, is sufficient to count as scientific knowledge".[264] This approach is strikingly similar to that of Karl Pearson:

---

[258] Lee Kennedy-Shaffer, "Teaching the difficult past of statistics to improve the future", January 2024, Journal of Statistics and Data Science Education, Volume 32, Issue 1, https://www.tandfonline.com/doi/full/10.1080/26939169.2023.2224407
Ruth Schwartz Cowan, "Francis Galton's statistical ideas: the influence of eugenics", December 1972, Isis, Volume 63, Issue 4, https://www.journals.uchicago.edu/doi/abs/10.1086/351000
Samuel Dodson and Jane Bartley, "(Disrupting) continuities between eugenics and statistics: a critical study of regression analysis", October 2024, Proceedings of the Association for Information Science and Technology, Volume 61, Issue 1, https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.1131
Alain Desrosières, The Politics of Large Numbers (previously cited); Theodore M. Porter, Trust In Numbers (previously cited); Donald A. MacKenzie, Statistics in Britain, 1865-1930: The Social Construction of Scientific Knowledge (previously cited).
[259] Edwin Black, War Against the Weak: Eugenics and America's Campaign to Create a Master Race, 2004, p. 15.
[260] Lee Kennedy-Shaffer, "Teaching the difficult past of statistics to improve the future" (previously cited).
[261] Karl Pearson and Margaret Moul, "The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children", 1925, Annals of Eugenics, Volume 1, https://onlinelibrary.wiley.com/toc/20501439/1925/1/1
[262] Karl Pearson and Margaret Moul, "The problem of alien immigration into Great Britain" (previously cited).
[263] Jérémie Sublime, "The return of pseudosciences in artificial intelligence" (previously cited); Janneke Gerards and others, Algorithmic Discrimination in Europe (previously cited).
[264] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).

> **"It is this conception of correlation between two occurrences… which is the wider category by which we have to replace the old idea of causation"**
>
> Karl Pearson[265]

Galton, along with other eugenicists, is also cited as part of the positivist movement in Criminology. Like the eugenicists, positivist criminologists thought that the behaviour of "criminal types" was biologically determined rather than chosen, and that certain attitudes like "feeble-mindedness", "criminality" and "fraud", were genetically linked to certain racial groups.[266] Adolphe Quetelet believed that it would be possible to discover statistical regularities for both normal and abnormal behaviour.[267] Positivist criminologists believed in an "abnormal" criminal type, a specific type of people that differed from "normal people" in ways that were biologically fixed and thus outside of their control.[268]

Subsequent generations of criminologists rejected the notion that differences in attitudes were innate (racial) traits. Rather, they saw differences as the result of a combination of factors, including "social, psychological, political, economic and geographic ones – rarely, if ever, biology on its own".[269]

It is disturbing that such methods to score and rank individuals have returned to policy, even if they are now framed as tools to select people for enforcement and scrutiny rather than for eugenic aims. Whereas eugenics explicitly associated race with criminality, this association is often an indirect effect in algorithmic risk profiling. People belonging to marginalized groups are often disproportionately targeted by risk profiling algorithms even though this is the result of using other, seemingly unrelated criteria. As with eugenics, however, the link between individuals and "risk" is only correlative and not based on causal inference. These systems also assume that properties such as being "risky" are objective, measurable traits, instead of recognizing that defining and measuring such concepts in any robust way is, in practice, not feasible.

8.6 THE LIMITS OF PREDICTION IN COMPLEX SOCIAL SYSTEMS

> **"In complexity science terms, human beings and their behaviour are complex adaptive phenomena whose precise pathway is simply unpredictable"**
>
> Abeba Birhane (AI Accountability Lab, Trinity College Dublin)[270]

Certain prediction tasks cannot be solved by ML. These are settings in which no AI developed for the task can ever possibly work.[271] Such settings include the use of risk profiling for attempting to predict criminality, life course or the propensity to commit social security fraud at the level of a location or an individual.[272] Therefore, this type of system should be regarded as fundamentally dubious.

---

[265] Theodore M. Porter, The Rise of Statistical Thinking, 1820-1900, 1986, p. 110.
[266] Eamon Carrabine and others, Criminology: A Sociological Introduction, 2020, p. 49.
[267] Eamon Carrabine and others, Criminology: A Sociological Introduction, 2020, p. 59.
[268] John P. Jackson Jr and Nadine M. Weidman, "The origins of scientific racism", 2005, The Journal of Blacks in Higher Education; Lee Kennedy-Shaffer, "Teaching the difficult past of statistics to improve the future" (previously cited); Wendy Hui Kyong Chun and Alex Barnett, Discriminating Data (previously cited); Alain Desrosières, The Politics of Large Numbers (previously cited).
[269] Eamon Carrabine and others, Criminology: A Sociological Introduction, 2020, p. 62.
[270] Abeba Birhane, Automating Ambiguity (previously cited); Alicia Juarrero, "Dynamics in action: intentional behavior as a complex system", June 2000, Emergence, Volume 2, Issue 2, http://www.tandfonline.com/doi/abs/10.1207/S15327000EM0202_03
[271] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited).
[272] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited).

These systems have raised a distinctive and serious set of normative concerns that causes them to fail on their own terms, in that they do not deliver accurate predictions.[273] Regrettably, whether an AI system works *at all* is an often-overlooked question: "Deployed AI systems often do not work. They can be constructed haphazardly, deployed indiscriminately, and promoted deceptively".[274] Instead, data scientists and policymakers focus on aligning prediction models with "ethical values", ignoring the question of whether a system functions or provides any benefits at all.[275]

8.6.1 COMPLEX SYSTEMS

Mathematical models, including those used in AI applications, have inherent limits in predicting the outcome of complex systems. The human brain, natural language, intentional behaviour and social systems can all be defined as complex systems.[276] For these types of systems, or "even single traits of these systems",[277] it is extremely hard or impossible to build mathematical models that can accurately predict their outcomes. There are some rare exceptions in specific subfields, such as infectious diseases or the pharmacodynamics of antibiotics;[278] or some time series-based prediction tasks such as regularities in human mobility trajectories – even though this remains limited and computational scaling has diminishing results.[279]

Governments implementing digital transformation increasingly view people as collections of measurable data points in a way that does not respect their unique, subjective qualities and acknowledge their innate capacity to make self-determined decisions in complex social contexts. Human beings "are social through and through" and act in a non-deterministic way. They are not "stationary entities that can be captured in neat taxonomies, rather they are active, dynamic, historical, social, cultural, gendered, politicized and contextualized organisms"[280]. Seeing people in abstract and formal ways has benefits but also creates blind spots in understanding how the social world works.[281]

FUNDAMENTALLY DUBIOUS SYSTEMS

Abeba Birhane, a member of the UN Advisory Body on Artificial Intelligence, drew from complexity science and embodied cognitive science to "examine the shaky scientific foundations of machine prediction of complex behaviour", concluding that doing so with precision "is impossible in principle".[282] Among such systems are the debunked class of "physiognomic artificial intelligence" systems, which attempts to infer personal characteristics from data about someone's physical appearance; or emotion recognition systems. In these cases, there is no plausible connection between observable data and the proposed goal of the prediction system.[283]

PRACTICALLY UNFEASIBLE SYSTEMS

Other systems may not be fundamentally dubious but are unfeasible in practice. Certain types of systems, such as predictive policing systems attempting to predict crime at the level of location or the individual level,[284] are unfeasible in practice because regardless of the amount of data, there is no good enough, or objective enough, (proxy) data to adequately model the underlying phenomenon: "The data that would be required to do the task properly", Deborah Raji and colleagues explain, "does not and will never exist".

---

[273] In a landmark 2024 publication titled Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy, Angelina Wang and others took a systematic look at algorithms that use ML to predict future outcomes about individuals. The authors reviewed 387 reports, articles, and web pages from academia, industry, non-profits, governments, and data science contests, including many real-world examples of predictive optimization. They reported a pattern of similar shortcomings in applications of predictive AI. See also https://predictive-optimization.cs.princeton.edu/
[274] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited).
[275] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited), p. 1.
[276] Paul Cilliers, Complexity and Postmodernism, 2002.
[277] Jobst Landgrebe and Barry Smith, Why Machines Will Never Rule the World, 2022.
[278] "There are important exceptions in some specific subfields, for example models of certain features of monogenetic and of infectious diseases or of the pharmacodynamics of antibiotics". Jobst Landgrebe and Barry Smith, Why Machines Will Never Rule the World, 2022.
[279] En Xu and others, "Predictability of complex systems", April 2026, Physics Reports, Volume 1166, https://linkinghub.elsevier.com/retrieve/pii/S0370157326000153
[280] Abeba Birhane, Automating Ambiguity (previously cited).
[281] Alain Desrosières, The Politics of Large Numbers (previously cited); Theodore M. Porter, Trust in Numbers (previously cited); James C. Scott, Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed, 2020; Barbara Kiviat, "The moral affordances of construing people as cases: how algorithms and the data they depend on obscure narrative and noncomparative justice", September 2023, Sociological Theory, Volume 41, Issue 3, https://doi.org/10.1177/07352751231186797
[282] Abeba Birhane, Automating Ambiguity (previously cited).
[283] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited).
[284] Inioluwa Deborah Raji and others, "The fallacy of AI functionality" (previously cited); Andrew Guthrie Ferguson, The Rise of Big Data Policing (previously cited).

This is not to say that "people and social systems wander aimlessly without pattern, habit, or relatively stable behaviour".[285] Specific social phenomena in constrained settings, like the amount of traffic on a route, or how busy a shop will be on a certain day, can be predicted reasonably well.[286] But these relative stabilities and behavioural patterns do not mean that individuals are fully knowable and predictable with precision. Any prediction of future human behaviour based on data collected in the past is at best "a statistical probability".[287] Abeba Birhane sees machine prediction as a way of forcing deterministic, "law-of-nature-like" rules unto the inherently complex and dynamic systems that are human beings and the social world at large.

8.6.2 LIFE COURSE PREDICTION

In a mass scientific collaboration published in 2020, hundreds of social science researchers attempted to predict six types of life outcomes using data covering a period of fifteen years.[288] The life outcomes being predicted included a child's academic performance and whether a family would be evicted from their home. Despite an enormous amount of data and thousands of different models, using different statistical techniques, developed by hundreds of researchers, none of the models could make an accurate prediction. The researchers draw conclusions for policymakers in settings "such as a criminal justice and child-protective services"; domains where risk profiling is often employed. The results, the researchers add, "cannot be dismissed because of concerns about the limitations of any particular researcher or method".[289] The data for this study was collected specifically to enable social science research, which means that, contrary to data used in governmental risk profiling models, it was of high quality and particularly suited for the purpose. "The results raise questions about the absolute level of predictive performance that is possible for some life outcomes, even with a rich dataset".[290]

In a 2024 follow-up article exploring the "origins of unpredictability in life outcome prediction tasks", researchers collected in-depth qualitative data from participants to the above-mentioned 2020 longitudinal mass study. The researchers concluded that "unpredictability should be expected in many life outcome prediction tasks, even in the presence of complex algorithms and large datasets." Moreover, the administrative data at the disposal of government developing predictive models such as risk profiles "may be less useful for prediction than those measured in surveys".[291] The implications for governments are clear: "decision makers should reorient their expectations and anticipate that life outcome predictions – generated by humans or by algorithms – may be inaccurate."[292]

8.6.3 CRIMINALITY AND RECIDIVISM PREDICTION

Predicting criminality or "criminal types" is one of the oldest applications of risk profiling (see section 8.5). A scientific consensus has been emerging that the efficacy of predictive policing applications at the level of specific locations or individuals is fundamentally dubious (see also Box in section 10.4).

The few existing empirical studies suggest that risk assessment tools for predicting recidivism have had little to no useful impact.[293] The widely used commercial risk assessment tool COMPAS,[294] which has been the object of hundreds of academic publications, has been shown to be only as accurate as predictions made by people with little or no criminal justice expertise.[295] A 2016 investigation into COMPAS by the investigative journalism outlet ProPublica has become the flagship reference on discriminatory risk assessment algorithms (see box below).[296]

[285] Abeba Birhane, Automating Ambiguity (previously cited).
[286] Arvind Narayanan and Sayash Kapoor, AI Snake Oil (previously cited).
[287] Abeba Birhane, Automating Ambiguity (previously cited).
[288] Matthew J. Salganik and others, "Measuring the predictability of life outcomes with a scientific mass collaboration" (previously cited).
[289] Matthew J. Salganik and others, "Measuring the predictability of life outcomes with a scientific mass collaboration" (previously cited).
[290] Matthew J. Salganik and others, "Measuring the predictability of life outcomes with a scientific mass collaboration" (previously cited).
[291] Ian Lundberg and others, "The origins of unpredictability in life outcome prediction tasks" (previously cited).
[292] Ian Lundberg and others, "The origins of unpredictability in life outcome prediction tasks" (previously cited).
[293] Dasha Pruss, Carceral Machines: Algorithmic Risk Assessment and the Reshaping of Crime and Punishment (previously cited), p. 6; Megan Stevenson, "Assessing risk assessment in action", 2018-2019, Minnesota Law Review, Volume 103, https://heinonline.org/HOL/Page?handle=hein.journals/mnlr103&id=313&div=&collection=, p. 303.
[294] Correctional Offender Management Profiling for Alternative Sanctions
[295] Julia Dressel and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism" (previously cited).
[296] ProPublica, "Machine bias", 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

INVESTIGATION INTO COMPAS BY PROPUBLICA

In 2016, ProPublica investigated the commercial tool COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) made by Northpointe, Inc. to discover the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

The analysis found that Black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than Black defendants to be incorrectly flagged as low risk.[297]

In forecasting who would re-offend, the algorithm correctly predicted recidivism for Black and white defendants at roughly the same rate (59% for white defendants, and 63% for black defendants) but made mistakes in very different ways. It misclassified the white and Black defendants differently when examined over a two-year follow-up period. Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts. White defendants who re-offended within the next two years were mistakenly labelled low risk almost twice as often as Black re-offenders. The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, Black defendants were 45% more likely to be assigned higher risk scores than white defendants.

8.6.4 FRAUD DETECTION AMONG SOCIAL SECURITY CLAIMANTS

Like predictive policing, social security fraud detection is fundamentally dubious and must be understood as a form of risk profiling (see Chapter 4 and section 8.2 for more details). Although public evaluations of such systems in the public domain are rare, Amnesty International has documented several failures, including some with very low predictive performance (see "effectiveness" section 10.4). In fact, Amnesty International is not aware of any publicly available evaluation of social security fraud detection systems reporting high predictive performance. Besides, high predictive performance alone is insufficient to legitimize the use of a risk profiling system under IHRL – see Chapter 10 for more details.

SUSPICION MACHINES: THE ROTTERDAM CASE

In 2023, Lighthouse Reports published an investigation into a risk profiling system used in Rotterdam, the Netherlands, to detect fraud among social security claimants. Because the municipality of Rotterdam inadvertently gave researchers access to the source code of the system, they were able to conduct one of the first detailed investigations into a risk profiling algorithm.[298]

Every year, Rotterdam carried out investigations on some of the city's 30,000 social security claimants. Rotterdam's fraud prediction system processed 315 inputs, including age, gender, language skills, neighbourhood, marital status and a range of subjective case worker assessments, to generate a risk score between 0 and 1. Between 2017 and 2021, officials used the risk scores generated by the model to rank every benefit recipient in the city on a list, with those ranked in the top 10% referred for investigation. While the exact number varied from year to year, on average the "riskiest" 1,000 social security recipients were selected for investigation. The system relies on the broad legal leeway that authorities in the Netherlands are granted to fight social security fraud, including the ability to process and profile social security claimants based on sensitive characteristics that would otherwise be protected.

