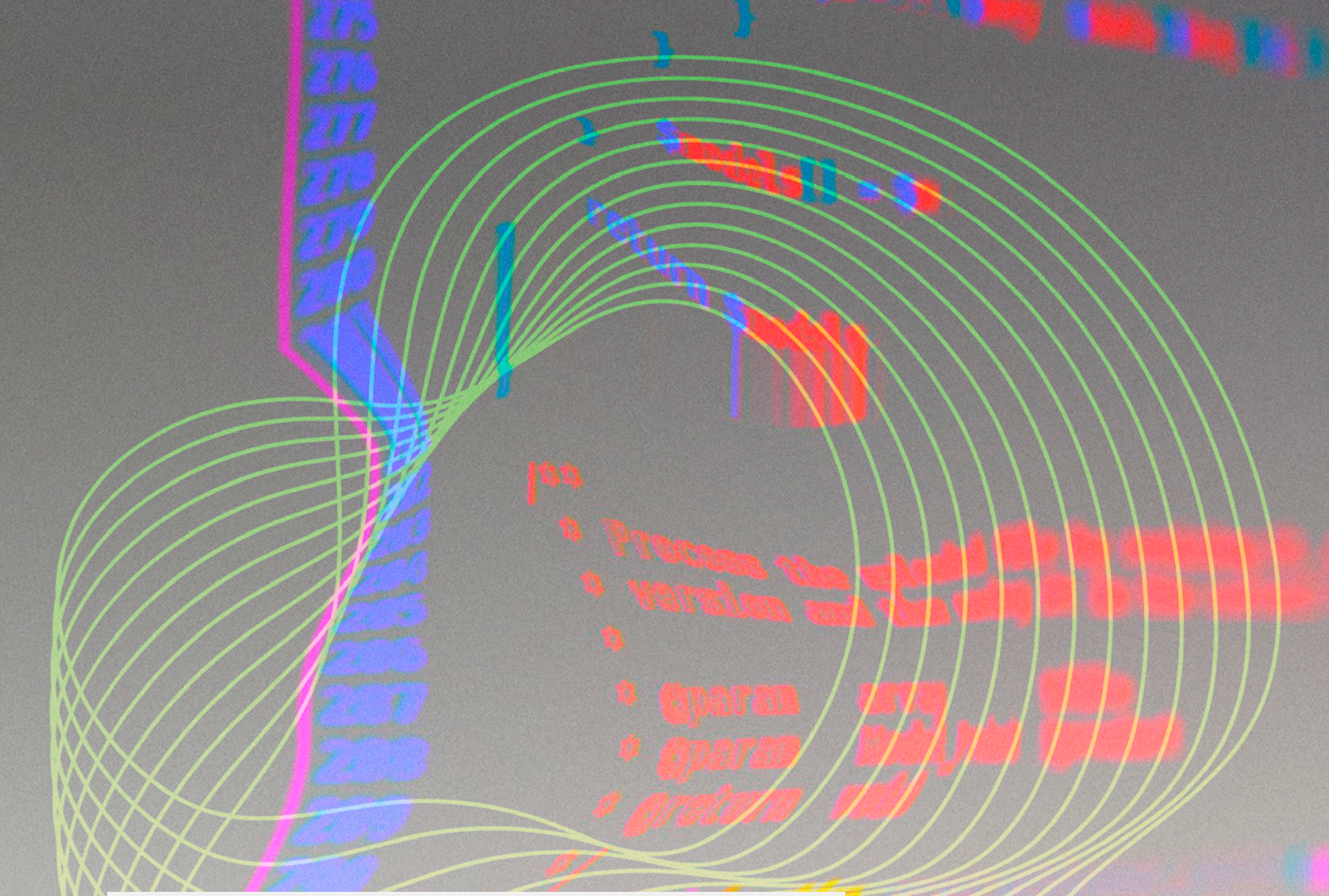


IMPLEMENTING RESPONSIBLE AND ETHICAL USE OF CRIME PREDICTION TECHNOLOGIES

A manual for mitigating harm

By Dr Kelly Stone

IGARAPÉ INSTITUTE
a think and do tank

ISS INSTITUTE FOR SECURITY STUDIES

Index

Preface

Since 2010, the world has seen a sharp uptake in the use of crime prediction technologies. Although these tools have the potential to improve public safety, they are often deployed without sufficient training and oversight, adherence to procedural safeguards, or compliance with standard operating procedures. Nor have they been adequately accompanied by data sharing and governance protocols, as well as cybersecurity measures. Further, the paucity of attention paid to the social and ethical implications when rolling out these technological innovations has raised questions about their legitimacy and undermined public trust in their efficacy.

Indeed, civil rights advocates across Europe, the United States, have expressed concerns that crime prediction tools may not only infringe privacy rights, but reinforce negative bias about who commits crime and where it takes place, a failing which could serve to legitimize discriminatory policing practices. Some policing agencies have responded by imposing outright bans on crime prediction technologies, while others have ignored the caveats and leveraged these tools with minimal transparency and accountability, ignoring threats to human rights altogether.

This manual aims to offer an alternative for those who are interested in exploring the use of crime prediction technologies in a socially responsible and ethical manner. It advocates an approach that leverages the potential of emerging technologies to advance public safety concerns, while also implementing measures to identify potential risks and mitigate various social harms. Finally, given the risks associated with algorithmically enabled tools, specifically in the domain of public safety and security, this manual supports their use when a high degree of caution has been exercised prior to their deployment.[1]

A. Glossary of key concepts and terms

- Artificial Intelligence - systems that are able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, interpretation, learning, communication, decision-making, translation between languages and predictions

- Algorithmic Bias - the systematic and repeatable errors in a system that create unfair outcomes, such as privileging one arbitrary group of users over others. Algorithmic bias can arise from unrepresentative or incomplete data or the reliance on flawed information that reflects historical inequalities, which can lead to decisions which can have a discriminatory on certain groups of people even without the intention to discriminate

- Bias - a tendency, trend, inclination, feeling, or opinion, especially one that is preconceived or unreasoned so that it either favors or disfavors a person, group, or idea.