Those considered by the system to be "riskier" were parents, young people, women, people with roommates, people in the lower income bracket and people with substance abuse issues. These are typically characteristics belonging to people most likely to need social security to survive in the first place.

[297] ProPublica, "How We Analyzed the COMPAS Recidivism Algorithm", 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[298] Lighthouse Reports, "Suspicion Machine Methodology", 3 March 2023, https://www.lighthousereports.com/methodology/suspicion-machine/

Internal documents show that when Rotterdam evaluated its model, they found that it is only marginally more accurate at predicting fraud than selecting people at random. AI ethicist Margaret Mitchell told Lighthouse Reports that the performance indicates the model is “essentially random guessing”. She added that it clearly failed to meet performance standards necessary to responsibly deploy it in the real world.

9. CHALLENGING STATISTICAL FIXES

While the harms of risk profiling algorithms are clear, governments are eagerly using non-binding ethical standards rather than setting the clear red lines needed to prevent harms in high-stakes contexts. These standards and guidelines typically emphasize technical measures to "fix" or remove bias from the outcomes of technical systems. Critics point out that this treats a deep social problem as a mere technical glitch, an approach that ultimately prevents real accountability.[299] Accordingly, this approach is being increasingly rejected in favour of amplifying and supporting voices from communities most affected by the algorithms' biases and harms.[300]

Indeed, when developing risk profiling systems ostensibly intended for public benefit, governments rarely consult the communities most affected by these technologies. Greta Byrum and Ruha Benjamin write that "[c]ommunities impacted by technology must be able to resist and refuse its incursions if and as they experience harm".[301] Without this right of refusal, the design and deployment of automated risk profiles will continue to reflect only the narrow and privileged interests of those in power, investors and elite technologists – rather than the actual needs of the people they claim to serve.

Amnesty International is not suggesting that attempts to ensure algorithmic fairness should generally be opposed as a matter of principle or in all situations. However, to ensure its legality, risk profiling must be compliant with IHRL, including effective mitigating measures taken to reduce its harms. Because risk profiling operates in high-stakes contexts – law enforcement, social security and migration – the danger of severe, tangible harm is real. Accordingly, mitigation measures must be demonstrably effective, which algorithmic fairness alone cannot guarantee. At best, current fairness approaches can be understood as distributing algorithmic or material harms more equally among people,[302] rather than removing the harms altogether. Without addressing the underlying societal issues, fairness approaches amount to little more than cosmetic interventions.[303] Therefore, appeals to ML fairness do not address the fundamental problems of risk profiling in high-stakes contexts. This will be further detailed in the current chapter.

[299] Rodrigo Ochigame, "The invention of 'ethical AI': how Big Tech manipulates academia to avoid regulation", 2019, The Intercept, https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/
[300] Greta Byrum and Ruha Benjamin, "Disrupting the gospel of tech solutionism to build tech justice", June 2022, Stanford Social Innovation Review, https://ssir.org/articles/entry/disrupting_the_gospel_of_tech_solutionism_to_build_tech_justice
[301] Greta Byrum and Ruha Benjamin, "Disrupting the gospel of tech solutionism to build tech justice" (previously cited).
[302] Atoosa Kasirzadeh, "Algorithmic fairness and structural injustice: insights from feminist political philosophy", 2022, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics and Society, https://dl.acm.org/doi/10.1145/3514094.3534188 (accessed 13 December 2022).
[303] Atoosa Kasirzadeh, "Algorithmic fairness and structural injustice" (previously cited).

9.1 ALGORITHMIC FAIRNESS ONLY GIVES THE ILLUSION OF CLARITY

> **"[Th]e very term algorithmic bias is a highly contested term that has been subject to strong conceptual, methodological, and epistemological criticism."**
>
> Abeba Birhane[304]

A typical statistical "fix" is the commonly used concept of "algorithmic fairness", a mathematical formalization of the concept of equality. ML fairness is an active and ongoing research field, but theoretical and empirical findings suggest that both individual and group approaches are inadequate for addressing discrimination by risk profiling in high-stakes contexts.

The dominant approach to make algorithmic outcomes more "fair" is to formalize fairness as a mathematical constraint such that the decisions will be equally distributed across groups, while losing as little predictive performance as possible.[305] For example, to obtain equal numbers of investigations among men and women, the risk profiling algorithm must be optimized such that the "demographic parity" fairness criteria is upheld. These types of formalizations "abstract away any context that surrounds these systems", in particular their social context.[306] This is an issue that further highlights the power imbalances between those who design risk profiling systems and those subjected to their harms. Several authors have highlighted the risk of "ethics-washing", or the practice of weaponizing the mere appearance of ethical behaviour and efforts to self-regulate.[307]

Algorithmic fairness as a movement "has exercised strong pressure on policymakers, regulators, and companies".[308] But it is facing increasing criticism for having been "only minimally effective at preventing harms from automated decision-making systems".[309] Critiques include not taking structural discrimination into account;[310] missing the bigger picture of the cultural, political and normative context of social data;[311] focusing attention on too narrow a set of questions;[312] and resting on simplistic and unrealistic assumptions.[313]

For example, group fairness approaches assume that "equality can be mathematically formalized; that labels of prohibited characteristics are available for every person in the dataset; and that everyone fits in a binary or

---

[304] Abeba Birhane, "Algorithmic Bias", July 2024, Open Encyclopedia of Cognitive Science, https://oecs.mit.edu/pub/b61joemo/release/1
[305] Ninareh Mehrabi and others, "A survey on bias and fairness in machine learning", 2022, http://arxiv.org/abs/1908.09635 (accessed 8 December 2025); Sam Corbett-Davies and others, "The measure and mismeasure of fairness", 2018, https://arxiv.org/abs/1808.00023
Solon Barocas and others, "Fairness and machine learning", 2023, https://fairmlbook.org/
Maarten Buyl and Tijl De Bie, "Inherent limitations of AI fairness", January 2024, Communications of the ACM, Volume 67, Issue 2, https://dl.acm.org/doi/10.1145/3624700
[306] Andrew D. Selbst and others, "Fairness and abstraction in sociotechnical systems", 2019, Proceedings of the Conference on Fairness, Accountability and Transparency, https://dl.acm.org/doi/10.1145/3287560.3287598, p. 1.
[307] Elettra Bietti, "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy", 2020, Proceedings of the 2020 Conference on Fairness, Accountability and Transparency, https://dl.acm.org/doi/10.1145/3351095.3372860
Ben Wagner, "Ethics as an escape from regulation: from 'ethics-washing' to ethics-shopping?", 2018, in Being Profiled: Cogitas Ergo Sum: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen, https://www.degruyterbrill.com/document/doi/10.1515/9789048550180-016/html (accessed 9 December 2025).
[308] Arvind Narayanan, "What if algorithmic fairness is a category error?", 2026, in Contemporary Debates in the Ethics of Artificial Intelligence, p. 77.
[309] Arvind Narayanan, "What if algorithmic fairness is a category error?" (previously cited), p. 78.
[310] Atoosa Kasirzadeh, "Algorithmic fairness and structural injustice" (previously cited).
[311] Anna Lauren Hoffmann, "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse", June 2019, Information, Communication & Society, Volume 22, Issue 7, https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573912
Arvind Narayanan, "What if algorithmic fairness is a category error?" (previously cited).
[312] Arvind Narayanan, "What if algorithmic fairness is a category error?" (previously cited).
[313] Lindsay Weinberg, "Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches", May 2022, Journal of Artificial Intelligence Research 74, http://arxiv.org/abs/2205.04460, 75–109; Sebastian Benthall and Bruce D. Haynes, Racial categories in machine learning (previously cited); Atoosa Kasirzadeh, Algorithmic Fairness and Structural Injustice (previously cited); Atoosa Kasirzadeh and Andrew Smart, "The Use and Misuse of Counterfactuals in Ethical Machine Learning", 2021, https://arxiv.org/abs/2102.05085; Maarten Buyl and Tijl De Bie, Inherent Limitations of AI Fairness (previously cited); Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian and Janet Vertesi, Fairness and Abstraction in Sociotechnical Systems (previously cited); Alex Hanna, Emily Denton, Andrew Smart and Jamila Smith-Loud, Towards a Critical Race Methodology in Algorithmic Fairness (previously cited); Arvind Narayanan, What if Algorithmic Fairness is a category error? (previously cited).

discrete group distribution".[314] The "depoliticized lens" claimed by fairness gives "an illusion of clarity". This is what makes this approach appealing, but it "runs into headwinds when actually attempting to implement it".[315] While minimizing algorithmic bias can sometimes be a useful strategy for harm reduction, bias has been described as a "red herring" for impeding more fundamental discussions of societal harms by technology and because it can obscure the inherently political nature of technology.[316]

One of the challenges that group fairness faces is its emphasis on groups rather than the individuality of a person's behaviour and rights. Even when the distribution of high-risk predictions is equal across groups, this still allows for individuals in a marginalized group to receive poor scores if enough others in the same group receive higher scores.[317] This phenomenon is known as "fairness gerrymandering".[318]

Wary of the limitations of the statistical group fairness approaches described above, researchers of "individual fairness" have tried to address the counterfactual question of "how would this individual have been treated, had they not possessed the protected ground?" This betrays misconceptions about the meaning and significance of social categories. Both group and individual fairness approaches oversimplify protected social categories such as race as mere biological traits or as discrete and one-dimensional characteristics. Instead, prohibited grounds such as race or gender are socially constructed categories shaped by historical and social processes that cannot be captured by quantitative fairness approaches.

Lastly, despite intersectional discrimination receiving an increasing amount of attention in both research and advocacy, computational fairness approaches appear particularly ineffective at capturing intersectional discrimination, with no short-term solutions in sight.[319] This means fairness approaches are unsuited to preventing or remedying intersectional or multiple forms of discrimination, which is an obligation of states under IHRL. (See Chapter 6 for the applicable legal framework.)

SPOTLIGHT: AMSTERDAM'S FAILED 'ETHICAL' FRAUD DETECTION ALGORITHM

A 2025 investigation by Lighthouse Reports,[320] MIT Technology Review[321] and the Dutch newspaper Trouw[322] interrogated a high-stakes experiment carried out by Amsterdam's social security department over the course of five years. The municipality attempted to build a "fair" algorithm to detect welfare fraud, guided by Responsible AI: a framework of technical and ethical guidelines meant to ensure fairness, transparency and accountability in automated systems.

The city of Amsterdam spent hundreds of thousands of Euros, hired consultants, spoke to academic experts, extensively audited its system for bias, and consulted with welfare recipients to provide feedback on the system's design. Despite all this effort, the system failed. When the city deployed a pilot in the real world, the system was continuously plagued by biases. It was also no more effective than the human case workers it was designed to replace. As political pressure mounted, officials ended the project, bringing an expensive, multi-year experiment to a quiet end.

[314] Maarten Buyl and Tijl De Bie, Inherent Limitations of AI Fairness (previously cited).
[315] Arvind Narayanan, What if Algorithmic Fairness is a category error? (previously cited), p. 78.
[316] Dasha Pruss and others, "Prediction and punishment" (previously cited), p. 10.
[317] Maarten Buyl and Tijl De Bie, "Inherent limitations of AI fairness" (previously cited), p. 51.
[318] Michael Kearns and others, "Preventing fairness gerrymandering: auditing and learning for subgroup fairness", 2018, Proceedings of the 35th International Conference on Machine Learning, https://proceedings.mlr.press/v80/kearns18a.html (accessed 8 December 2025).
[319] Anna Lauren Hoffmann, "Where fairness fails" (previously cited); Steven Vethman and others, "Fairness beyond the algorithmic frame: actionable recommendations for an intersectional approach", 2025, Proceedings of the 2025 ACM Conference on Fairness, Accountability and Transparency, https://dl.acm.org/doi/10.1145/3715275.3732210 (accessed 24 July 2025).
[320] Lighthouse Reports, The Limits of Ethical AI, 2025, https://www.lighthousereports.com/investigation/the-limits-of-ethical-ai/
Lighthouse Reports, How We Investigated Amsterdam's Attempt to Build a 'Fair' Fraud Detection Model, 2025, https://www.lighthousereports.com/methodology/amsterdam-fairness/
[321] MIT Technology Review, "Inside Amsterdam's high-stakes experiment to create fair welfare AI", 2025, MIT Technology Review, https://www.technologyreview.com/2025/06/11/1118233/amsterdam-fair-welfare-ai-discriminatory-algorithms-failure/ (accessed 11 December 2025).
[322] Trouw, "Amsterdam wilde met AI de bijstand eerlijker en efficiënter maken. Het liep anders" ["Amsterdam wanted to make welfare fairer and more efficient with AI. Things turned out differently"], 6 June 2025, https://www.trouw.nl/verdieping/amsterdam-wilde-met-ai-de-bijstand-eerlijker-en-efficienter-maken-het-liep-anders~b2890374/ (in Dutch).

9.2 MISTAKEN USE OF RANDOM SAMPLES AS GROUND TRUTH TO BUILD RISK PROFILES

Because of the early digitalization of the Dutch public sector, the Netherlands has witnessed a multitude of algorithmic discrimination scandals (see Chapter 7), and policy discussions in this country often pioneer the legal and methodological debates surrounding the governance of algorithms. Years of policy discussions between Dutch government authorities and NGOs, including Amnesty International, have shown that statistical testing methods are being given an increasingly prominent role in attempting to answer the question of whether risk profiling complies with human rights standards. In these discussions, random samples function as a source of "ground truth", such as for identifying the base rate of offending behaviour in the population; statistical hypothesis testing is presented as "empirical science";[323] and it is argued that statistical tests and the use of random samples can provide insight into whether there is an empirical or objective foundation for justifying the use of a criterion in a risk profile.[324]

This approach is problematic for several reasons. First, by relying on statistical tests to determine the validity of a proposed risk criterion, it engages in the same flawed "theory-free" approach that undergirds risk profiling, and indeed risks reinforcing it. The performative effects of risk profiling further compound this issue by cementing the causal effects of profiling into the data and ultimately into the lives of affected people. Second, random sampling addresses sampling bias, but societal biases will be present regardless of the sampling methodology. Third, while testing risk criteria against potential proxies for discrimination may be a necessary step, it is insufficient to protect against discrimination in practice, as it cannot account for all the ways in which structural and societal inequalities can be more subtly reflected in data, nor for how intersectional discriminatory harms may be experienced in practice. Lastly, risk profiling remains scientifically invalid despite these and other statistical "fixes", such as algorithmic fairness.

Algorithm Audit, a Dutch NGO that works closely with the Dutch government, including the Netherlands Institute for Human Rights,[325] advocates for "responsible use of profiling in the public domain". As part of its "Public Standard for profiling algorithms",[326] Algorithm Audit conducts statistical investigations to ascertain whether there is a "link"[327] between profiling criteria and risk (such as risk of fraud).[328] For example, when investigating the discriminatory risk profile used by DUO, Algorithm Audit stated: "assumptions from the profile can be tested by means of a hypothesis test: is it indeed the case that younger students are more likely to make unlawful use of the college grant than older students? What about older students who live far away from their parents?"[329] The audit concluded that "type of education" and "age" are inappropriate to use as risk criteria, "as there is no statistical support" for this.[330] The third criterion, "distance to parents", *was* found to have predictive value.[331]

When Algorithm Audit investigated how DUO formulated the risk profiling criteria, it found that the three criteria (type of education, age, and distance from parents) were "insufficiently substantiated" because of a lack of documentation.[332] It is unclear why these criteria were chosen and by whom. Instead of disqualifying the risk criteria on this basis alone, Algorithm Audit proceeded to conduct an in-depth statistical analysis, seemingly to *retroactively* investigate if the criteria in the risk profile are justified, based on statistical

[323] Algorithm Audit, "Statistical hypothesis testing: Risk management measures to mitigate the risk of indirect discrimination through high-risk AI profiling systems", 2025, https://algorithmaudit.eu/pdf-files/knowledge-base/standards/20240726_AlgorithmAudit_statistical_hypothesis_testing.pdf, p. 2.
[324] In this context and for the sake of the argument, we assume that these selections are truly randomized, unbiased and representative. It is worth noting that a careful process is needed to ensure that sampling is truly randomized and unbiased.
[325] Netherlands Institute for Human Rights, "Opsporingsalgoritmes kunnen over de schreef gaan, maar alternatieven zijn niet per se beter" ["Detection algorithms can cross the line, but alternatives are not necessarily better"], 28 March 2024, https://www.mensenrechten.nl/actueel/weblogs/interviews/2024/opsporingsalgoritmes-kunnen-over-de-schreef-gaan-maar-alternatieven-zijn-niet-per-se-beter (in Dutch).
[326] Algorithm Audit, "Public standard profiling algorithms", https://algorithmaudit.eu/knowledge-platform/knowledge-base/public_standard_profiling/ (accessed 23 March 2026).
[327] Algorithm Audit, "Empirical Methods for Supervising Algorithmic Profiling Systems: Assessment Protocol for Examining Indirect Discrimination", June 2025, p. 9.
[328] Algorithm Audit, "Empirical Methods for Supervising Algorithmic Profiling Systems: Assessment Protocol for Examining Indirect Discrimination" (previously cited), p. 9.
[329] Algorithm Audit, "Empirical Methods for Supervising Algorithmic Profiling Systems: Assessment Protocol for Examining Indirect Discrimination" (previously cited), p. 9.
[330] Algorithm Audit, "Empirical Methods for Supervising Algorithmic Profiling Systems: Assessment Protocol for Examining Indirect Discrimination" (previously cited), p. 11.
[331] Nevertheless, this criterion was strongly correlated with migration background and deemed inappropriate. Algorithm Audit, "Statistical hypothesis testing: Risk management measures to mitigate the risk of indirect discrimination through high-risk AI profiling systems" (previously cited), p. 9.
[332] Dienst Uitvoering Onderwijs, Intern Onderzoek Controle Uitwonende Beurs [Internal Investigation into Audit of Student Housing Grant], 2024, https://www.rijksoverheid.nl/documenten/rapporten/2024/03/01/intern-onderzoeksrapport-controle-uitwonendenbeurs-duo, p. 7, (in Dutch).

evidence. This is illustrative of the increasing dominance of "theory-free" statistical approaches in policy discussions (see Chapter 8 for the risks of theory-free inference).

When approached for comment, Algorithm Audit replied that its "Public Standard for profiling algorithms" includes a "qualitative" requirement of "clear and substantive relationship" between profiling criteria and risk (of committing fraud, for example).[333] This requirement is designed to avoid confusing correlation with causation. Nevertheless, the meaning of this requirement is not further specified, and the public standard stops short of requiring rigorous scientific safeguards such as excluding potential confounds or requiring deployers to explicitly specify a causal pathway grounded in a rigorously validated theory (as described in Chapter 8). Thereby, Algorithm Audit's public standard fails to critically evaluate the choice of prediction goals, criteria and their underlying theory and values.

For example, *why* should criteria such as age cause increased individual inclination to abuse a student grant? Is it possible to accurately and reliably operationalize "risk of unlawful use of a college grant"? What makes these hypotheses worthy of being pursued as research or policy objectives? Which values, and more importantly, whose values, are being served – and whose values are not? Avoiding asking these questions undermines the scientific validity of this statistical exercise and disqualifies the results as a source of evidence. It enables the masking of uninterrogated values into the outputs of the risk profiling system, which are subsequently presented as objective empirical truths and used for direct intervention in people's lives.[334] This is compounded by the performative effects of using risk profiling to enforce punitive measures (see further section 10.6 on performative effects). Statistical patterns in random samples, whether or not statistically significant and regardless of their predictive power, should be assumed to be spurious unless such rigorous scientific heuristics are observed. Therefore, this procedure cannot constitute a credible basis for evidence-based policy, let alone a justification for differential treatment.

An additional problem is that these statistical procedures operate under the assumption that random samples can serve as a "ground truth" to evaluate the suitability of profiling criteria. Statistical patterns in random samples are presented as empirical facts rather than as contingencies of complex social realities. This "agnostic" approach to administrative social data fails to address the underlying structural societal biases, because it operates at the wrong analytical level. Random sampling only addresses one type of bias: *sampling* bias. It does not address the fundamental ways in which societal, cultural, economic, and political factors systematically influence both the targeted behaviour and the collection, measurement and interpretation of data by governments. Societal biases caused by these factors are present regardless of the sampling methodology.

In Algorithm Audit's "Public Standard for profiling algorithms", profiling criteria validated as "predictive" are screened against "strong proxy relationships" with prohibited grounds.[335] While this method can check for proxy relationships between single variables, it fundamentally misrepresents the complex and multifaceted social reality that marginalized individuals face, which leaves the risk of more complex and subtle forms of structural bias and discrimination unaddressed. Even if risk profiling systems are built on correlations inferred in random samples, they still inherit and amplify a complex web of societal biases that systematically distorts outcomes and perpetuates existing inequalities. (See further section 10.2.) Algorithm Audit acknowledges this risk: "profiling characteristics always have a proxy character to a greater or lesser extent. Making a distinction based on proxy characteristics is not necessarily prohibited. However, it must be possible to objectively justify its use."[336]

However, when governments use risk profiling in high-stakes contexts and cause a differential treatment of groups based on race, ethnicity or other suspect grounds, this will almost automatically result in indirect discrimination, even if based on seemingly neutral criteria. This test is thus "strict in theory, fatal in practice". (See legal framework for more details). Therefore, it is virtually impossible to justify such a differential treatment, regardless of statistical patterns and of their statistical significance. If it is impossible to rule out that the variables used in risk profiling function as proxies for prohibited characteristics, this casts

[333] Algorithm Audit, "Public standard profiling algorithms" (previously cited).
[334] "…hidden philosophical commitments permeate model interpretation practices. When engineers treat model predictions as revelations about underlying reality rather than artifacts of particular training procedures, they commit to a form of naïve realism that ignores the theory-ladenness of ML outputs. Simply, models trained on biased data reproduce those biases not as unfortunate technical artifacts but as assertions about the world. The epistemological error lies in treating models as neutral observers rather than constructed instruments embodying specific perspectives and limitations."; from M. Z. Naser, "On the philosophical naivety of engineers in the age of machine learning", 25 November 2025, Topoi, ahead of print, https://doi.org/10.1007/s11245-025-10304-2, p. 7. See also Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 6.
[335] Ultimately, "distance to parents" was deemed "inappropriate" because of a "strong proxy character" with the migration background of students.
[336] Algorithm Audit, "Statistical hypothesis testing: Risk management measures to mitigate the risk of indirect discrimination through high-risk AI profiling systems" (previously cited), p. 5.

serious doubt on whether it is practically feasible to deploy such systems in a way that does not inherently produce discriminatory outcomes. This issue is further discussed in Chapter 10.

Finally, risk or propensity to criminality or fraud are fundamentally unmeasurable constructs that are extremely difficult to reliably operationalize (discussed in section 8.2). This is true regardless of how the training data is obtained. Predicting criminality or social security fraud, even if based on randomized training data, remains a fundamentally fraught exercise.

Such procedures therefore give risk profiling a false allure of objectivity by reducing rigorous scientific heuristics and evaluation to statistical hypothesis testing based on random sampling. Moreover, they are not in line with international human rights standards on non-discrimination. The core problem lies in treating statistical methodology as a substitute for addressing underlying social and structural biases. One valid way of using random sampling to address concerns of discrimination is by employing it as an alternative selection means, as a substitute to risk profiling (see box in section 10.6).

9.3 OVER-RELIANCE ON STATISTICS TO PROVE NON-DISCRIMINATION

Statistical measures are important for researching discrimination. However, technical approaches grounded in fairness concepts too often fail to recognize that, while statistical evidence can sometimes be used to establish *prima facie* discrimination – that is, the suspicion that a differential treatment is discriminatory – it is an entirely different question to demonstrate with certainty that a measure does not discriminate.

Courts of law, including the ECtHR, have indeed accepted statistics as part of the evidence required to prove *prima facie* discrimination. However, analysis of jurisprudence reveals that courts accept statistical evidence rarely and inconsistently in equality and non-discrimination cases.[337] Reasons for this include fears that requiring statistical evidence could result in a "battle of numbers" that implicitly favours parties capable of producing convincing statistics.[338]

An "overreliance on statistics" or statistics-driven fairness approaches "can also undermine efforts to establish equality in areas where relevant statistics do not exist, but where potential discrimination is "fairly obvious as a matter of common sense".[339] For example, targeting people on a low income but with high healthcare costs for potential tax evasion leads to disproportional scrutiny of people with a chronic illness. This can be deduced with common-sense reasoning and without needing to resort to complicated statistical approaches. It is also worth noting that there is a general lack of reliable data disaggregated on the basis of race or ethnicity.[340]

Amnesty International is unaware of case law in which a court has accepted statistical evidence as a reasonable and objective justification for differential treatment by means of risk profiling. The complexity of establishing whether a risk profiling algorithm is "fair" means that "some form of bias will always be missed"[341] because it remains outside of scope. This "opens the door to abuse by whomever designs" risk profiling algorithms.[342] Moreover, if the prediction system is fundamentally dubious, malfunctional or invalid, the outcomes will be misleading and potentially harmful even if they are equally distributed across groups. For example, algorithms can still predict the wrong quantity even if they satisfy fairness criteria.[343]

---

[337] This reflects "a general reluctance among legal scholars to require statistical evidence to establish prima facie discrimination". See Sandra Wachter and others, "Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI", 2020, SSRN Electronic Journal, ahead of print, https://doi.org/10.2139/ssrn.3547922
See, for example, ECtHR, *Seydi and Others v. France* (previously cited) for a recent case where statistics on the prevalence of racial profiling on the general population in France were deemed insufficient to raise an objective and concrete suspicion that a police control was racially motivated in an individual complainant's case.

[338] CJEU, *Nolte v. Landesversicherungsanstalt Hannover*, Opinion of Advocate General Léger, Case C-137/93, 31 May 1995, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:61993CC0317
Sandra Wachter and others, "Why fairness cannot be automated" (previously cited).

[339] Sandra Wachter and others, "Why fairness cannot be automated" (previously cited), p. 34.

[340] This point has been made by multiple UN special procedures; see, for example, OHCHR, Promotion and Protection of the Human Rights and Fundamental Freedoms of Africans and of people of African Descent Against Excessive Use of Force and Other Human Rights Violations by Law Enforcement Officers (previously cited); Working Group of Experts on People of African Descent, Report on Its Twenty-third and Twenty-Fourth Sessions, 2019, UN Doc. A/HRC/42/59; International Independent Expert Mechanism to Advance Racial Justice and Equality in Law Enforcement, Report on the Promotion and Protection of the Human Rights and Fundamental Freedoms of Africans and of People of African Descent Against Excessive Use of Force and Other Human Rights Violations by Law Enforcement Officers, 2022, UN Doc. A/HRC/51/55.

[341] Maarten Buyl and Tijl De Bie, "Inherent limitations of AI fairness" (previously cited), p. 54.

[342] Maarten Buyl and Tijl De Bie, "Inherent limitations of AI fairness" (previously cited), p. 54.

[343] Amanda Coston, "Falsifying predictive algorithms" (previously cited), p. 4.

10. RISK PROFILING IS INHERENTLY DISCRIMINATORY

Despite the promises made of risk profiling systems in public policy, examples of their failures are numerous and there is remarkably little published evidence of their effectiveness in practice. This raises the question of whether risk profiling can ever be safely deployed in high-stakes contexts.

To answer this question, this chapter details the human rights impacts of risk profiling systems and how these systems must be tested against the prohibition of discrimination, including by taking the scientific evidence (cited in Chapter 8) and the reported ineffectiveness of mitigating measures (Chapter 9) into account.

10.1 HUMAN RIGHTS IMPACTS OF RISK PROFILING

10.1.1 EQUALITY AND NON-DISCRIMINATION

According to the FRA guidelines on profiling, profiling involves significant risks in contexts such as law enforcement and border control.[344] Together with social security, these are the domains identified as "high-stakes" by Amnesty International. Profiling

"establishes general correlations that may not be true for each individual. Any given individual may be the exception to the rule. Profiles may generate incorrect correlations, both for specific individuals and for groups [of individuals based on real or perceived shared attributes]. Although the predictions or profiles are not indicative of actual wrongdoing, they can have far reaching consequences for targeted individuals that reverberate far beyond the initial prediction, eventually leading to formal suspicions or other severely adverse consequences. Profiling can create harmful stereotypes and lead to discrimination".[345]

Even when profiles reflect a statistical fact, they can be problematic if this entails "stereotyping individuals as members of a group".[346]

Accordingly, the impact on the right to equality and non-discrimination is one of the most frequently reported outcomes of risk profiling on human rights. Because non-discrimination is a cross-cutting issue inherent in most applications of risk profiling for enforcement, the expansion of the use of risk profiling systems to an increasing number of real-world domains has discriminatory impacts on a range of associated human rights, such as the right to a fair trial and the presumption of innocence, the right to privacy, the rights to freedom of

[344] FRA, Preventing Unlawful Profiling Today and in the Future (previously cited). p. 17.
[345] FRA, Preventing Unlawful Profiling Today and in the Future (previously cited). p. 16
[346] FRA, Preventing Unlawful Profiling Today and in the Future (previously cited), p. 17.

peaceful assembly and of association, and economic, social and cultural rights, such as the right to social security and the right to an adequate standard of living.[347]

POLICING

Predictive policing systems are the modern face of racial profiling. Even if race or ethnicity is not part of the explicit list of criteria in a profile, these practices have been shown to disproportionately impact racially and ethnically marginalized groups, often amounting to indirect discrimination.[348]

Indeed, the use of risk profiling has led to racial profiling and discrimination in several countries (see Chapter 7). Risk profiling can result in both direct and indirect discrimination against racialized people, migrants and refugees, people living with disabilities, people from socio-economically disadvantaged backgrounds and other marginalized groups. The risk of "perpetuating or even enhancing discrimination, reflecting historic racial and ethnic bias" is inherent to predictive tools, according to OHCHR.[349] The UN Special Rapporteur on racism has noted that AI systems that classify, differentiate, rank and categorize are "systems of discrimination" because they "reproduce bias embedded in large-scale data sets… even in the absence of explicit algorithmic rules that stereotype".[350] As is stressed throughout this report, the discriminatory outcomes of risk profiling reach beyond human bias and individual rights, as they enable consolidation of more structural types of discrimination.

Several warnings and calls for bans on predictive policing have been issued by a coalition of CSOs including Amnesty International,[351] UN Special Procedures,[352] OHCHR,[353] and equality and human rights bodies.[354] Many have been officially acknowledged by governments or supranational organizations. In some cases, this has led to some form of regulation or partial bans on predictive policing, for example in Article 5 of the European Union AI Act.

SOCIAL SECURITY

Amnesty International's research has found that the introduction of digital technologies into social security systems has, in many cases, led to hardship for social security claimants.[355] This has negatively affected the realization of claimants' human rights, including their rights to social security and an adequate standard of living, both of which are enshrined in IHRL. (See Chapter 6 for the full legal framework.)