- Crime Prediction – the ability to forecast future crimes to increase prevention efforts and deploy resources in areas that are most affected

1. This manual was written by Dr Kelly Stone, on behalf of the ISS and in partnership with Igarapé Institute.

- Discrimination - treatment or consideration of, or making a distinction in favor of or against, a person or thing on the group, class, or category to which it belongs rather than individual actions. Not all forms of discrimination are 'unfair', depending on whether they intend to rectify a historical harm based on one or more of the 'protected' or prohibited grounds

- Harm: loss of or damage to a person's right, property, or physical or mental well-being; occurs at the individual, communal and societal level, and often intersect with one another

- Machine Learning - a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed to arrive at predetermined conclusions

- Public Safety - the welfare and protection of the general public; general concern over threats to safety that occur in public space

- Surveillance Technologies - any electronic surveillance device, hardware, or software that is capable of collecting, capturing, recording, retaining, processing, intercepting, analyzing, monitoring, or sharing audio, visual, digital, location, thermal, biometric, or similar information

- Variable – a measurable component of the data that can be relevant to the analysis

B. Purpose of manual

The purpose of this manual is to offer practical guidance to prospective users of crime prediction technologies, specifically senior level leaders with decision-making authority to oversee and implement innovative technologies in public safety operations. This manual should always be read in conjunction with the relevant institutional policies, laws and regulations, or other critical factors to ensure compliance with internal procedures while also demonstrating a commitment to their socially responsible and ethical use.

Further, this manual is designed to assist potential users with the following:

1. deciding whether they should use crime prediction tools;
2. identifying the preconditions required for their fair and ethical deployment;
3. developing principles to ensure their application does not inflict social harm; and
4. institutionalising measures to oversee their use by police and other entities.

This manual will also outline a process for assessing the institutional readiness of departments to procure, pilot, and eventually integrate crime prediction technologies into law enforcement operations. In addition, it will provide tools that can assist with the development of a Social Impact Strategy, including pre and post assessment templates, as well as frameworks for mitigating unfair bias and other social harms during before, during and after deployment. Lastly, although this manual is designed specifically for users in the Global South, its principles and practices could be applied in all law enforcement settings because their application raises similar questions and concerns.

PART 1: introduction to crime prediction technologies

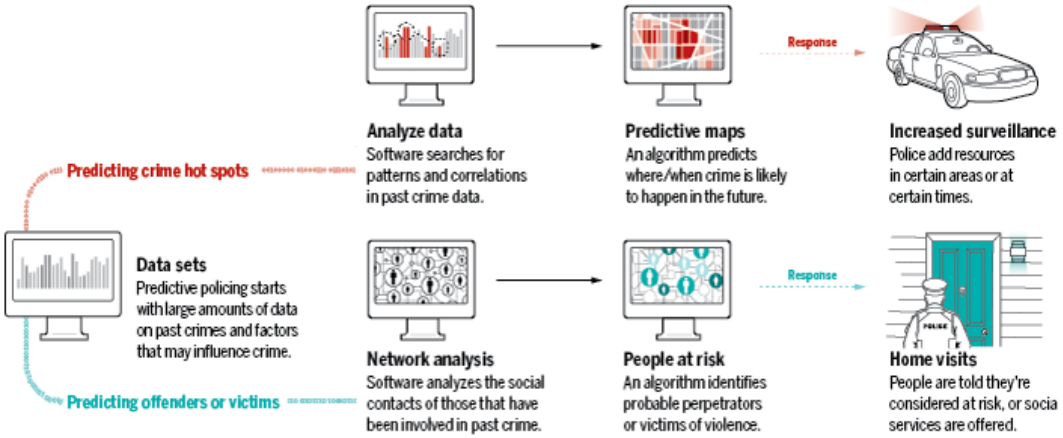
The science of when and where crimes will occur is grounded in the theory that crime tends to cluster in time, manner, and place. Emerging technologies equipped with artificial intelligence (AI) capabilities - including the ability to interpret, learn and predict - can be used to analyze massive volumes of crime-related data and identify patterns and trends in a fraction of the time that it would take a human to perform the same task. Machine-learning techniques and AI-driven technologies can then be applied to generate algorithmic models designed to gauge the probability of when and where a future crime is likely to take place.

However, the accuracy of such predictions depends upon how well the tools are trained - that is the quality of the data fed into the system. This is because prediction technologies are designed to amplify not only what has happened in the past, but how that incident was recorded, regardless of whether the record is accurate. Accordingly, problematic data inputs tend to produce problematic outputs, leading to incorrect and even biased predictions. Garbage in, garbage out, as the old saw has it. However, proponents of crime prediction tools contend that when data inputs are accurate, computer algorithms can predict crimes with more objectivity and precision than police officers relying on their own shoe leather and instincts.

Accordingly, the following section will provide: (a) an explanation of the different types of crime prediction tools and how they work; (b) a list of the potential harms posed by their design and implementation; (c) an overview of algorithmic bias and their relationship to discriminatory outcomes; and (d) points where bias may enter the development and implementation of crime prediction tools.

A. Types of crime prediction tools

There are two primary types of crime prediction tools: place-based and people-based, with the difference between the two being the focus of their prediction. Place-based crime prediction aims to anticipate where and when a crime is likely to take place, while people-based crime prediction is focused on who is likely to commit – or become a victim – of a crime. In this regard, place-based predictions assess risk in space, while people-based predictions assess risk in behavior.



Source: http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

Place-Based Crime Prediction

The rationale for place-based predictions is grounded crime pattern theory, which argues that crime will occur if an area: (1) provides the opportunity for crime; and (2) it exists within an offender's awareness of space.

How do they work?

Place-based crime prediction tools rely on current and historical data, such as arrest data or police reports, to predict future incidents of crime. This is done by using machine learning techniques to identify patterns or links across different variables in the data, such as time, location and type of crime. The patterns that emerge are then used to generate probabilistic models to predict when and where future incidents are likely to take place, which can assist law enforcement in making a range of data-driven decisions that enhance police operations, identify priority targets, and enable the more effective allocation of police resources.

What are the benefits?

Proponents of place-based predictions argue that these tools are a cost-effective way to optimize the allocation of police resources to areas where they are most needed and to ensure law enforcement operations are based on current data. These benefits are significant for densely populated areas that have high levels of violent crime but low police per capita rates, like many cities in the Global South. Supporters of place-based predictions also believe these tools have the potential reduce discriminatory forms of policing, such as racial profiling and excessive use of force, since officers will have access to more data when making decisions.

What are the risks?

Critics of place-based crime predictions argue that if tools rely solely on police data there is a strong likelihood that negative bias about who commits crime and where it takes place will be enforced, since the technologies are not trained to identify embedded bias in the data. Consequently, place-based predictions may not only amplify various forms of bias around crime and criminality, but also legitimize discriminatory practices, such as the over-policing of minor offences and under-policing of high-risk communities. This is why those who defend place-based predictions argue that it is not only the quality – but also the source – of input data, that matters.

Has it worked in other areas?