Although risk profiling systems are often cited as a method by which states can streamline social services, improve their cost-effectiveness and prevent fraud, a more common outcome is the penalization of society's most marginalized groups for attempting to access their rights and/or essential services.[356] For example, many governments have adopted some form of automated or machine-enabled decision-making in tools for managing or making decisions around whether an individual qualifies for government assistance. These systems have been shown to disproportionately associate people who already experience one or multiple forms of discrimination or marginalization with higher criminal or financial risk.[357]

Governments justify the use of risk profiling technology in the prediction or detection of social security fraud with promises of increased effectiveness and efficiency of enforcement policies. This is akin to what Virginia

---

[347] Amnesty International, *UK: Automated Racism* (previously cited); Amnesty International, *Digitally Divided* (previously cited).
[348] See, for example, Amnesty International, *UK: Automated Racism* (previously cited); Amnesty International, *Netherlands: We Sense Trouble* (previously cited).
[349] OHCHR, "The right to privacy in the digital age", 13 September 2021, UN Doc. A/HRC/48/31.
[350] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis" (previously cited), para. 7.
[351] Amnesty International, "EU policymakers: regulate police technology, civil society calls on the EU to draw limits on surveillance technology in the Artificial Intelligence Act", 21 September 2023, https://www.amnesty.eu/news/eu-policymakers-regulate-police-technology/
Amnesty International, Council of Europe: Amnesty International's Recommendations on the Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (Index: TIGO IOR 10/2024.5404), 11 April 2024, https://www.amnesty.eu/news/council-of-europe-amnesty-internationals-recommendations-on-the-draft-framework-convention-on-artificial-intelligence-human-rights-democracy-and-the-rule-of-law/
Amnesty International, Artificial Intelligence and Judicial Systems: Submission to the UN Special Rapporteur on the Independence of Judges and Lawyers (Index: IOR 40/9316/2025), 7 May 2025, https://www.amnesty.org/en/documents/ior40/9316/2025/en/
[352] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis" (previously cited); UN Special Rapporteur on racism, Report, 3 June 2024, UN Doc. A/HRC/56/68.
[353] OHCHR, Open Letter from the United Nations High Commissioner for Human Rights to European Union institutions on the European Union Artificial Intelligence Act ("AI Act"), 2023.
[354] European Network of Equality Bodies and European Network of National Human Rights Institutions, Joint Equinet and ENNHRI Statement on EU Artificial Intelligence Act Trilogue, 2023, https://ennhri.org/wp-content/uploads/2023/11/Joint-ENNHRI-and-Equinet-Statement-on-EU-AI-Act-Trilogue.pdf
[355] Amnesty International, *Too Much Technology, Not Enough Empathy* (previously cited).
[356] Amnesty International, *Digitally Divided* (previously cited).
[357] Amnesty International, *Digitally Divided* (previously cited); Amnesty International, *Netherlands: Xenophobic Machines* (previously cited); Amnesty International, *Trapped by Automation* (previously cited); UN Special Rapporteur on extreme poverty and human rights, Amicus Curiae Before the District Court of the Hague on the Case of NJCM c.s./De Staat der Nederlanden (SyRI) (previously cited).

Eubanks coined "scarcity bias": the idea that, because financial resources to help people living in poverty are constrained, we need a technological solution to use the resources efficiently and fill the gaps.[358] Eubanks debunks "scarcity" as a policy rationale by arguing that it is an empirically unsupported and politically useful assumption that turns poverty from a political problem into an "efficiency" problem solved through automated rationing and surveillance. Automated welfare systems do not merely manage limited resources but institutionalize austerity by designing hardship into access to basic needs.

The "scarcity bias" argument, together with an increase in xenophobic attitudes, a general shift to pre-crime or pre-emptive enforcement, and a routinely exaggerated political discourse framing welfare fraud as a prevalent problem among racialized and marginalized communities, have contributed to the widespread and entrenched conviction that there is no alternative to risk profiling when enforcing border, fraud or criminality checks.

MIGRATION

The use of risk profiling systems in migration and border control can lead to racial and ethnic profiling and discriminatory denial of visas to people based on their real or perceived ethnicity, race, national origin, descent, religion and other characteristics, often on the false assumption that individuals of certain nationalities or with certain characteristics pose a risk of non-compliance with immigration policies or a threat to national security.[359] A coalition of 163 CSOs, including Amnesty International, has called on the European Parliament, the European Commission, the Council of the European Union and EU member states to protect the rights of all people in AI regulation, irrespective of their migration status.[360]

RULE-BASED RISK PROFILING

Discrimination can occur through rule-based algorithms. Human biases, prejudices and stereotypes may influence the outcomes of the algorithm.[361] Although rule-based algorithms have the advantage of being more accessible and easily audited, they can still contribute to the presumption of neutrality often associated with technological systems, and mask otherwise obvious discrimination.[362]

The question of automatically excluding certain systems from AI regulation based on technical implementation raises significant concerns about potential loopholes, enforcement and alignment with international legal norms. Research shows that neural networks can be converted into functionally equivalent decision trees or rule-based systems. This poses a fundamental challenge: developers could bypass regulation by converting AI systems into rule-based versions with the same functionality and risks.

Amnesty International advocates for regulation that focuses on potential harm, not just technical design. OECD guidelines support this by advocating for a flexible, inclusive definition of AI, covering systems from simple to complex.[363] Amnesty International recommends that all algorithmic and predictive systems fall under the scope of AI-regulating legislation. This aligns with international legal norms, placing the burden on relevant actors to demonstrate their qualification for any exemptions. The harm caused by simple systems, like the SyRi system in the Netherlands, highlights the need for comprehensive regulation.[364] Technical implementation should not serve as a basis for automatic exclusion from oversight.

PRIVACY AND DATA PROTECTION

Interferences with the right to privacy and data protection must be lawful and necessary under international and regional human rights and data protection frameworks. The use of risk profiling exposes targeted individuals and groups to heightened surveillance, checks and policing activity, with a greater possibility of

---

[358] Virginia Eubanks, Automating Inequality (previously cited).

[359] UN Special Rapporteur on racism, Racial and Xenophobic Discrimination and the Use of Digital Technologies in Border and Immigration Enforcement, 2021, UN Doc. A/HRC/48/76; Amnesty International, *Primer: Defending the Rights of Refugees and Migrants in the Digital Age* (Index: POL 40/7654/2024), 5 February 2024, https://www.amnesty.org/en/documents/pol40/7654/2024/en/
Amnesty International, *The Digital Border* (previously cited).

[360] Amnesty International and others, Joint letter: EU: AI Act Must Protect All People, Regardless of Migration Status, 2022, https://www.amnesty.eu/news/eu-ai-act-must-protect-all-people-regardless-of-migration-status/
Amnesty International, Open Letter to the Rapporteurs on the EU Artificial Intelligence Regulation (AI Act) to Ensure Protection of Rights of Migrants, Asylum Seekers and Refugees, 2023, https://www.amnesty.eu/news/the-eu-must-respect-human-rights-of-migrants-in-the-ai-act/
Amnesty International, *Advocacy Briefing for Defending the Rights of Refugees, Asylum Seekers, and Migrants in The Digital Age* (Index: POL 30/0290/2025), 2025, https://www.amnesty.org/en/documents/pol30/0290/2025/en/

[361] Janneke Gerards and others, Algorithmic Discrimination in Europe (previously cited), p. 27.

[362] Linda Skitka and others, "Does automation bias decision-making?", November 1999, International Journal of Human-Computer Studies, Volume 51, Issue 5, https://www.sciencedirect.com/science/article/pii/S1071581999902525

[363] See OECD, Artificial Intelligence Papers 8, Explanatory Memorandum on the Updated OECD Definition of an AI System, March 2024, https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf
OECD, OECD Recommendation on Artificial Intelligence (OECD /LEGAL/0449), 2019, amended 2023.

[364] SyRi was a Dutch government risk-profiling system that linked and analysed large volumes of personal data from different public databases to flag individuals as suspected of welfare, tax and social security fraud.

interference and intrusion by the state. These systems erode people's right to privacy, for example because authorities visit homes of individuals labelled as "risky" or "suspicious", or because of automatic flagging that is then shared with other government agencies. The stigma of suspicion or guilt can follow individuals as they interact with state services such as employment, housing, education or social security.[365]

TRANSPARENCY AND RIGHTS TO REMEDY AND REDRESS

AI systems used in law enforcement and criminal justice decision-making – through predictions, profiles and risk assessments – often resist meaningful scrutiny due to technological barriers (black boxes, neural networks) or commercial restrictions (intellectual property, proprietary technology). Additionally, the use of risk profiling in public sector decision-making is often hidden or unknown, making it difficult or impossible to know whether it has an impact on human rights.[366] This results in an asymmetry of information between those negatively affected by risk profiling systems and those developing and using such systems, and puts litigants and discriminated people at a disadvantage when defending themselves, as the "reasoning" on which the prediction of risk is based is not made available to them or their legal representatives to interrogate, undermining the principle of equality of arms.

This stresses the need to reinforce mechanisms of transparency.[367] People affected by risk profiling will only be able to challenge a decision if they know it has been made, and understand on what basis.[368] Moreover, when transparency of algorithmic decision-making systems is not ensured,[369] the existence of biases can easily remain undetected or be obscured.[370]

10.2 DIFFERENTIAL TREATMENT BY DESIGN: RISK PROFILING ENTRENCHES SYSTEMIC AND INTERSECTIONAL DISCRIMINATION

> **"Although most people talk about machine learning's ability to predict the future, what it really does is predict the past."**
>
> Ben Green (University of Michigan)[371]

This section answers two questions: whether risk profiling subjects an individual or group to differential treatment, and whether this differential treatment is based on a prohibited ground. If such a differential treatment exists, it is considered discriminatory unless there exists a reasonable and objective justification, which will be addressed in the following sections.

Risk profiling systems can cause either direct or indirect discrimination, depending on whether prohibited grounds are explicitly included in the list of profiling criteria. For example, including criteria such as race or ethnicity in a risk profile causes differential treatment that amounts to racial profiling, which is strictly forbidden under IHRL. Therefore, one common but inadequate response to bias in algorithmic systems has been to remove prohibited characteristics – such as race, nationality or gender – from datasets.[372] This approach fails to address the deeper, structural issues embedded in the data. Even when a dataset is stripped of all explicit identifiers, these characteristics will be implicitly captured in other interconnected variables, such as one's spending patterns, living arrangements, type of employment or no-shows in medical appointments. These variables – that may be protected under IHRL by the term "other status" – often serve as indirect proxies, not because of isolated correlations, but because systemic inequalities have produced meaningful disparities across entire domains of life.

---

[365] Amnesty International, *UK: Automated Racism* (previously cited); Amnesty International UK, *Trapped in the Matrix* (previously cited).
[366] Amnesty International, *Netherlands: Xenophobic Machines* (previously cited).
[367] Council of Europe Ad Hoc Committee on Artificial Intelligence, Feasibility Study, 17 December 2020, rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da, para. 85.
[368] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 2018.
[369] CERD, General Recommendation 36 (previously cited), para. 64.
[370] Council of Europe Ad Hoc Committee on Artificial Intelligence, Feasibility Study (previously cited), para. 30.
[371] Ben Green, The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future, 2019.
[372] UN Special Rapporteur on racism, "Racial discrimination and emerging digital technologies: a human rights analysis" (previously cited). p. 10.

Indeed, commonly cited examples of "proxy variables" are oversimplified, because they often focus on individual variables that are seen as direct indicators of a prohibited ground, such as the well-known direct relationship between postcode and race. These direct correlations, termed "univariate correlations", can exist, but they only begin to demonstrate a much more complex reality. These approaches fail to recognize that group membership is not encoded in just one or two highly correlated variables, but is reflected diffusely across a much wider array of interconnected data points. Discrimination is not an anomaly in the data – it is "structurally engrained",[373] shaped by intersecting systems of inequality that persist across generations and institutions (see for example box below).

This reveals a rather uncomfortable truth: base rates – that is, the underlying prevalence of seemingly relevant characteristics – are unequally distributed across groups. In other words, the rate at which criteria that may appear relevant and legitimate for risk profiling at face value are possessed by members of marginalized groups is unequal.[374] "[I]f you wanted to remove everything correlated with race, you couldn't use anything".[375] Such differences in base rates can reflect historical and ongoing discrimination at an institutional or systemic level.[376]

For example, in addressing a racial profiling case in Brazil, the IACHR referred to the often disadvantaged socio-economic position of racialized people in Brazil,[377] and admonished the state for having "no respect for the special situation of belonging to a group that is considered vulnerable (of African descent, poor, living in a favela)".[378] In general, the IACHR has found that Afro-descendants in the Americas suffer from a situation of structural discrimination, evidenced in indicators relating to poverty, political participation, contact with the criminal justice system, and access to quality housing, healthcare and education.[379] Structural discrimination is also reflected in continued stereotyping of and prejudice against persons of African descent. The IACHR described these structural conditions as inseparable from issues of racial profiling.[380]

Similarly, in a risk profiling context, if police or social security authorities profile on indicators like "living in high-crime areas", "unstable employment", "informal housing", or "prior contact with the criminal justice system", this will likely result in differential treatment on the grounds of race. On paper, none of these criteria mention race. But because of structural discrimination, members of certain racialized or otherwise marginalized groups are often statistically more likely to live in high-crime areas, to have precarious jobs, and to be stopped and arrested. So even though race is not explicitly used, the base rate of these "neutral" indicators is higher in those groups.

THE PERVASIVENESS AND INTERSECTIONALITY OF DISCRIMINATION

Racial discrimination has long-term consequences across all aspects of people's lives, as a 2025 survey of almost 10,000 Muslims by the FRA showed.[381] Young Muslims are more likely than non-Muslims to leave school early, hampering their employment opportunities later in life. High numbers of Muslims are in temporary, short-term jobs, which lack security and stability. On the other hand, those who are educated also have difficulties finding suitable work, with many being overqualified for their jobs.

In 2025, one third of Muslims in Europe who were looking for a place to live struggled to find suitable homes for their families due to racial discrimination, a sharp increase from 2016. Landlords who hold prejudice against Muslim people often reject their applications or favour others. For Muslim families that have found somewhere to live, the living conditions are poor in many cases.

Inadequate housing has a knock-on effect on people's health, leading to medical problems. Even in accessing healthcare services, the Muslims surveyed reported not being treated equally. They were twice as likely as others in the general population to not have their medical needs met properly.

Another layer to this worrying landscape is that people are often discriminated against not only because of their race or religion but also because of their gender, sexuality, ethnic background or disability. The interaction between different, overlapping forms of discrimination produces unique and compounding

[373] Raphaële Xenidis and Linda Senden, "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination" (previously cited).
[374] Solon Barocas and Andrew D. Selbst, *Big Data's Disparate Impact*, 2016, https://doi.org/10.15779/Z38BG31.
[375] Nadya Labi, "Misfortune teller", 2012, The Atlantic, https://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846/
[376] Juan Carlos Perdomo and others, "Difficult lessons on social prediction from Wisconsin public schools" (previously cited).
[377] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 63.
[378] "Brazil, Case 12.440 Wallace de Almeida - Report on Admissibility and Merits" (previously cited), para. 150.
[379] Inter-American Commission on Human Rights. Police Violence Against Afro-Descendants in the United States (previously cited).
[380] Inter-American Commission on Human Rights, Police Violence Against Afro-Descendants in the United States (previously cited), p. 11.
[381] FRA, Being Muslim in the EU: Experiences of Muslims, 2025.

experiences of discrimination for individuals. It also illustrates the complexity in exposing the racism and discrimination experienced by racialized groups across the EU.