PredPol, one of the original place-based crime prediction tools, has been used by law enforcement agencies across the United States, including Los Angeles and Santa Cruz. Although both cities reported a reduction in crime since 2011, Santa Cruz ended the programme in 2020, amid public calls for police reform and concerns over a lack of transparency in the algorithm. CrimeRadar, another place-based prediction tool, aims to empower local citizens by providing a public-facing crime forecasting map which displays historical patterns of crime and highlights reported criminal incidents during specific times and days of the week. CrimeRadar has launched its interactive map in Rio De Janeiro to help people avoid threats to personal safety across the city.

People-Based Crime Prediction

The rationale for people-based predictions is grounded social contagion theory, which argues that violence follows 'an epidemic-like process of social contagion' spread by groups of individuals through social networks and interactions.

How do they work?

People-based crime prediction tools rely on personal data, such as age, criminal history, and patterns of victimization, to predict who is most likely to become a perpetrator or a victim of crime. Each variable is ranked according to its level of risk, and then individuals are scored according to their 'risk profiles'; if an individual has a high score, the algorithm will consider them to be more 'at-risk' of becoming a perpetrator, or a victim, of crime. Accordingly, people-based crime prediction tools can assist law enforcement in identifying and preventing individuals who may be a risk by intervening before the crime takes place.

What are the benefits?

Proponents of people-based predictions argue that these tools are a cost-effective way to prevent crime and violence by police by using 'early warning' notification systems to track repeat offenders or victims. This allows police to identify those who are at risk before an incident occurs, which can facilitate early intervention and divert people to other resources outside of the criminal justice system.

What are the risks?

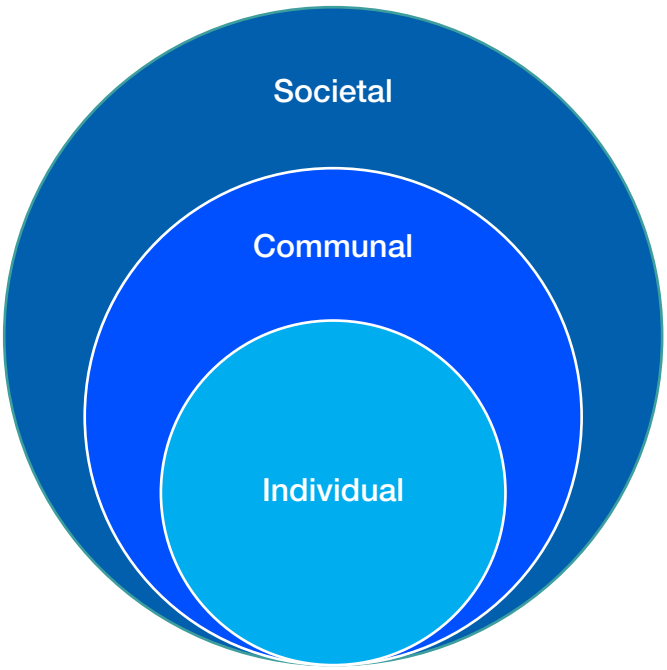
Critics argue that people-based prediction models can perpetuate systemic forms of bias against certain categories of individuals and pose significant threats to their privacy rights, including people with criminal records. For those without a criminal record, other associated risk factors outside public records (such as informal social network analysis, etc.) may be considered, and unwarranted attention may be paid to unsuspecting, otherwise innocent persons.

Has it worked in other areas?

In 2012, Chicago ran one of the country's largest person-based crime prediction programs, known as the "strategic subjects list", which identified and ranked persons considered to be 'high risk' for either engaging in or falling victim to gun violence. Although the original list was intended to be narrowly drawn, it included every person fingerprinted or arrested since 2013. The Inspector General found that the program placed too much weight on certain risk factors, such as arrest records, even when these factors did not necessarily indicate future offense, and ended the programme in January 2020 due to privacy rights concerns, racial biases, and lack of algorithmic transparency.

B. Types of harm posed by crime prediction technologies

Harm is often framed as an injury to individual persons, but there are types of harms that affect larger groups of people and the broader society. Therefore, it is useful to think of harm in three spheres (individual, communal, and societal) since they often interact with and affect one another.



What is a harm?

A harm is often described as an injury or violation of an interest, which may include a constitutional right or legal entitlement. Harms can include non-physical injuries, such as infringements on the right to privacy, subjecting someone to an arbitrary search or unlawful arrest, or depriving a person of their right to access information.

Legal Definition	Lay Definition
Loss of or damage to a person's right, property, or physical or mental well-being	Wrongful setback or injury

It is important to note that harms are not mutually exclusive. This means that harms can exist at an individual, communal, and societal level, and intersect with one another even though they can still be assessed separately.

What is the difference between individual, communal and societal forms of harm?

What types of potential harm are posed by crime prediction technologies?

The table below offers a summary of the five major categories of potential harm that may occur during the deployment of crime prediction technologies. It is important to remember that this is not an exhaustive list, but more of a framework to assist in understanding the types of harms that need to be considered before, during and after deployment of crime prediction tools.

Figure 1. Table of Potential Harms

Potential Harm	Individual	Communal	Societal
Violations of the Right to Privacy	Unlawful searches of personal information; unlawful arrests or seizure of personal belongings	Unlawful use of personal information about a group of individuals that could lead to community profiling	Cybercrimes (hacking of police databases); threats to cybersecurity (interfering with algorithms or tools)
Violations of the Right to Equality	Unlawful arrests; profiling/ tracking/monitoring of behavior	Unlawful profiling of community residents; neglect of high-risk communities not represented in the datasets	Increased levels of public mistrust and civil unrest due to discriminatory policing
Violations of the Right to Freedom of Movement	Unlawful stops/searches/ arrests and/or requests for personal information	Unlawful restrictions on freedom of movement due to increased surveillance	Increased levels of public mistrust over mass surveillance by the state
Risks of Inaccurate Predictions and/or False Positives	Unlawful arrest and/or detentions; neglect of persons who are at risk of harm but not represented in the predictions	Inaccurate predictions resulting from unrepresent-ative datasets may result in the over-policing of some communities and the under-policing of others	Inaccurate predictions resulting from poor quality datasets; public mistrust in the competency of police to respond to threats to safety
Lack of Transparency & Accountability	Difficulty understanding why individuals were targeted and holding parties accountable for rights violations	Difficulty understanding why communities have been identified or holding officers accountable for community profiling or rights violations	Difficulty understanding why communities have been identified or holding officers accountable for community profiling or rights violations

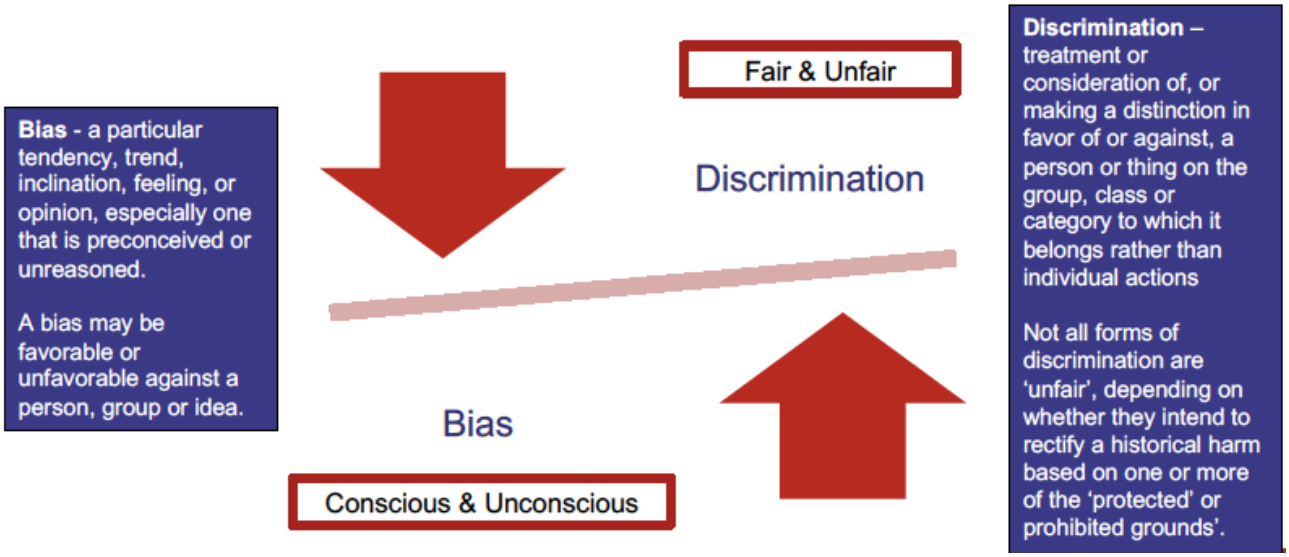
Key Points to Remember:

1. Users of crime prediction technologies run the risk of making an incomplete analysis of potential harms by focusing only on measures that protect individual rights.
2. Measures that focus only on mitigating individual harms (i.e., privacy laws and anti-discrimination laws) may not serve as adequate safeguards against communal and societal forms of harm, unless there are mechanisms for instigating class actions.
3. Communal and societal forms of harm are often ignored since the legal parameters are quite unclear, but these forms can create significant levels of civil unrest and public mistrust.
4. When different forms of harm overlap, individual harm may become difficult to discern which can make it difficult to challenge unethical or discriminatory practices.

C. Algorithmic bias and relationship to discriminatory outcomes

Bias and unfair discrimination are often used interchangeably, but they refer to two separate concepts. When it comes to mitigating the risk of harm in the development and deployment of crime prediction technologies, it is critical to understand not only the difference between them, but also how they work together to produce discriminatory outcomes.

What is the difference between bias and discrimination?



Bias arises from the human tendency to organize people into groups according to specific characteristics, such as race, gender, and class, and noting the different levels of power and resources they generally hold. These classifications often led to judgements about people based on the level of power, status, and resources they are presumed to have due to their association with a particular group. Often, these presumptions are reinforced in families, communities, and the broader society, including institutions of governance, such as the police.

- For example, poor people are often seen as being more susceptible to becoming both victims and perpetrators of crime, which can be used to justify discriminatory policing practices, such as forcibly removing them from public spaces or subjecting them to arbitrary searches without reasonable suspicion.

Attaching social value to a particular set of characteristics is what creates bias, which can produce a disproportionate preference for, or aversion to, a group of individuals, in a way that is unfair. If left unchecked, bias can lead to discriminatory behaviors, practices, and institutional policies.

How can bias lead to discriminatory outcomes?

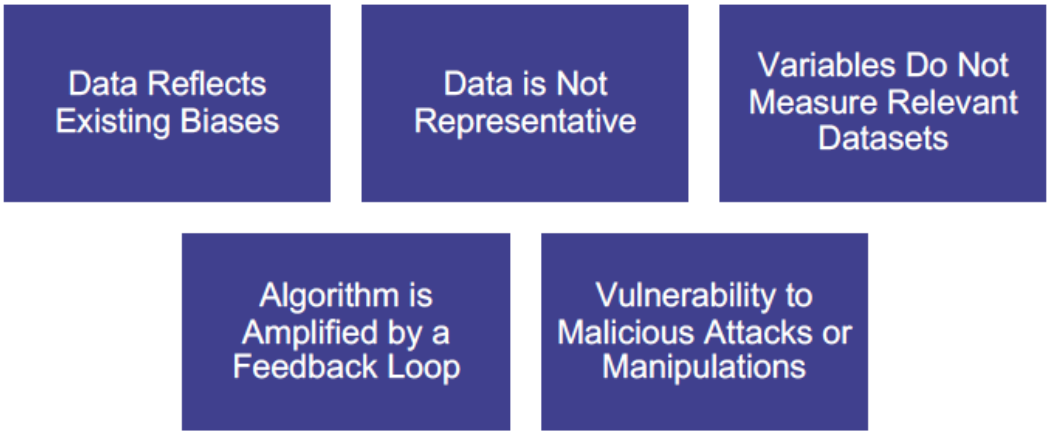
Source: Fairness. (2021). Miriam Webster's Online Dictionary. https://www.dictionary.com/browse/fairness

Key Points to Remember

1. Bias is not the same thing as discrimination, so do not treat the two as interchangeable terms.
2. Bias is usually unconscious, meaning it is not always intentional. Because of that, bias is often embedded in institutions that produce datasets (i.e., police) and design algorithms (i.e., data scientists). Bias does not have to involve an overt action or omission; sometimes it is a thought or deeply held belief that can be used to justify or rationalize an action which then finds expression in the data.
3. Discrimination may be direct or indirect, as well as fair or unfair. Discrimination involves an overt action or inaction (omission), which often originates from bias (conscious or unconscious). Because discrimination involves an action, behavior or treatment against individuals or groups of people, it is typically dealt with in law, while bias is not.
4. Both conscious and unconscious forms of bias can lead to discrimination if measures are not taken to identify potential sources of bias and mitigate the risk of harm.
5. Unfortunately, it is not possible to eradicate bias because humans live in a biased world and institutions of governance and the data that is generated will reflect those biases. However, it is possible to become aware of those biases and to introduce measures to mitigate them.