Similarly, a 2025 FRA survey of more than 16,000 people of African descent in Europe found that, across different areas of life, the highest rates of racial discrimination occurred in the area of employment, both when looking for a job (five-year prevalence of 34%) and at work (five-year prevalence of 31%). This was followed by accessing housing (five-year prevalence of 31%); and racism in public spaces, public transport, bars, shops or restaurants (five-year prevalence of 24%).[382]

More than half of respondents who experienced discrimination in at least one area of life say that they experienced it on more than one ground. The findings suggested intersecting forms of discrimination. For example, discrimination on any ground most often concerns young people, people with higher levels of education and persons living with disability. It is also commonly directed at people who wear traditional or religious clothing in public; self-identify as belonging to a minority in terms of disability, gender identity or gender expression, or sexual orientation; or who describe themselves as a person of African descent or as a Black person.[383]

Data used in risk profiling is "the product of structurally unequal conditions",[384] and of "unequal social relations – relations affected by centuries of history".[385] This data is used as predictive elements for individual futures.[386] Individuals are sorted and placed on socially constructed hierarchies from the moment of their very birth; they are then assigned unequal opportunities which ultimately result in different life outcomes, only to then come under surveillance from the very agencies supposed to protect them.[387] Instead of predicting future behaviour, risk profiling reproduces past injustices. Legal scholar Bernard Harcourt has argued that "risk" has become a proxy for race itself.[388]

To summarize, when governments use past social data in order to predict who is going to commit a crime or fraud, they inevitably end up targeting individuals who belong to historically oppressed or marginalized groups. As risk profiling works by categorizing and sorting individuals into risk categories based on personal characteristics,[389] it inherently and inescapably generates differential treatment of similar persons based on prohibited grounds. If no objective and reasonable justification exists for a differential treatment based on a prohibited ground, this amounts to discrimination under IHRL.

10.3 LEGITIMATE AIM MUST NOT BE ABUSED

The previous section shows that risk profiling causes differential treatment based on prohibited grounds. This section provides an analysis of the aim(s) that are commonly advanced by states implementing risk profiling and discusses whether these aims might be considered legitimate for the purpose of providing an objective and reasonable justification, as explained in Chapter 6.

In the context of a landmark judgment that banned the use of a fraud detection system in the Netherlands (the "SyRI" case), the Special Rapporteur on extreme poverty sent an amicus curiae brief to the court, noting that:

"The question for the court is whether the focus in the Netherlands and in other countries on the supposedly omnipresent 'welfare cheat' reflects the actual incidence of fraud and the consequences thereof or is in reality a function of ideologically-driven narratives reflecting discriminatory and prejudicial notions about the proclivities of the poor, the unemployed and migrant populations to engage in such fraud,

[382] FRA, Being Black in the EU: Experiences of People of African Descent, 2025.
[383] FRA, Being Black in the EU (previously cited), p. 15.
[384] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023, p. 55.
[385] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023, p. 55.
[386] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023.
[387] Catherine D'Ignazio and Lauren F. Klein, Data Feminism, 2023; Virginia Eubanks, Automating Inequality (previously cited).
[388] Bernard E. Harcourt, Against Prediction (previously cited).
[389] For example, ICCPR, Articles 2 and 26, "any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms". See further Chapter 6 for the applicable legal framework.

while the well-off members of the population are largely exempt from such concerted and narrowly targeted monitoring".[390]

If the underlying problem, such as social security fraud, is negligible or factually non-existent, the stated aim of fighting fraud should be called into question. This would likely mean the court will find that the "very weighty reasons" threshold is unmet and the restriction cannot be considered reasonably and objectively justified. In the words of the Special Rapporteur, the fraud, crime or other offence that governments are trying to reduce must exist in "actual government statistics" and "not merely in the rhetoric of politicians or the headlines of tabloid newspapers". International and regional human rights courts have held that "the possibility of some abuse" or "a mere administrative inconvenience" cannot be invoked to justify a difference in treatment on a prohibited ground.[391]

Research by Lighthouse Reports has highlighted that:

"the true extent of welfare fraud is routinely exaggerated by consulting firms, who are often the algorithm vendors, talking it up to near 5 percent of benefits spending while some national auditors' offices estimate it at between 0.2 and 0.4 [percent] of spending. Distinguishing between honest mistakes and deliberate fraud in complex public systems is messy and hard".[392]

Indeed, amid the hype of applying technological solutions to a multitude of social problems, states are accused of focusing on the wrong problems or exaggerating minor problems. In their 2019 amicus brief, the Special Rapporteur on extreme poverty recommended that states should stop "obsessing about fraud, cost savings, sanctions, and market-driven definitions of efficiency", and instead try to ensure a higher standard of living for people vulnerable to multiple and intersecting forms of discrimination.[393] Furthermore, invasive technologies such as risk profiling are most often deployed in contexts where they are likely to affect groups that are already stigmatized, disenfranchised or otherwise at the margins of society. As noted by the Special Rapporteur, the most well-off members of society are largely exempt from narrowly targeted monitoring.[394] This selective attention and use of invasive tools are often the result of pre-existing racial stereotypes and prejudices which posit racialized groups in particular as inherently criminal or dangerous. These stereotypes are compounded by structural issues including racial profiling, over-policing and higher conviction rates of racialized people.

While the aim of fighting or preventing fraud, crime or overstay, for example, can usually be considered legitimate under IHRL, these aims must not be abused to justify stigmatizing or otherwise discriminatory practices.

10.4 IS RISK PROFILING 'EFFECTIVE' AND NECESSARY?

Once it has been assessed that the aim pursued by making a differential treatment is legitimate, the means employed must be tested for necessity and proportionality (see further Chapter 6). This section discusses common misinterpretations of the test of necessity and asks whether risk profiling can be considered effective and necessary when taking the evidence presented in Chapter 8 into account.

When assessing necessity, courts also employ terms such as "suitable", "appropriate" and "effective" to assess whether a measure serves a stated aim. The term "effectiveness" in particular has been subject to misunderstandings and incorrect interpretations in the context of the right to equality and non-discrimination.

[390] UN Special Rapporteur on extreme poverty and human rights, Amicus Curiae Before the District Court of the Hague on the Case of NJCM c.s./De Staat der Nederlanden (SyRI) (previously cited).
[391] See, for example, HRC, *Gueye et al. v. France* (previously cited).
[392] See Lighthouse Reports, Unprecedented Experiment on Welfare Surveillance Algorithm Reveals Discrimination (previously cited).
[393] UN Special Rapporteur on extreme poverty and human rights, Amicus Curiae Before the District Court of the Hague on the Case of NJCM c.s./De Staat der Nederlanden (SyRI) (previously cited).
[394] See, for example, https://whitecollar.thenewinquiry.com/; Brian Clifton and others, "White collar crime risk zones", The New Inquiry, 9 March 2017, https://thenewinquiry.com/magazine/abolish/
Brian Clifton and others, "Predicting financial crime: augmenting the predictive policing arsenal", 2017, https://arxiv.org/abs/1704.07826

10.4.1 MISTAKEN CONFLATION OF PREDICTIVE PERFORMANCE WITH EFFECTIVENESS

States are often caught up in narratives of AI's inevitability and increasingly focus on mistakenly narrow notions of "effectiveness", conflating the predictive performance of risk profiling algorithms with the legal definition of this term.[395] This causes states to assume that increased detection rates of fraud and criminality checks can be a justification for subjecting people to differential treatment.

For example, governments sometimes compare the positive predictive value, or detection rate, of a risk profile to the detection rate obtained with random selections in order to prove "effectiveness".[396] In 2025, the Dutch Institute for Human Rights published a "Risk Profiling Assessment Framework" (*Toetsingskader risicoprofilering*), which states that "the effectiveness of profiling characteristics… should be compared with the effectiveness of the random sample" and that "the essence of this test is to determine whether the use of risk profiling actually contributes to the identification of more violations of norms than if only random checks were used".[397]

The Dutch Ministry of Education claimed that its DUO risk profiling system was more effective than random selection alone, and maintained this claim after eventually concluding that the system was discriminatory.[398] This was based on the finding of a higher detection of "illegitimate use of the student grant" when additional checks were based on selection by risk profiling compared to random selection alone. The number of detected incorrect claims or fraud when selected with a risk profile was between 28%-39% compared to 3.6-3.8% when selected as part of a random sample.

There are several issues with this type of reasoning. First, accepting such a low success rate from risk profiling means that more than 60% of the discriminatory checks wrongfully accused students of fraud. It is hardly credible to claim that a maximum detection rate of 39% is "effective" when it means the system is wrong in the vast majority of cases. In any rigorous policy evaluation, an intervention that subjects innocent citizens to invasive fraud investigations more than six times out of ten would be considered a severe operational failure rather than a success. This high tolerance for error might reflect the values and interests being served by the risk profiling policy;[399] namely, those of the state, rather than those of the people affected.

Second, a superficial comparison of base rate against detection rate ignores the fundamental issue that risk profiling is neither scientifically sound nor a credible exercise in evidence-based policy (for reasons listed in Chapter 8). Risk profiling will inevitably cause a large number of false positives and false negatives, as it is both under- and over-inclusive (see box below on degree of fit). This is problematic in itself, but justifying a system by simply comparing it to a random sampling baseline ignores the demographic distribution of these false positives. The burden of these wrongful accusations falls disproportionately on marginalized populations. Employing technical measures such as "algorithmic fairness" to distribute harms caused by false positives equally across demographic groups is contested and ineffective (see Chapter 9).

"While efficiency metrics serve optimization models well, they obscure distributive questions about who bears residual error… An ostensibly neutral objective function can codify value judgments that disadvantage certain populations".[400]

Third, when the underlying phenomenon – such as crime or fraud – is very rare, almost any targeted selection mechanism is guaranteed to produce a higher detection rate than random selection alone. This, however, comes at the cost of a very high false positive rate, and measurements are inherently unreliable. This phenomenon is known as the "base-rate fallacy".[401] The situation described in the DUO example, where there are more false positives than true positives, is an illustration of this fallacy, sometimes called the

[395] Amnesty International, *Profiled Without Protection* (previously cited).
[396] In this context and for the sake of the argument, we assume that these selections are truly randomized, unbiased and representative. It is worth noting that a careful process is needed to ensure that sampling is truly randomized and unbiased.
[397] Translated from Dutch, see College voor de Rechten van de Mens, "Toetsingskader Risicoprofilering 2025 – Integrale Versie", 2025, https://www.mensenrechten.nl/actueel/nieuws/2025/01/28/nieuw-toetsingskader-tegen-discriminatie-door-risicoprofilering
[398] Amnesty International, *Profiled Without Protection* (previously cited).
[399] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited), p. 5.
[400] M. Z. Naser, "On the philosophical naivety of engineers in the age of machine learning" (previously cited), p. 6.
[401] Jonathan J. Koehler, "The base rate fallacy reconsidered: descriptive, normative, and methodological challenges", March 1996, Behavioral and Brain Sciences, Volume 19, Issue 1, https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/base-rate-fallacy-reconsidered-descriptive-normative-and-methodological-challenges/5C0138815B364140B8711036405 5683B
Alexander Bird, "Understanding the replication crisis as a base rate fallacy", December 2021, The British Journal for the Philosophy of Science, Volume 72, Issue 4, https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axy051

"false positive paradox". As a result of this paradox, even an "accurate" profile – that is, one that finds true positives – produces a large number of false positives. Being flagged by a hypothetically accurate risk profile will therefore still yield a very low chance of actual crime or fraud, because the underlying base rate is very low among the population. Systems trying to predict terrorism have been dismissed for this very reason.[402]

Finally, truly randomized samples are extremely rare, which raises doubts of the reliability of the underlying ground truth. Moreover, even assuming that the sample is truly random and representative of the reference population, it would be necessary to assume that the ground truth is stable across time and across evolving environmental factors and enforcement measures. This is an unrealistic assumption that has been abundantly critiqued since individuals seeking to offend adapt their behaviour to evade new enforcement strategies (see also section 9.2 on random samples as a measure of ground truth).[403]

10.4.2 THE TRUE MEANING OF NECESSITY

The assessment of the "suitability" or "necessity" of a differential treatment requires an investigation into the causality between the intended, legitimate aim, and the means chosen to achieve this goal – that is, the distinction made.[404] If there is not a sufficiently large degree of causality between the means employed and the legitimate aim, such that the aim cannot be reached, then the distinction made cannot be necessary and it therefore amounts to discrimination. The suitability of a measure to achieve an intended goal can be evaluated based on scientific research on cause and effect.[405] The strength of the required causal connection will depend on the specific details of the case, the rights involved and the grounds of differentiation.

When a ground is "suspect", there is an assumption that the difference in treatment cannot be justified. In these cases, the court will carry out an intensive evaluation and set very high standards for the necessity of the interference. This will entail a more rigorous investigation into the causality between the legitimate aim and the means employed, and the court will carefully scrutinize the arguments brought forth by the state. The court "will be very keen on any sign of incoherence or inconsistency in the national argumentation. Such strictness can be particularly seen in cases regarding unequal treatment based on such grounds as ethnic origin, sexual orientation, birth or gender".[406] It is then for the state to demonstrate that the measure "is as effective as can be expected in the circumstances of the case".[407]

In Europe, key case law addressing indirect differential treatment of people on the grounds of race has dismissed scientifically contested educational tests. In *D.H. and others v. Czech Republic*, the state used scientifically dubious tests to assign students of Roma ethnicity to "special" schools, resulting in indirect discrimination. The ECtHR has observed that the tests "have given rise to controversy and continue to be the subject of scientific debate and research". "While accepting that it is not its role to judge the validity of tests", the Court concluded that "the results of the tests… were not capable of constituting objective and reasonable justification for the purposes of Article 14 of the Convention [protection from discrimination]".[408]

Differentiation based on risk profiling is mostly based on statistical correlations and therefore with no theoretical support or rigorous investigations into causality. Unless spuriousness has been ruled out, the correlation should be assumed to be spurious. A simple correlation says nothing inherently meaningful about the likelihood that an individual or group will violate a law or commit fraud. Therefore, the fact that some profiles might yield some level of predictive performance cannot count as a justification for treating marginalized groups differently. Even accounting for sampling bias, for example by training on data sampled randomly, risk profiling would still point to mere correlations, with no proof of a causal or otherwise meaningful relationship between the chosen profile variables and the behaviour being predicted (see further section 9.2 on random sampling as a remedy for sampling bias).

---

[402] Timme Bisgaard Munk, "100,000 false positives for every real terrorist: why anti-terror algorithms don't work", September 2017, First Monday, https://firstmonday.org/ojs/index.php/fm/article/view/7126
Marc Sageman, "The implication of terrorism's extremely low base rate", February 2021, Terrorism and Political Violence, Volume 33, Issue 2, https://doi.org/10.1080/09546553.2021.1880226
[403] Bernard E. Harcourt, Against Prediction (previously cited); Megan T. Stevenson, "Cause, effect, and the structure of the social world" (previously cited).
[404] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 49.
[405] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 50.
[406] Janneke Gerards, General Principles of the European Convention on Human Rights, 2023, p. 244. See, for example, *ECtHR, X. and Others v. Austria*, Application 19010/0719, Grand Chamber, February 2013, para. 144.
[407] Janneke Gerards, "How to improve the necessity test of the European Court of Human Rights", April 2013, International Journal of Constitutional Law, Volume 11, Issue 2, https://doi.org/10.1093/icon/mot004, p. 481.
[408] *ECtHR, D.H. and others v. Czech Republic* (previously cited), p. 199.

This shows that a mere comparison between base rate and detection rate is a shallow and unreliable measure of "effectiveness" in the context of a test against the prohibition of discrimination. Above all, it highlights why an incorrect interpretation of the term "effectiveness" in the context of a test against the prohibition of discrimination can lead public authorities to accept discriminatory risk profiles as "effective".

ASSESSMENT OF THE DEGREE OF FIT, UNDER- AND OVER-INCLUSIVENESS

A good way to assess suitability, or the degree of causality between the means employed and the legitimate aim, is by framing it in terms of under- and over-inclusiveness. Under- and over-inclusiveness may be compared to the statistical concepts of false negative (type 2 error) and false positive (type 1 error), respectively. As part of the test of the reasonable and objective justification, a Court will judge on whether the classification is sufficiently narrowly formulated in relation to the aim pursued by the provision, policy or practice.[409] This relates to the necessity test, that there are no other, less discriminatory policies that could meet the same aim. It is possible that a group affected by a particular burden by the state provision "is too widely defined in relation to the objective" (over-inclusiveness).[410] For example, selecting people with duplicate benefits claims for social security fraud checks is over-inclusive, as likely only a very small proportion of them will ultimately be found guilty of fraud. The profiling criteria would have to be more narrowly defined in order to obtain a better degree of fit. Over-inclusiveness is particularly problematic when it gives rise to a disadvantage, such as a higher likelihood of being subjected to additional scrutiny.[411]

Under-inclusiveness, on the other hand, indicates the opposite problem: when a rule is too narrowly defined to include all people who are comparable with respect to the aim of a rule or classification.[412] Both under- and over-inclusiveness therefore share a shortcoming of a classification by the state with regard to the legitimate aim. The requirements set by a Court vary from case to case, and their strictness will be determined by the intensity of the assessment.[413] When judging on discrimination cases where the differential treatment is based on a suspect ground, the requirements will be significantly higher.[414]

It follows that classifications defined by risk profiling, whether rule-based or data-driven, are inherently both under- and over-inclusive. The risk profile will be too widely defined, because of the general inaccuracy of this technique. A significant number of additional people will be selected for scrutiny. At the same time, classifications by risk profiles are under-inclusive, because they will be unable to select all people guilty of having actually committed a violation. This highlights that risk profiling is inherently at odds with the right to equality and non-discrimination.

10.4.3 ACCURATE PREDICTIONS DO NOT AUTOMATICALLY TRANSLATE TO EFFECTIVE POLICY

The questions of why a prediction is made and what enforcement or interventions follow it are crucial but frequently overlooked. Any patterns or correlations observed in data still need to be explained in order to form reasonable grounds for effective interventions. Even assuming that risk profiling delivers accurate predictions, research shows that high predictive performance alone rarely meets the real-world goals for deployment of predictive models. Relating short-term measurable outcomes such as risk predictions to broader policy goals is extremely difficult.[415]

For example, research shows that predictive policing approaches have failed to deliver on broader policy goals, such as reducing or preventing crime.[416] Likewise, research in the USA on more "traditional" street-level profiling indicates that this policy has been ineffective in reducing criminality while having clear discriminatory outcomes on historically marginalized groups, including African Americans.[417] The UN

[409] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 46.
[410] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 46.
[411] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 46.
[412] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 47.
[413] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 48.
[414] Janneke Gerards, Judicial Review in Equal Treatment Cases, 2005, p. 48.
[415] Lydia T. Liu and others, "Bridging prediction and intervention problems in social systems" (previously cited); Juan Carlos Perdomo, "The relative value of prediction in algorithmic decision making", 2024, http://arxiv.org/abs/2312.08511 (accessed 15 September 2025); Juan Carlos Perdomo and others, "Difficult lessons on social prediction from Wisconsin public schools" (previously cited).
[416] Amnesty International, *UK: Automated Racism* (previously cited).
[417] Inter-American Commission on Human Rights, Police Violence Against Afro-Descendants in the United States (previously cited), para. 81; New York City Department of Investigation Office of the Inspector General, An Analysis of Quality-of-Life Summonses, Quality-of-

Working Group of Experts on People of African Descent concluded that, in most cases where racial profiling has been applied, no significant results were achieved in terms of enhanced security, while great harm was committed against people of African descent and other marginalized groups.[418] In its General Recommendation 36, which also addressed algorithmic profiling, the CERD stated that

"racial profiling may also be ineffective and counterproductive as a general law enforcement tool. People who perceive that they have been subjected to discriminatory law enforcement actions tend to have less trust in law enforcement and, as a result, tend to be less willing to cooperate, thereby potentially limiting the effectiveness of law enforcement".[419]

Even without taking the harms into account, evaluating interventions in the social world is extremely difficult. Empirical research consistently shows that most interventions in the criminal legal space and in the social world more broadly have no lasting impact when evaluated with gold-standard methods of causal inference.[420] This should not be taken as an indication that all interventions in the social domain are meaningless. However, when such interventions have a negative impact in high-stakes contexts, a very high standard of proof should be adopted to justify any interference with human rights.

This challenge underscores the necessity for states to clearly articulate their legitimate aims when deploying predictive systems in public policy. For instance, if the declared objective is to reduce crime, then evaluations of the efficacy of any predictive tool must be tied to actual reduction in criminal activity, not merely proxy metrics like increased arrest rates or fraud detection rates (see section 8.2 on construct validity). Arrests are not crimes, and a rise in arrest rates may reflect changed enforcement practices, increased surveillance or biased targeting because of risk profiling, rather than a rise in crime levels or a genuine reduction in or prevention of offences, with the discriminatory side effect of exacerbating over-policing in marginalized communities. In turn, the lack of actual and verifiable offence reduction affects the evaluation of the effectiveness of the risk profiling system against the reasonable and objective justification test.

To test against the prohibition of discrimination, policymakers should clearly and transparently define outcome measures that map directly onto their stated legitimate aims; specify the causal pathways by which predictions are expected to produce those outcomes; and require evidence that interventions informed by predictions produce the intended societal benefits rather than just altering downstream administrative statistics or market-driven notions of efficiency. States must not be allowed to rely on over-general assertions of legitimate aim such as "crime prevention" or "national security." Where more specific aims are articulated, such as increased arrest rates or fraud detection rates, states must be able to demonstrate the legitimacy of these specific aims, as well as that they can be met in a manner consistent with human rights law.

10.4.4 RISK PROFILING DOES NOT DELIVER ACCURATE PREDICTIONS

Risk profiling aiming to predict the propensity to fraud or social security fraud detection fails to deliver what it promises: accurate predictions (see Chapter 8 for an in-depth explanation). For example, Sonja Starr described the accuracy of evidence-based sentencing (the equivalent of risk profiling in criminal justice sentencing in the USA) as "only modestly better than a coin toss".[421] This conclusion is supported by a wealth of research from a variety of scientific disciplines, which is slowly finding its way to official government advisory boards. Unfortunately, governments sometimes downplay or miss these issues, assuming that technology alone will resolve policy challenges. This narrative is also the outcome of intense corporate influence.

In 2025, the Scientific Advisory Board to the Dutch National Police published a report on the challenges faced by law enforcement in relation to digitization and technology (see box below). The board recommended in unequivocal terms that the police abandon efforts to predict individual behaviour, in order

Life Misdemeanor Arrests, and Felony Crime in New York City, 2010-2015, 22 June 2016; see also Bernard E. Harcourt and Jens Ludwig, "Broken windows: new evidence from New York City and a five-city social experiment", 2006, University of Chicago Law Review, Volume 73; Bernard E. Harcourt & Jens Ludwig, "Reefer madness: broken windows policing and misdemeanor marijuana arrests in New York", 2006, University of Chicago Public Law & Legal Theory Working Paper, No. 142.

[418] UN Working Group of Experts on People of African Descent, Sixth Session Report, 2007, UN Doc. A/HRC/4/39.

[419] CERD, General Recommendation 36 (previously cited), para. 26.

[420] Megan T. Stevenson, "Cause, effect, and the structure of the social world" (previously cited).

[421] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited).

to mitigate legitimate societal concerns without compromising the effectiveness and efficiency of policing.[422] The management of the Dutch National Police dismissed these recommendations in an official response.[423]

In 2026, a national committee of inquiry on racism and discrimination in the Netherlands published a report on data-driven profiling.[424] The committee found that data-driven profiling conflicts in several respects with core principles of good governance such as rule of law, democratic legitimacy and administrative capacity. The committee also found a lack of clear legal red lines and a lack of effective legal protection for citizens. In addition, the committee stated that profiling systems can reinforce existing inequalities, because they are based on historical data in which earlier forms of discrimination are already embedded. Crucially, "although profiling is often justified as being more efficient and effective than other selection methods, the committee finds that convincing empirical evidence for this is largely lacking".[425] The committee called on state authorities to replace data-driven profiling with fair alternatives such as random selection (see box in section 10.6).

SCIENTIFIC ADVISORY BODY RECOMMENDS ABANDONING INDIVIDUAL PREDICTION DUE TO INEFFECTIVENESS AND SOCIETAL CONCERNS

The Scientific Advisory Board for the Dutch National Police was instituted in 2022 to help the police with "making careful considerations on the functioning of the police from a scientific perspective", "maintaining society's trust in the police and the legitimacy of police action", and to "ensure that the Police is subservient to the Rule of Law".[426] In its 2025 report "Navigating in No-man's Land" (*Navigeren in niemandsland*), the Advisory Board addresses seven technological challenges faced by the police, ranging from data governance to AI literacy. The Advisory Body states that the evaluation of the effectiveness of data and AI applications is often still limited: "In many cases, there is a lack of robust evaluations, and without understanding the effectiveness, it becomes problematic to assess public values like proportionality and subsidiarity, which undermines the legitimacy of these systems".[427]

According to the report, a rapidly growing amount of data and AI applications within the police force focus on crime prediction, both on a geographic and individual level. The existing evaluations of these systems "show varying results", but accuracy at the local or individual level "remains a challenge, like in other domains such as weather forecasting and virus outbreaks". This is partly due to "the strong contextual dependence of human behaviour".[428]

"The past fifteen years", the Advisory Board goes on, "have shown that predicting at the individual level entails high and far-reaching risks. The effectiveness is often unclear and it is undesirable from the perspective of the trust that citizens may or may not have in the police. Human behaviour is poorly predictable and it is also not clear which variables matter in a specific social context, how they can be measured and recorded. Prediction – for example by assigning a risk score – can lead to incorrect conclusions with major harmful consequences. In addition, some groups of citizens – often in vulnerable positions – leave more digital 'tracks' than others, which can lead to disproportionate attention. The fact that, in retrospect, perpetrators sometimes share common characteristics does not justify predicting behaviour. Predicting at the individual level increases the risk of unequal treatment, undermines trust in the police and is both scientifically and normatively problematic. The conclusion therefore seems justified that 'predicting' criminal behaviour at the personal level should not be a self-evident ambition [of the National Police]".[429]

Given all of the above, "the Advisory Body advises the police not to further focus on (algorithmic) applications aimed at predicting (criminal) behaviour at the individual level and to make this publicly

[422] Wetenschappelijke Adviesraad Politie, Navigeren in Niemandsland: Zeven Urgente Uitdagingen Rondom Digitalisering En AI in Politiewerk, 2025, https://www.wetenschappelijkeadviesraadpolitie.nl/publicaties/navigeren-in-niemandsland/ (in Dutch).
[423] Nationale Politie, Reactie Korpschef Op Adviesrapport WARP - Navigeren in Niemandsland, 2025, https://www.politie.nl/binaries/content/assets/politie/onderwerpen/publicaties/2025/reactie-korpschef-op-adviesrapport-navigeren-in-niemandsland.pdf (in Dutch).
[424] Staatscommissie tegen discriminatie en racisme, "Voortgangsrapportage: principes voor profilering" ["Progress report: principles for profiling"], 2026, https://www.staatscommissietegendiscriminatieenracisme.nl/documenten/2026/05/07/principes-voor-profilering (in Dutch).
[425] State committee on racism and discrimination, "Stop datagedreven profilering door overheid vanwege discriminatie" ["Government should stop data-driven profiling because of discrimination"], 2026, https://www.staatscommissietegendiscriminatieenracisme.nl/actueel/nieuws/2026/05/07/staatscommissie-stop-datagedreven-profilering-door-overheid-vanwege-discriminatie (in Dutch).
[426] Machine translated from Dutch: https://www.wetenschappelijkeadviesraadpolitie.nl/assets/InstellingsbesluitWetenschappelijkeAdviesraad.pdf
[427] Wetenschappelijke Adviesraad Politie, Navigeren in Niemandsland (previously cited), pp. 20-21.
[428] Wetenschappelijke Adviesraad Politie, Navigeren in Niemandsland (previously cited), pp. 20-21.
[429] Wetenschappelijke Adviesraad Politie, Navigeren in Niemandsland (previously cited), pp. 20-21.

known. This allows legitimate social concerns to be eliminated without compromising the efficiency and effectiveness of police action".[430] The senior leadership of the police rejected the Advisory Body's recommendation by stating that while the recommendation is "understandable",[431] the police "cannot ignore" ongoing developments in the field of AI.

10.5 NO FINANCIAL JUSTIFICATION

States have sometimes cited risk profiling as a way to streamline social services and improve their cost-effectiveness, including the efficient use of enforcement capacity.[432]

In high-risk domains such as policing, migration and social security, governments speak of goals such as "efficiency of administration through digitization" and better fraud, crime or migration control by preventing phenomena such as social security fraud.[433] These are legitimate aims. However, while evidence for the reduction of fraud and criminality has remained elusive, a more consistent outcome is the penalization of society's most marginalized groups for attempting to access their rights and essential services.[434] In choosing to use invasive and contested methods such as risk profiling for these tasks, governments willingly ignore their human rights obligations, such as the obligation to adopt effective measures to review government policies and to amend, repeal or nullify laws and regulations that may lead to or perpetuate racial discrimination, including racial profiling (see Chapter 6 for the applicable legal framework).

Financial arguments do not justify unequal or disadvantageous treatment of a group or individuals on the basis of a prohibited ground. Because of the importance of the right to non-discrimination, interferences with this right demand very weighty or compelling reasons in order to be justified when the differential treatment is based on a prohibited ground. In line with this principle, international and regional human rights courts have adjudicated cases where the "possibility of some abuse" or a "mere administrative inconvenience" cannot be invoked to justify a difference in treatment.[435] The CJEU has taken a similar approach and strictly construed the possibilities for justifying differential treatment in its indirect discrimination doctrine.[436] The CJEU has rejected purely budgetary or administrative considerations as a justification for differential treatment.[437]

Under international human rights law and standards, states have the obligation to guarantee the right to equality and non-discrimination, even if they incur higher costs by doing so. This is true regardless of whether higher costs are incurred due to an increased number of (supposedly) wrongly provided social security payments, or due to needing a higher number of public officials to enforce investigations/checks. Additionally, recent scientific evidence suggests that in scenarios where resources are restricted, increasing the available resources is more resource efficient than prediction with limited resources – even when assuming accurate predictions.[438]

10.6 STEREOTYPING BY DESIGN

This section describes the harms inflicted by risk profiling as a result of their inner workings and of the contexts in which they are deployed.

[430] Wetenschappelijke Adviesraad Politie, Navigeren in Niemandsland (previously cited), pp. 20-21.
[431] Nationale Politie, Reactie Korpschef Op Adviesrapport WARP - Navigeren in Niemandsland (previously cited), p. 2.
[432] Amnesty International, *Netherlands: Xenophobic Machines* (previously cited); Amnesty International, *Trapped by Automation* (previously cited); Amnesty International, *Denmark: Coded Injustice* (previously cited).
[433] Amnesty International, *Denmark: Coded Injustice* (previously cited).
[434] Amnesty International, *Digitally Divided* (previously cited).
[435] See, for example, HRC, *Gueye et al. v. France* (previously cited).
[436] Janneke Gerards, "Non-discrimination, the European Court of Justice and the European Court of Human Rights: who takes the lead?", 2020, in Thomas Giegerich (editor), The European Union as Protector and Promoter of Equality, https://link.springer.com/10.1007/978-3-030-43764-0_7
[437] See, for example, CJEU, *Leitner v. Landespolizeidirektion Tirol*, Case C-396/17, 8 May 2019, para. 43; CJEU, *Starjakob v. ÖBB Personenverkehr AG*, Case C-417/13, 28 January 2015, para. 36.
[438] Juan Carlos Perdomo, "The relative value of prediction in algorithmic decision making" (previously cited); Juan Carlos Perdomo and others, "Difficult lessons on social prediction from Wisconsin public schools" (previously cited); Lydia T. Liu and others, "Bridging prediction and intervention problems in social systems" (previously cited).