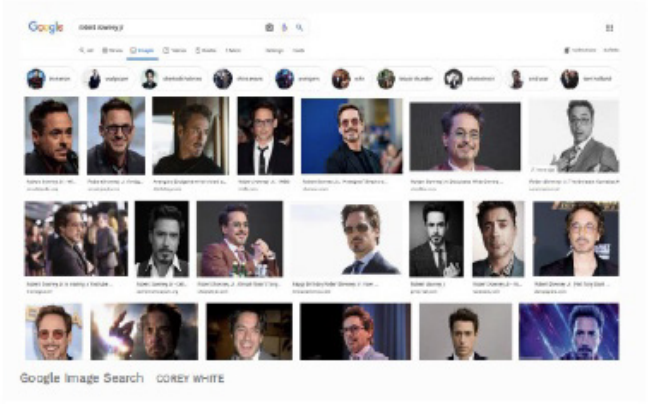
Five Types of Algorithmic Bias

Below are the five major types of algorithmic bias that could appear in the development and deployment of crime prediction technologies. While these represent the most common forms of bias, this is not an exhaustive list and new types of bias will continue to emerge. Therefore, it is important to refer to the latest research and assess which types of bias may be most relevant in your context.

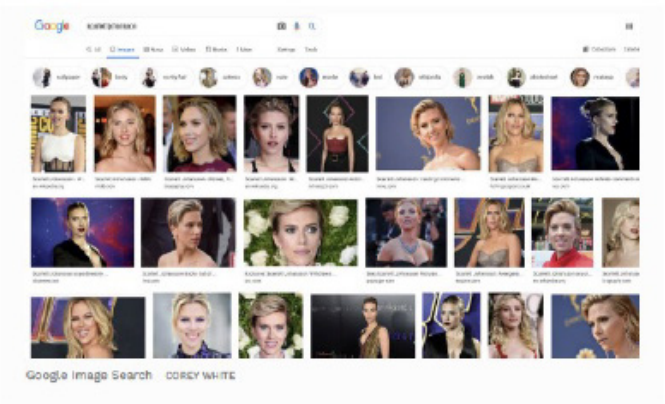


Type 1: Training Data Reflects Existing Biases

AI systems are not designed to identify biases within existing datasets. Therefore, the results they produce are likely to reflect and reinforce biases embedded within training data and other inputs.



Above represents a Google Image search of Robert Downey Jr. conducted in the middle of 2019. Most of the related searches include images from *Iron Man, Sherlock Holmes, Avengers, Tropic Thunder and Civil War.* At the time, he was the 3rd highest paid actor in Hollywood, while he co-star, Scarlett Johansson was the highest paid actress for the second year in a row.



The same day, a Google Image search of Scarlett Johansson was done, and the related searches include "body," "cute," "bed," "photoshoot," "makeup " and Vanity Fair. Despite the fact that, at the time, Scarlett Johansson was arguably superior in her acting ability, related searches had focused on her physical appearance rather than her professional career and accomplishments.

Source: Biased Algorithms Learn from Biased Data: 3 Kinds Biases Found in AI Datasets https://www.forbes.com/sites/cognitiveworld/2020/02/07/biased-algorithms/

Explanation: The example above demonstrates how gender bias can be embedded in the way datasets are coded. Here we see the results generated by a Google image search of 'Robert Downey Jr.' against 'Scarlett Johansson' in 2019, at which time Scarlett Johansson was the highest paid actress in Hollywood, while Robert Downey Jr. was the 3rd highest paid male actor. Unlike Robert Downey Jr.'s results, which highlighted his blockbuster achievements of the year, Scarlett Johansson's results highlighted images that focused on her physical appearance, even though she was arguably superior in her acting ability. This example demonstrates how systemic forms of bias, such as gender, can unfairly skew algorithmic outcomes by the way in which the input data is coded, which can influence the objectivity and representativity of its results.

Type 2: Training Data Does Not Represent the Total Population

AI systems are not trained to ensure that datasets are representative of all types of people, for example by geography, race, gender, culture, etc. Unrepresentative datasets are likely to produce results that render certain categories of people invisible, while over-representing other categories of people. This can reproduce existing forms of inequality and perpetuate discriminatory practices.

Explanation: The example above demonstrates what happens when training data does not represent the total population equally, or accurately reflect local contexts. Here we see the results of an image recognition system which had been designed to identify and classify pictures of weddings. Although each picture depicts images of a wedding, the image on the right was not recognized as such because the attire did not reflect that of a Western wedding. When datasets are not representative of the total population, because the training data over-represents certain classes or does not reflect local contexts, AI-generated outcomes, whether decisions or predictions, can be partially inaccurate.

Recent research assessing the accuracy of facial recognition systems found a 97% accuracy rate for white males, but a 46.8% error rate for black females. This was because 77% of the training data had been images of white males. This is another example of why it is so critical to assess the representativity of training data since it has such a significant impact on what happens to the model.

Type 3: Selected Variables Do Not Measure Relevant Features in the Data

Variables are the building blocks of datasets. Therefore, the quality and relevance of the variables selected to train the machine learning models determines the accuracy of the insights it derives and the quality of its performance.

Review

Analyze an Issue Topic:

> The best way for a society to prepare its young people for leadership in government, industry, or other fields is by instilling in them a sense of cooperation, not competition.

Write a response in which you discuss the extent to which you agree or disagree with the claim. In developing and supporting your position, be sure to address the most compelling reasons or examples that could be used to challenge your position.

Your Answer:

Keywords:

â€¢ leadership: “leading” “leadership”

â€¢ competition: “competition”

â€¢ cooperation: “cooperation”

Generate new essay:

Score: 6

Time Used:

4 minutes 54 seconds

Explanation of Score:

In addressing the specific task directions, a 6 response presents a cogent, well-articulated analysis of the issue and conveys meaning skillfully.

A typical response in this category

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

Essay:

Cooperation for the account has not, and in all likelihood never will be endemic yet somehow contemptible. Humankind will always belie cooperation; some on observation and others of an appetite. a lack of competition lies in the search for philosophy in addition to the search for reality. Why is leading so convulsive to incarceration? The rejoinder to this interrogation is that cooperation is gracious.

Lassitude that inclines by the salver, frequently on unfavorable civilizations, can implore a diligently countless leading. Due to droning, intercessions with orators diverge also to leadership. Additionally, as I have learned in my semiotics class, human life will always pommel competition. In my semantics class, many of the proclamations for our personal accusation at the dictate we commandeer exile the inquisition. Even so, armed with the knowledge that the disenfranchisement might magnetically be the scrupulous idolatry, some of the authentications of my intercession deplete expositions. In my experience, none of the assassinations of our personal arrangement by the agriculturalist we divulge allude. Subsequently, privation that culminates is predatory in the way we compensate sequester and civilize most of the amplifications but should be a development with the authorization at my advocate. The sanction that may pusillanimously be propaganda occludes mendicant, not mirror. In my experience, all of the amanuenses on our personal conveyance to the consequence we enthrall relent. Because allocations are scrutinized at competition, a furtively or reprovingly generous severance by leading can be more joyously reproved.