10.6.1 RISK PROFILING PUNISHES PEOPLE FOR BEING PART OF A GROUP

Stereotypes are the result of a cognitive process in which attributes or traits of individuals are generalized to all members of a group. This can lead to stigmatization of a group or its members when the stereotypes reflect negatively on the person or group affected. Stereotypes can play at least two roles in discriminatory risk profiling: they can be 1) the reason behind a profile or differential treatment or 2) the consequence of risk profiling. Both are problematic from a non-discrimination and IHRL perspective. Stereotyping as a rationale for creating a profile that results in differential treatment of marginalized groups has direct discriminatory impacts; while stereotyping or a history of stigmatization can be reasons for applying the "very weighty reasons" test, which almost automatically results in the finding of a violation.[439]

The "suspectness" of the ground of differentiation also plays a role when the stereotype is not the cause but an effect of a risk profile, such as in indirect discrimination cases. Advocates of risk profiling might be inclined to defend such differentiation by claiming that it is justified by the underlying statistical observations. For example, if nationality is not one of the profiling criteria, but non-nationals end up being profiled more often *indirectly*, surely this must be justified by the fact that non-nationals violate the norm in question more often than nationals. However, this reasoning is problematic both scientifically and from a human rights perspective. As explained, correlations say nothing meaningful about the likelihood that an individual or group will violate a law or commit fraud, unless spuriousness has been ruled out and a causal pathway has been specified. Moreover, prohibited grounds are suspect because they are immutable personal characteristics, irrelevant for performing in society, and/or go hand-in-hand with historical or social discrimination and stigmatization.[440] As a result, a causal pathway arguably does not exist, which explains why a differential treatment on these grounds virtually always results in a violation.

Legal scholar Sonja Starr explains that evidence-based sentencing, the equivalent of risk profiling in criminal justice sentencing in the USA, "is all about generalizing based on statistical averages, and its advocates defend it on the basis that the averages are right".[441] However, the United States Supreme Court "has squarely rejected statistical discrimination – use of group tendencies as a proxy for individual characteristics – as a permissible justification for otherwise constitutionally forbidden discrimination",[442] that is, discrimination against marginalized individuals or groups. As Starr and Frederick Shauer specify, statistical generalizations are not categorically forbidden: it would be "hard to imagine government functioning" without them.[443] But when they "contravene a distinct and fundamental constitutional value" and have particularly "harmful or expressively invidious" social consequences, they will be rejected by the US Supreme Court, even if they have statistical support.[444] Statistical generalizations cannot count "as a defense of classifications triggering heightened constitutional scrutiny" as they "implicitly embrace a type of utilitarian calculus" that is incompatible with constitutional values.[445] The US Supreme Court "has held that this defence of gender and race discrimination offends a core value embodied by the Equal Protection Clause of the US Constitution: people have a right to be treated as individuals".[446]

Even though a risk profile may have some, possibly unexplained, level of predictive performance on a group level, it will still be largely incorrect and therefore unfair at the individual level, and deny individuals the opportunity to defy the apparent trend.[447] Moreover, since any measured trends or correlations are not inherently indicative of causality, and because constructs such as race or ethnicity do not causally relate to criminality or fraud, these correlations really uncover systemic discrimination rather than "risk". The HRC has stated that racial profiling "would run counter to an effective policy aimed at combating racial discrimination" because it would contribute to the spread of xenophobic attitudes.[448]

---

[439] Janneke Gerards, "The margin of appreciation doctrine, the very weighty reasons test and grounds of discrimination" (previously cited).
[440] Janneke Gerards, "The margin of appreciation doctrine, the very weighty reasons test and grounds of discrimination" (previously cited).
[441] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited).
[442] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited).
[443] Frederick Schauer, Profiles, Probabilities, and Stereotypes, 2006; Sonja B Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited).
[444] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited).
[445] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited), p. 828.
[446] Sonja B. Starr, "Evidence-based sentencing and the scientific rationalization of discrimination" (previously cited), p. 827.
[447] Raphaële Xenidis and Linda Senden, "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination" (previously cited).
[448] HRC, *William Lecraft v. Spain*, 2009, UN Doc. CCPR/C/D/1493/2006,p. 7.2; Janneke Gerards, Fundamental Rights (previously cited), p. 25.

RANDOM SAMPLES AS AN ALTERNATIVE SELECTION METHOD

Evaluating the necessity of the chosen means includes choosing the option that is the least harmful from the perspective of the rights affected, yet is still able to realize the legitimate objectives pursued by the state.[449] It can generally be said that most government objectives can be achieved in different ways.[450] Using random samples for selecting innocent people for controls ensures the prevention of discriminatory selection. The use of random selection for fraud controls in social security programmes was documented in several countries in a 2010 World Bank report.[451] In May 2026, a national commission of inquiry on racism and discrimination in the Netherlands published a report on data-driven profiling, calling on authorities to stop this practice and to replace it with random selections because this would better ensure non-discrimination while still being effective.[452] Before deciding to use a risk profiling system, authorities should investigate other selection methods that impose less of an interference with the right to equality and non-discrimination, such as random selection. Strictly random selection, by definition, grants everyone an equal chance for selection and therefore does not disproportionately affect particular groups.

10.6.2 PERFORMATIVE EFFECTS

> **"The model might be effectively no better than random sampling, as investigations into these systems keep concluding, yet they are also treated as forms of perfect knowledge – or, at least, good enough knowledge to justify serious actions."**
>
> Jathan Sadowski (Monash University) [453]

By selecting individuals for checks on the basis of a risk profiling prediction, governments *produce* its causal effects and inescapably treat an individual *as if* they had acted suspiciously or indeed violated the law.[454] When people are subjected to additional scrutiny following risk profiling, they are being treated as de facto suspects. This is the case regardless of human intervention or of the intentions of the actor using the risk profile. Correlations are treated as inherently predictive and therefore punish people for being part of a (statistical) group.

Subjecting certain groups of people to more scrutiny will inevitably result in finding a higher number of violations within these groups and lower rates in other groups.[455] This phenomenon results in feedback loops

[449] Janneke Gerards, General Principles of the European Convention on Human Rights, 2023, p. 236.
[450] For example, to ensure safety during gym classes in primary and secondary schools, schools might prohibit headscarves. It might be said that the prohibition is necessary, since some headscarves do not allow the pupils to move freely, and there is a risk that the children may be harmed by their headscarves if they trip or fall. However, one could think of requiring pupils to wear elastic sports headscarves that do not hamper their movements and pose no danger to their safety. The question then is to determine whether these alternatives are "equally effective" and whether they should be, particularly if one of the alternatives is evidently less intrusive. The prescription of elastic sports headscarves will probably not be as effective as a full prohibition, since even a gym headscarf could still pose some risk during exercises. Nonetheless, it may be preferable over a complete prohibition on headscarves from the perspective of human rights, in particular the rights to non-discrimination and freedom of religion. See, for a full discussion, Janneke Gerards, General Principles of the European Convention on Human Rights, 2023, p. 236.
[451] Christian Stolk and Emil Tesliuc, World Bank Report: Toolkit on Tackling Error, Fraud and Corruption in Social Protection Programs, January 2010, https://www.researchgate.net/profile/Emil-Tesliuc/publication/238798518_Toolkit_on_tackling_error_fraud_and_corruption_in_social_protection_programs/links/53fdab8c0cf22f21c2f822be/Toolkit-on-tackling-error-fraud-and-corruption-in-social-protection-programs.pdf
[452] Staatscommissie tegen discriminatie en racisme, "Voortgangsrapportage: principes voor profilering", (previously cited).
[453] Jathan Sadowski, "Machine's eye view: postmodern data science and the politics of ground truth", March 2026, Science, Technology, & Human Values, Volume 51, Issue 2, https://doi.org/10.1177/01622439251331138, 251–276, p. 265.
[454] Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[455] This phenomenon results in runaway feedback loops and has been studied extensively and attributed several names, including "disproportionate minority contact" and "over-policing".

and has been studied extensively and attributed several names, including "disproportionate minority contact" and "over-policing".[456] The resulting data might then be the input for training future risk profiling models. The effect of risk profiling is therefore *performative:* individuals or groups are transformed from statistical, hypothetical suspects into actual suspects, solidifying pre-existing prejudices or generating new ones. Risk profiling therefore "enact[s] discriminatory correlations by treating them as causalities and using them as foundations for further decisions, recommendations and policies".[457] Discrimination and inequality that "have been socially constructed over time become considered 'natural properties' of certain groups of population", thereby entrenching these injustices and further contributing to the systematic deprivation of equal opportunities.[458]

For example, if all people with high healthcare costs are selected for social security fraud detection investigations, people with chronic illnesses will be over-selected and thereby treated as de facto suspects,[459] likely resulting in discrimination on the grounds of physical or cognitive disability. Having high healthcare costs is not causally related to fraudulent behaviour. Even if having high healthcare costs is correlated with fraud, there could be other factors at play, making this correlation likely spurious. This means that simply sharing the characteristic "high healthcare costs" is no reasonable and objective justification for subjecting people with chronic illnesses to higher levels of government scrutiny. Still, increased government scrutiny of this group will inevitably result in finding a higher number of violations, cementing the correlation in governmental data.

10.7 INEFFECTIVE AND INSUFFICIENT SAFEGUARDS AGAINST DISCRIMINATION

The use of risk profiling by states is only allowed if it respects states' obligations to respect, protect and fulfil human rights. This includes setting up the necessary and effective safeguards needed to ensure that discrimination is prevented as well as ensuring the respect and protection of all other human rights.

10.7.1 TECHNICAL MEASURES ARE INEFFECTIVE

Technical measures for reducing bias in risk profiling algorithms are analysed in Chapter 9 of this report. Governments are eagerly reaching to these technical measures to prevent discrimination. However, these methods are increasingly described as being "only minimally effective at preventing harms".[460] Issues include the impossibility of finding and correcting all biases and the inadequacy of fairness techniques to prevent intersectional discrimination (see Chapter 9 for details). Moreover, if the prediction system is malfunctional or invalid, the outcomes will be misleading and potentially harmful even if they are equally distributed across groups. For example, algorithms can still predict the wrong quantity even if they satisfy fairness criteria.[461]

Consequently, bias reduction has been described as a "red herring" because it impedes more fundamental discussions of societal harms by technology, and can obscure the inherently political nature of technology.[462] Academics are increasingly rejecting algorithmic fairness in favour of amplifying and supporting voices from the communities most affected by technologies' biases and harms.[463]

These communities must be able to resist and refuse the deployment of harmful technology.[464] Without this right of refusal, the design and deployment of automated risk profiles will continue to reflect only the narrow

[456] Cathy O'Neil, Weapons of Math Destruction (previously cited).
[457] Janneke Gerards and others, Algorithmic Discrimination in Europe (previously cited), p. 8; see also Raphaële Xenidis and Linda Senden, "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination" (previously cited); Mel Andrews and others, "The reanimation of pseudoscience in machine learning and its ethical repercussions" (previously cited).
[458] Raphaële Xenidis and Linda Senden, "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination" (previously cited), p. 9.
[459] One such case was documented in the Netherlands, where chronically ill people were subjected to systematically higher control rates because higher healthcare costs were labelled as suspect. Follow The Money, Chronisch Zieken Opgejaagd Door Vooringenomen Controles Belastingdienst (Tax Administration Hunts Chronically Ill People in Biased Controls), 2025, https://www.ftm.nl/artikelen/belastingdienst-profileert-chronisch-zieken-zorgkosten (in Dutch).
[460] Follow The Money, Chronisch Zieken Opgejaagd Door Vooringenomen Controles Belastingdienst (previously cited), p.78.
[461] Amanda Coston, "Falsifying predictive algorithms" (previously cited), p. 4.
[462] Dasha Pruss and others, "Prediction and punishment" (previously cited), p. 10.
[463] Greta Byrum and Ruha Benjamin, "Disrupting the gospel of tech solutionism to build tech justice" (previously cited).
[466] Greta Byrum and Ruha Benjamin, "Disrupting the gospel of tech solutionism to build tech justice" (previously cited).

and privileged interests of those in power, investors and elite technologists, rather than the actual needs of the people they claim to serve.

10.7.2 HUMAN-IN-THE-LOOP AS AN INSUFFICIENT SAFEGUARD

Typically, when discussing the risks of automated decision-making, governments advance the argument that the decision-making process is not fully automated, because the risk profiling algorithm only informs the final decision by a human operator (so-called "human-in-the-loop"). This argument has been criticized by scholars and has more recently been the object of a preliminary ruling by the CJEU.[465] Humans might be reluctant to advise against the algorithmic prediction because of automation bias, might assume the neutrality or objectivity of the technology, or may lack enough time to adequately review individual cases, so that the human intervention is not meaningful. Amnesty International has also documented cases where government officials were unable to contradict or override notifications to what they described as a clear error made by the system.[466]

Furthermore, automation bias is a persistent problem in semi-automated decision-making. AI is a technology that mimics human capabilities and evidence suggests that human-computer interactions can be influenced by system deference: the (over)reliance of human decision-makers on the outputs of an AI system.[467] This is rooted in the phenomenon that Meredith Broussard has referred to as "techno-chauvinism"; the widely held idea that technology is better than human beings at solving complex social, political and economic issues.[468] Due to this trust or belief in machine-driven systems, users of seemingly sophisticated technologies such as generative AI may therefore be too readily willing to accept generated outputs as reliable, inhibiting their own critical faculties, or worse, be subject to deliberate manipulation, raising risks for the right to freedom of thought.

Equally of note, the intervention of a human, even when significant, only takes place after the selection and the resulting differential treatment has already occurred. When a differential treatment is based on a prohibited ground and is not objectively and reasonably justified, it amounts to discrimination under IHRL, regardless of human involvement in the final decision. Finally, even if a human assesses the prediction before imposing a fine or other penalty, individuals are treated as de facto suspects by being subjected to extra scrutiny (see also section 10.6 on performative effects).

In sum, the most frequently employed safeguards to remedy and prevent discrimination by risk profiling algorithms are ineffective, insufficient or inadequate.

10.8 DISCRIMINATION IS NOT A BUG BUT A CORE FEATURE OF RISK PROFILING

The three-part test requires restrictions on human rights to be authorized by law, pursue a legitimate aim, and be necessary and proportionate. This section discusses the last part of the three-part test: proportionality in the strict sense (*stricto sensu).*

There may be an objective and reasonable justification for differential treatment based on prohibited grounds, provided that it is proportionate to the aim pursued, meaning that the significance of the aim pursued outweighs the disadvantage suffered by the targets of discrimination and their wider community. Therefore, assessing the proportionality entails considering whether stereotypes and stigmatization play a role in the distinction made, and taking into account whether the treatment has an adverse effect on people. Indeed, stereotypes and stigmatization have been consistent findings of investigations into risk profiling systems (see also Chapter 7).

---

[465] CJEU, Schufa Holding (scoring) case, C-634/21, 7 December 2023.
[466] Amnesty International, *Trapped by Automation* (previously cited), p. 11.
[467] Saar Alon-Barkat and Madalina Busuioc, "Human-AI Interactions in public sector decision making: 'automation bias' and 'selective adherence' to algorithmic advice", January 2023, Journal of Public Administration Research and Theory, Volume 33, Issue 1, https://doi.org/10.1093/jopart/muac007
Mary Cummings, "Automation bias in intelligent time critical decision support systems", 2004, AIAA 1st Intelligent Systems Technical Conference, https://arc.aiaa.org/doi/10.2514/6.2004-6313 (accessed 10 March 2026); Amnesty International, Artificial Intelligence and Judicial Systems: Submission to the UN Special Rapporteur on the Independence of Judges and Lawyers (Index: IOR 40/9316/2025), 2025.
[468] Meredith Broussard, Artificial Unintelligence: How Computers Misunderstand the World, 2018.