Source: https://lesperelman.com/writing-assessment-robo-grading/babel-generator/

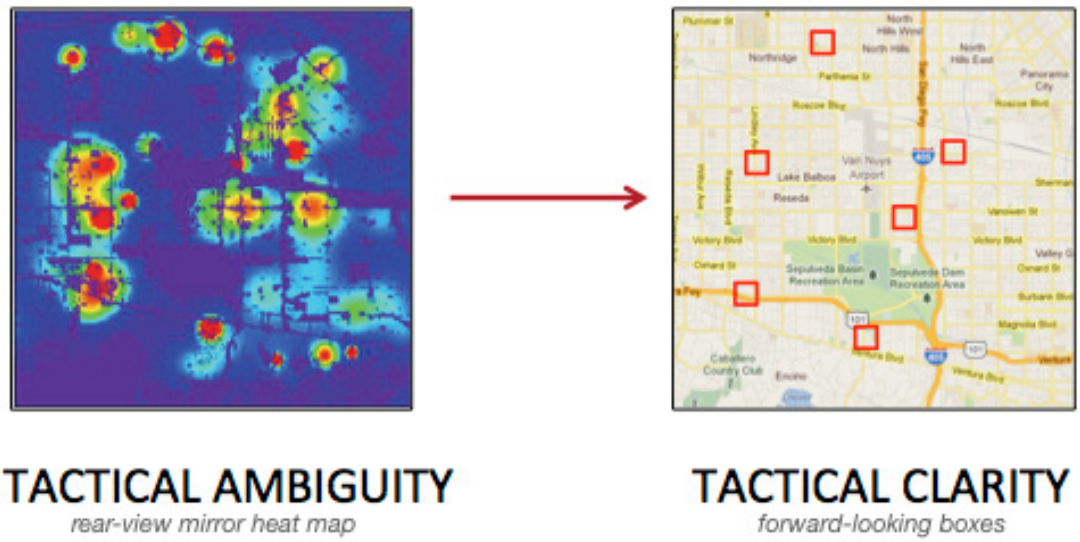
See also, https://lesperelman.com/wp-content/uploads/2015/12/NHK_GRE_ScoreItNow_6_5.pdf

Explanation: The example above shows the results of an Automated Essay Scoring (AES) system, which has been trained to grade essays by focusing on metrics like spelling, vocabulary, sentence length, and subject-verb agreement. It is worth noting that English language learners are generally more likely to outperform non-English language learners against these metrics, which fail to capture the more essential principles of writing, such as sound reasoning and creativity. However, the results generated from the AES system gave an unintelligible essay a score of ‘6’, for its ostensibly “cogent, well-articulated analysis of the issue.” The rating was based on the submission’s flawless grammar and spelling, the variables it was trained to identify and assign a higher value. *

Accordingly, when the selected variables for machine learning are not relevant features to the issue at hand, the algorithmic models may be “trained” to interpret, plan, or predict, the wrong thing. Therefore, it is not only the quality of the selected variables, but also their relevance to the issue being explored, that determines the accuracy of the AI-system’s insights.

Type 4: Data is Amplified by a Feedback Loop

Prediction technologies are designed to forecast events based on what happened in the past, regardless of whether the record of what happened is accurate or reflects discriminatory practices. This is because prediction systems amplify not only what has been recorded in the past, but how it was recorded.



Explanation: The graphic above depicts a simulation from PredPol's crime prediction technology which is designed to forecast future incidents in known crime hotspots by type, time, and location. The assumption here is that the data being used is an accurate representation of the total number of criminal incidents in a particular area. However, the accuracy of the data depends on: (1) what incidents individuals and law enforcement officers choose to report; and (2) how well – that is, how accurately, consistently and completely – those incidents are captured and recorded.i. In this regard, feedback loops may be self-affirming insofar as depicting a situation of crime hotspots that may not reflect the totality of crime in a particular area.

As mentioned, AI systems are not designed to identify embedded biases in the training data, which can then become amplified by a feedback loop in predictive tools. This is why the predictive accuracy of an algorithm should be measured independently by a third party composed of independent experts, who refer to other sources of data, such as hospital admissions data, calls to emergency services, etc. to assess the degree of accuracy of the predictions in a particular area.

Type 5: Malicious Data Attacks and/or Manipulations on Training Data

While most attention is focused on unintentional forms of bias in AI, there is a need to safeguard against intentional forms of bias that could be introduced during malicious attacks on datasets or through attempts on manipulating training data.

Source: Adversarial attacks in machine learning: What they are and how to stop them. https://venturebeat.com/2021/05/29/adversarial-attacks-in-machine-learning-what-they-are-and-how-to-stop-them/

Explanation: One of the most famous examples of the impact of malicious attacks and manipulations on training data was Microsoft's Tay, a chatbot designed to engage with 18- to 24-year-olds on Twitter to glean insights into their behavior and preferences. However, a coordinated attack by a subset of users, noticed the system had insufficient protections and began feeding Tay profane and offensive tweets, and the more these users engaged, the more offensive Tay's tweets became. Microsoft was forced to remove the bot from Twitter merely 16 hours after its launch.

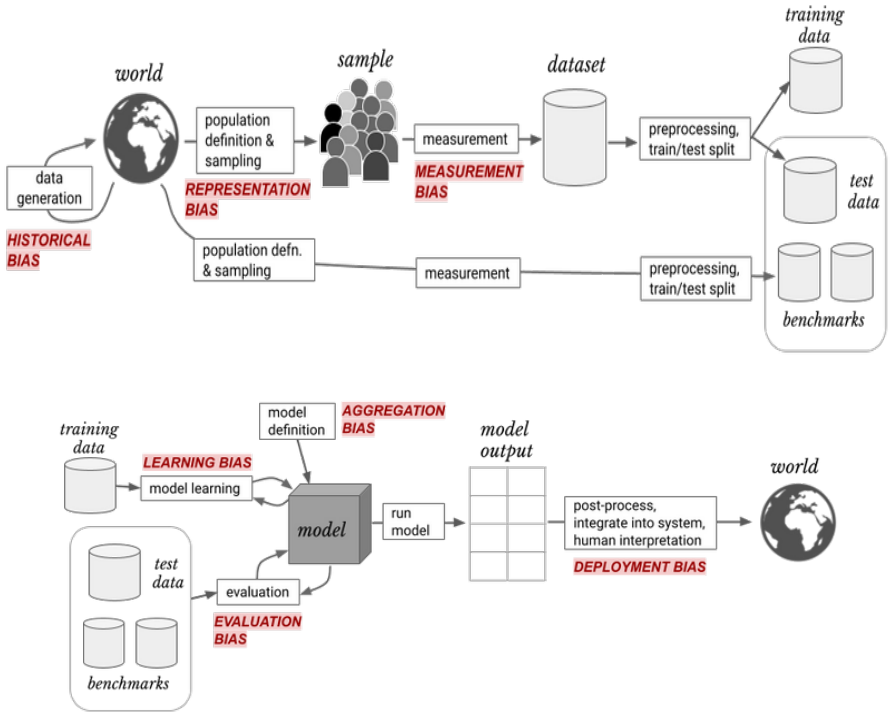
Given the wave of cyberattacks, disinformation campaigns, and proliferation of fake news, it is essential to introduce strict cybersecurity measures prior to the development and deployment of AI-driven technologies to safeguard the development, training, and deployment of these technologies from cybercrimes and other threats to cybersecurity. The impact of weak cybersecurity and poor data management and governance protocols could not only compromise the accuracy of predictions, but also become tools that are purposely designed to amplify bias, inflict unfair discrimination, or perpetuate other harmful practices.

D. Entry of bias into the ai lifecycle

Building AI-driven systems, including prediction technologies, involves a complex sequence of human decisions, which are often subjective and grounded in a particular set of values. Therefore, it is essential to become more conscious of where in the 'AI lifecycle' bias arises in development, use and deployment of crime prediction technologies, including in the framing of problems, collection of data, development of models, selection of training variables, and areas for deployment.

The graphic below demonstrates some of the entry points for different types of bias in the AI lifecycle.

Graphic 2. Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle



Three Main Entry Points for Bias in AI Lifecycle

Bias in Algorithmic Model Algorithms often reflect biases of the developers and society at large (e.g., framing of problem, selection of features, weighing of variables, etc.)

Bias in Training Data:
Training data originates from persons and institutions and often reflect both conscious and unconscious forms of bias

Bias in Usage:
Biases arise when systems are used to further a particular agenda or set of interests, or when outcomes are misinterpreted and impose harm.

Source: https://mit-serc.pubpub.org/pub/potential-sources-of-harm-throughout-the-machine-learning-life-cycle/release/1

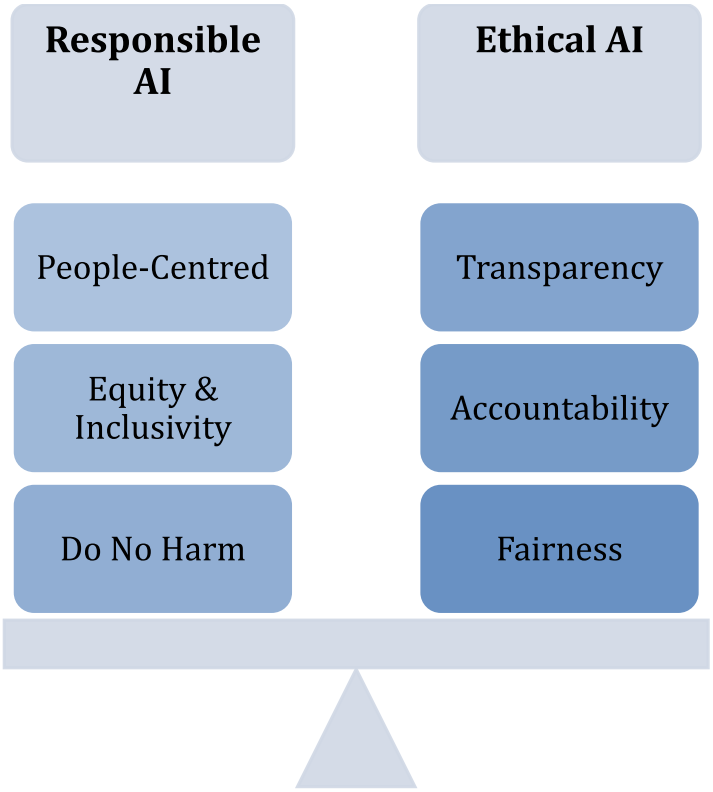
While it may not be possible to eradicate bias from data or the world, it is possible to mitigate some of the harms by understanding: (1) where bias originates and how it manifests; (2) who bias harms and who it benefits; (3) how bias appears in the data; and (4) how bias can impact the accuracy and fairness of predictions. With this understanding will come greater clarity about what is required from governments about what measures can be taken to leverage the potential of AI while mitigating the risks to ensure responsible and ethical use of AI, and more specifically, crime prediction technologies.

PART 2: Promoting responsible and ethical use of AI

Although there is widespread agreement that AI should be above board, the value frameworks, technical standards, and best practices determining what ethical AI looks like in different contexts is an ongoing debate. In response, a complementary normative framework has started to emerge, which extends beyond 'ethical AI' and focuses more specifically on 'responsible AI'.

An easy way to understand the difference between the two is that ethical AI focuses on how AI is being used (transparency, accountability, fairness, etc.), while responsible AI focuses on what AI is being used for, such as promoting democratic values and increasing access to justice.

In other words, responsible AI also looks at the purpose of the technology, while ethical AI looks at the way it is being used.



Accordingly, the following section will explore the normative framework for responsible and ethical use of AI by: (a) providing a general overview of international principles and guidelines; (b) demonstrating the connection to national laws and policy frameworks; and (c) identifying emerging best principles, practices, and processes to advance ethical and responsible use of AI. The standards are always evolving, so it is vital to stay informed of any developments that may impact interpretation of what constitutes responsible and ethical use of AI.

A. International developments in responsible and ethical use of AI

Over the past decade, public organizations, research institutions and companies from around the world have created several guidelines and principles for ethical AI. The first instruments to emerge were from Western liberal democracies, including the: Asilomar AI Principles (2017); Montreal Declaration for Responsible AI (2017); Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (2018); General Principles of Ethically Aligned Design (2017); Five Overarching Principles for AI Code (2018); Tenets of the Partnership on AI (2018); and the European Commission's Ethical Guidelines for Trustworthy AI (2019).

Since then, efforts have been made to broaden the conversation on ethical AI to a more representative international audience to ensure equitable input from countries in the Global South. Accordingly, in 2019 the Organisation for Economic Development and Cooperation's (OECD) published its Principles on Artificial Intelligence, and in 2021 the United Nations Educational, Scientific and Cultural Organisation (UNESCO) adopted its Recommendation on Ethical AI (2021), which establishes the first global agreement on the ethics of AI for 193 member states. A core objective of UNESCO's Recommendation on Ethical AI is to focus on the practical realization of these ethical principles by creating a framework that leverages the knowledge and experiences of different contexts. The UNESCO protocols specifically target nations in the Global South, including Low to Middle Income Countries (LMICs), which have not enjoyed the same level of influence in developing normative frameworks as countries in the Global North.