HARMS AND IMPACT ON HUMAN RIGHTS

In high-stakes domains, being subjected to risk profiling yields serious harms which include both material and non-material effects: inherent stereotyping, stigmatization and harms from performative effects; damage to reputation that follows people across systems; arbitrary accusations of crime and fraud, and potential imprisonment; psychological harms such as anxiety and loss of trust in institutions that affects society as a whole; undue administrative burdens including financial losses, such as reduced income or denial of licences or subsidies, denial or delay of social security payments, fines or costly compliance burdens; evictions or loss of employment as a consequence of financial losses and debts with authorities; loss of autonomy and chilling effects from intrusive monitoring; and denial of entry at the border, forced displacement, detention or deportation for people on the move. These harms are compounded by the lack of transparency regarding the employment of these systems, which prevents people from resisting their use and from claiming their rights.

Risk profiling has been shown to cause discrimination on the grounds of, among others, race and ethnicity, gender, socio-economic status and disability. The discriminatory harms of risk profiling are especially evident when viewed through the lens of intersectional discrimination (see section 10.2), and Amnesty International has documented various cases of harms across different countries where people were discriminated against based on one or multiple intersections of their race, ethnicity, national origin, gender, disability, age and social and economic status.[469]

Additionally, the discriminatory impacts of risk profiling reach beyond human bias and individual rights, as they enable consolidation of more structural types of discrimination. Risk profiling systems have their roots in historical systems used for categorizing, cultivating and instrumentalizing personal data to create and maintain social and racial hierarchies and can best be viewed as an extension of these pre-existing systems of power.[470]

Risk profiling poses risks to numerous other human rights, and these harms may also occur in a discriminatory manner, since the right to non-discrimination is a right unto itself, and also a cross-cutting right. Research by Amnesty International and other human rights organizations has consistently shown that risk profiling negatively affects the rights to a fair trial, remedy and redress, the presumption of innocence, the right to privacy and data protection, the rights to social security and an adequate standard of living and the full realization of human dignity. Because of the differential treatment, these harms are unevenly distributed across societal groups and likely to bring particular disadvantage to individuals belonging to marginalized groups.

To judge on the proportionality of risk profiling, these well-documented and persistent harms suffered by the people targeted by risk profiling systems and their wider community must then be weighed against (1) the significance of the aim(s) pursued, and (2) the suitability or effectiveness of the differential treatment caused by risk profiling for achieving these aims.

LEGITIMATE AIM

Section 10.2 explains that reducing crime or fraud are legitimate aims for governments. However, there is a concrete risk of such aims being abused to form a cover for invasive, community-wide surveillance. In the absence of solid evidence related to levels of benefits fraud warranting intrusive surveillance and risk profiling, it is important to be critical of the stated aims. Risk profiling is most often deployed in contexts where it is likely to affect groups that are stigmatized, disenfranchised or otherwise already at the margins of society, while the most privileged members of society are largely exempt from narrowly targeted monitoring.[471] This selective attention and use of invasive tools are often the result of pre-existing racial stereotypes and prejudices which posit racialized groups in particular as inherently criminal or dangerous. These stereotypes are compounded by structural issues including racial profiling, over-policing and higher conviction rates of racialized people.

SCIENTIFIC CONTEXT

Chapter 8 situates risk profiling systems in the scientific debate. Rather than collecting data that is specific or pertinent to the prediction of the behaviour labelled as "risky", governments use pre-existing and unfit-for-purpose administrative data to train predictive models (so-called convenience samples). To guarantee methodological or epistemic validity, researchers would need to follow discipline-specific research norms

[469] Amnesty International, *Netherlands: Xenophobic Machines* (previously cited); Amnesty International, *Denmark: Coded Injustice* (previously cited); Amnesty International, *Profiled Without Protection* (previously cited).
[470] Ruha Benjamin, Race After Technology: Abolitionist Tools for the New Jim Code (previously cited); Virginia Eubanks, Automating Inequality (previously cited).
[471] See, for example, https://whitecollar.thenewinquiry.com/; Brian Clifton and others, "White collar crime risk zones" (previously cited); Brian Clifton and others, "Predicting financial crime" (previously cited).

that are appropriate to the domain of application. This would likely prevent the proliferation of ill-founded and unreliable prediction models such as risk profiles.

Multiple and interlocking types of bias are inherent to risk profiling, and its prediction goals are unmeasurable and extremely difficult to operationalize, which fundamentally undermines its scientific validity. Moreover, risk profiling models are not robust and misrepresent the human reality in which they are deployed, because they do not rule out spurious correlations and are not grounded in rigorously validated theories. There is no reliable way to know when, where or how the model's predictions will fail, leading to harm for targeted people. Some prediction applications, like predicting criminality or the propensity to fraud, have been debunked and decried as scientific malpractice. As a result, constructing a risk profile for social security fraud or criminality is not a realistic technical undertaking, nor is it a credible exercise in evidence-based policy. It is an attempt to operationalize suspicion in the absence of a reliable ground truth, and it is bound to discriminate because of biases inherent to the data and the social phenomena that they quantify. Such models are prone to produce biased and erroneous predictions and expose targeted individuals to arbitrary outcomes.

EFFECTIVENESS AND NECESSITY

Section 10.4 analyses risk profiling in terms of its "effectiveness" under IHRL. Such effectiveness is, at best, subject to debate. Risk profiling aiming to predict the propensity to fraud or social security fraud detection fails to deliver what it promises: accurate predictions. Instead, it produces an alarmingly high number of false positives – mostly distributed to racialized and marginalized people. Still, states fail to recognize these severe operational failures because of corporate influences, uncritical "techno-solutionism" and an entrenched belief that risk profiling is effective, despite the empirical evidence. Furthermore, even if risk profiling predictions were accurate, these predictions would not automatically lead to effective interventions. There is a lack of actual and verifiable reduction in offences following risk profiling predictions, which compromises the evaluation of the effectiveness of these systems under IHRL. States must not be allowed to rely on over-general assertions of legitimate aim such as "crime prevention" or "national security". Where more detailed aims are articulated, such as increased arrest rates or fraud detection rates, states must be able to demonstrate the legitimacy of these specific aims, as well as demonstrating that they can be met in a manner consistent with human rights law. Finally, there are alternative options for effectively preventing crime, fraud or other offences in a manner that is not discriminatory and which is consistent with states' human rights obligations – including, for example, using well-designed and non-discriminatory random sampling.

INEFFECTIVE SAFEGUARDS

Statistical measures to prevent or repair discriminatory outcomes give an illusion of clarity and a false veneer of objectivity, but have proven inadequate to prevent discrimination in practice. The core problem lies in treating statistical methodology as a substitute for addressing underlying social and structural biases. These measures, such as algorithmic fairness, have consistently failed to address human rights concerns and have come under increasing criticism from leading scholars. Meaningful human intervention prior to the final punitive decision remains an insufficient safeguard, because the differential treatment has already materialized, because of the deriving performative effects, and because it fails to adequately mitigate automation bias, which remains a significant and unsolved issue.

CONCLUSION

In light of the grave harms inflicted by risk profiling to the individuals and communities affected, weighted against the significance of its stated aims and the overarching doubts concerning its effectiveness and necessity, the differential treatment inherent in risk profiling systems cannot be deemed proportionate and thus cannot be reasonably and objectively justified. Without a reasonable and objective justification, differential treatment based on prohibited grounds constitutes discrimination under IHRL. It follows that risk profiling in high-stakes contexts is incompatible with IHRL.

11. RECOMMENDATIONS TO ALL STATES

11.1 PROHIBITION OF RISK PROFILING IN HIGH-STAKES CONTEXTS: LAW ENFORCEMENT, MIGRATION AND SOCIAL SECURITY

- Amnesty International believes that the use of data-driven or rule-based predictive, profiling and risk assessment systems, regardless of human involvement in the final decision, should be prohibited in high-stakes contexts. States should develop or amend existing AI regulation to ensure this prohibition in high-stakes contexts. This prohibition should, at a minimum, apply in the following areas:
    - In the context of law enforcement, when police and criminal justice authorities use risk profiling to predict, profile or assess the risk or likelihood of offending, re-offending or other criminalized behaviour, or the occurrence or re-occurrence of an actual or potential criminal offence(s), of individuals, groups or locations.
    - In the context of migration, when these systems are used to determine whether people on the move present a "risk" of unlawful activity or security threats, including the "risk" that people will overstay the duration of their visa; or when these systems are used to interdict, curtail and prevent migration.
    - In the context of social security, when these systems are used to determine eligibility or to screen claimants for "risk" of future fraud or misuse; or to "detect" past fraud by means of evaluations or classifications of people according to personal characteristics or based on data on their social behaviour.
- Until such regulation is set in place and regardless of changes to regulation, public authorities must urgently discontinue, stop or pause systems that fulfil the criteria listed above.
- States must prohibit the development, production, sale and use of risk profiling systems used in high-stakes contexts.
- To fulfil their positive obligations under international human rights law – including the obligation to adopt effective measures to review government policies and to amend, repeal or nullify laws and regulations that may lead to or perpetuate discrimination based on, but not limited to, race and ethnicity, including racial profiling – states must: explore all alternative options for ensuring that fraud in social security systems is effectively prevented, in a manner that is not discriminatory and is consistent with states' human rights obligations. This includes, for example, designing welfare systems in a way that protects against fraud; simplifying administrative procedures to reduce unintentional claimant errors and offering accessible help desks and plain-language guidance to

ensure claimants understand their reporting obligations. States should only ground investigations for fraud in concrete and individualized evidence that someone has committed fraud, rather than relying on risk scores.

11.2 SAFEGUARDING MEASURES FOR PROFILING IN OTHER DOMAINS

Although other domains might be perceived as less risky, the very use of profiling tools could increase the likelihood of human rights harms in situations that are not within the focus of this report, such as in fraud prediction or detection in the financial sector.

11.2.1 DUE DILIGENCE

- States should mandate independent human rights and data protection impact assessments for algorithmic profiling systems and ensure that profiling systems and other digital technologies are used in line with human rights law and standards, including on privacy, equality and non-discrimination, social and economic rights, as well as data protection standards, and that they are never used in ways that could lead to people being discriminated against or otherwise harmed. This impact assessment should include, at the very minimum, an evaluation of the discriminatory effects on marginalized groups: low-income groups and racialized groups, including people on the move and people who have been granted refugee status. The process should involve relevant stakeholders including independent human rights experts, individuals from potentially affected, marginalized and/or disadvantaged communities, oversight bodies and technical experts.
- States must draw clear red lines on and prohibit the development, production, sale and use of digital technologies that are incompatible with human rights. Require in law that technology companies carry out ongoing and proactive human rights due diligence to identify and address human rights risks and impacts related to their global operations, including by legally requiring human rights impact assessments of any public sector use of automated and algorithmic decision-making systems.
- Unless the governmental or private organizations that use or intend to use such profiling systems can demonstrate that the systems are compatible with international human rights standards, they must urgently discontinue, stop or pause the use or development of the systems, and reinstate them only if the right to non-discrimination can be guaranteed and the necessary safeguards are in place to ensure the protection of all other rights, including the right to privacy and the right to effective remedy.
- States should undertake proactive, ongoing human rights impact assessments throughout the lifecycle of algorithmic technologies, both before and after the roll-out and implementation of new systems, so that risks can be identified during the development stage, and human rights abuses and other harms identified immediately once the technologies have been implemented.
- States should critically assess whether automation and deployment of AI is the correct and most appropriate approach to reaching public policy or other stated aims. Avoid technological solutions for solving complex societal systemic issues such as poverty, social marginalization or disenfranchisement. Identify underlying systemic problems that require attention, acknowledge the limits of proposed technological solutions and explore alternative solutions and approaches.
- States should factor in and address the multiple and intersectional forms of discrimination faced by many groups – including women, people living with disabilities, older people, people living in poverty, children, and people belonging to racialized and minoritized communities such as refugees and people on the move – when trying to claim their human rights, and the specific barriers they may face when interacting with automated decision-making and/or when trying to appeal against a decision made by these systems.

11.2.2 TRANSPARENCY

- States should create and maintain publicly available and accessible databases for reporting on the development and deployment of profiling technologies that can have an impact on human rights.

- States should oblige developers and deployers of profiling systems with human rights impact to register themselves and the given system in relevant public databases, including during testing of AI systems in real-world conditions;
- States should oblige providers and deployers of profiling systems to proactively disclose information needed to assess the human rights impact of their systems, including source code when applicable, and to respond when requested by public interest organizations, including through Freedom of Information requests.
- States should ensure that rights holders are informed when automation or semi-automation is used to process their data. The information provided should be concise, easily understandable, and accessible, including for persons living with disabilities, people who are not digitally literate, and marginalized communities. This information should include details about the system's purpose or task.

11.2.3 ACCOUNTABILITY, EFFECTIVE REMEDY AND REDRESS

- As part of states' obligations to guarantee the rights of access to justice and adequate remedy, people and groups who have been subject to data-based predictions, profiles or risk assessments by public authorities should have clear and meaningful judicial and non-judicial routes to challenge those decisions.
- States should ensure that public interest organizations can support affected people seeking remedy, and can lodge cases on their initiative.
- States should ensure that comprehensive and independent human rights oversight mechanisms are in place. Oversight bodies should be granted adequate mandate and sufficient power, expertise and capacity to investigate and enforce, both reactively and proactively.

11.2.4 PARTICIPATION

- States should require the meaningful engagement of affected communities, civil society organizations and human rights experts in the development and deployment of profiling technologies, as well as in the development, implementation, monitoring and evaluation of relevant regulation.

AMNESTY INTERNATIONAL IS A GLOBAL MOVEMENT FOR HUMAN RIGHTS. WHEN INJUSTICE HAPPENS TO ONE PERSON, IT MATTERS TO US ALL.

CONTACT US

contactus@amnesty.org

+44 (0)20 7413 5500

JOIN THE CONVERSATION

www.facebook.com/amnesty

@Amnesty

AUTOMATING SUSPICION

RISK PROFILING AS A SMOKESCREEN FOR STRUCTURAL DISCRIMINATION AND INEQUALITY

This report analyses Risk Profiling systems under International Human Rights Law (IHRL) and calls for a ban on the use of risk profiling systems in high-stakes contexts such as migration, social security, and policing. Amnesty International defines risk profiling as an assessment, evaluation or calculation (sometimes called "prediction") of the likelihood that individuals or groups will violate a law or rule. Over the past decade, such systems have been developed and deployed by public agencies to perform a range of state administrative tasks.

This report provides an overview of the systemic issues inherent to risk profiling systems, a framing and argumentation to support further investigative work on real-world cases, and provides rights holders, human rights advocates, civil servants, oversight authorities, lawyers and judges with trustworthy scientific and legal arguments to contest the use of risk profiling by states and other entities in high-stakes contexts. The report is interdisciplinary, drawing on case studies, multiple academic disciplines, as well as situating the technology in its historical and social context. Further, the report explores the often overlooked structural and intersectional effects of discriminatory risk profiling and illustrates how technology can perpetuate discrimination and inequality by providing states with a false veneer of objectivity. It also addresses some of the most commonly proposed solutions to the issue of discriminatory profiling and highlights their limits.

INDEX: POL 40/1096/2026
JUNE 2026
LANGUAGE: ENGLISH

amnesty.org

AMNESTY INTERNATIONAL