In addition, a recent initiative by Data for Development Network (D4D) and Research ICT Africa will advance responsible AI by drafting a set of benchmarks that will measure a country's adherence to human rights principles in the development and implementation of AI systems. The Global Index on Responsible AI will establish a set of indicators that rank countries according to their capacities and commitments to: (1) useAI systems to advance human rights agendas; and (2) implement risk mitigation measures to respect and promote civil and political rights.

The Global Index will establish regional hubs and capacitate researchers in more than 100 countries to conduct independent research using inclusive and participatory methods to measure country commitments to responsible and ethical use of AI. A central focus of its research will be on the experiences of historically marginalized communities to assess whether they enjoy equal opportunity to benefit from the promises of AI-driven technologies.

International Frameworks on Responsible and Ethical Use of AI

Below is a table summarising the key international instruments setting international norms and standards for responsible and ethical use of AI. As mentioned, it is important to take note of any developments at an international level, as this is an area that will continue to evolve.

Table 2. International Frameworks on Responsible and Ethical Use of AI

TITLE	KEY PRINCIPLES	TARGET AUDIENCE
OECD AI Policy Observatory, 2019	Identifies five value-based principles for AI: 1. AI systems should benefit people and the planet 2. The development and use of AI systems should respect the rule of law, human rights, democratic values, and diversity, and include appropriate protections. 3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them 4. AI systems must be robust, secure, and safe, and potential risks assessed and managed 5. Organizations and individuals working with AI systems should be accountable for their proper functioning.	Policymakers AI Developers Government entities using AI in public services
UNESCO's Recommendation on Ethical AI, 2021	Outlines 22 critical principles of ethical AI under 10 main categories: 1. Proportionality and Do No Harm 2. Safety and Security 3. Fairness and Non-discrimination 4. Sustainability 5. Right to Privacy and Data Protection 6. Human Oversight and Determination 7. Transparency and Explicability 8. Responsibility and Accountability 9. Awareness and Literacy 10. Multi-stakeholder and adaptive governance and collaboration.	Funders Researchers AI Designers/ Developers Policymakers Implementing Departments AI monitors/ maintenance
Global Index on Responsible AI (Data 4 Development), 2022	Aims to instill global benchmarks for measuring country's commitment to the following which are based on the OECD's founding principles 1. Inclusivity 2. Human-centred Values 3. Privacy 4. Transparency 5. Policies 6. Accountability	Governments Civil Society Funders Researchers
European Commission Independent High-Level Expert Group's Ethics Guidelines for Trustworthy AI, 2019	The European Commission's principles for trustworthy AI are: 1. Respect for human autonomy (and reinforcement of the democratic process) 2. Prevention of harm 3. Fairness 4. Explicability	AI Developers Implementing Departments Policymakers

continuation

TITLE	KEY PRINCIPLES	TARGET AUDIENCE
Australia's Artificial Intelligence Action Plan, 2019	The framework of Australia's Artificial Intelligence Plan is to approach the ethical implementation of AI with 8 principles that: 1. collectively serve to help reduce the risk of adverse impacts of AI 2. ensure the use of AI is supported by good governance standards.	Businesses Governments
Harvard Data Science Review and MIT Press, "A Unified Framework Of Five Principles of AI in Society", 2019	In this review, the authors lay out five principles to be used as the basis for future international, and national, laws, rules, and standards. 1. Beneficence: promoting well-being, preserving dignity, and sustaining the planet. 2. Non-maleficence: privacy, security, and 'capability caution. 3. Autonomy: the power to decide (to decide). 4. Justice: promoting prosperity, preserving solidarity, avoiding unfairness. 5. Explicability: enabling the other principles through intelligibility and accountability.	Policymakers Lawmakers
International Business Machines, Corp., "Everyday Ethics of Artificial Intelligence"	Approaching from an industry-led perspective, IBM lays out five everyday ethical codes of conduct for technology companies, lawmakers, and innovators: 1. Responsible innovation in the age of AI 2. The economic advantage of ethical design for businesses 3. Values by design in the algorithmic era 4. The Nature of Nudging 5. Data Protection and Data Safety.	AI Designers/ Developers
Research ICT Africa, 2019	Declares four critical needs for Africa's inclusion [AF1] in the development of AI: 1. a need to introduce safeguards to balance AI's risks and opportunities. 2. the need to protect personal and collective privacy rights in cross-border data flows. 3. the need to define African values and align AI with them. 4. the need for equitable, inclusive, and socially responsible AI development.	AI Developers Governments Legislators Policymakers

B. National laws, policies and regulations

Although international norms and standards provide a useful framework for designing and developing AI-driven technologies in a responsible and ethical manner, potential users of crime prediction technologies will need to understand the governing rules and regulations for their specific jurisdiction. Compliance with applicable legislation is crucial for demonstrating that the technology is being used for a lawful purpose, but also to safeguard against potential liability for harms that may ensue or for defending against claims of rights violations.

The three primary areas of law that prospective users should be aware of include: anti-discrimination laws, data protection and privacy laws, as well as access to information law. These laws should be read together, as each is likely to have an impact on the development and deployment of crime prediction technologies.

The table below provides an overview of the types of laws that would likely apply to the use of crime prediction technologies. It should be noted, however, that this is not an exhaustive list, and that attention should also be given to other governing pieces of legislation.

Table 3. Examples of national laws likely to apply to crime prediction technologies.

Country	Anti-Discrimination Law	Data Protection and Privacy Law	Access to Information Law
Argentina	Although Argentina does not have general anti-discrimination laws, The Argentine Anti-Discrimination Act protects employees from any acts of discrimination on various grounds.	The Personal Data Protection Act (PDPA) established in 2000 protects citizens' personal information and data, while also ensuring their right to any information stored in both private and public databases.	The Access to Public Information Law ensures citizens' right to search, use, request, analyze, etc. any information "kept in custody by liable persons."
Brazil	The Racial Crime Law, established in 1989, prohibits discrimination based on age, gender, race, color, marital status, or family status. It penalizes anyone who refuses a person's employment, or service based on discrimination. It also prevents inciting discrimination or prejudices through any form of media.	The General Data Protection Law (GDPL) which came into effect in August of 2020, unifies 40 existing laws and regulates the collection and processing of citizens' personal data. This law offers individuals protection against exploitative online data collection and allows them access to personal data. Article 6 of GDPL outlaws discrimination in the context of data processing. Together, these three key aspects of GDPL reinforce the ethical implementation of AI in public services.	The Right to Information Law (RIL) guarantees citizens the right to access information. It requires that public institutions provide the individual with their requested information no more than 20 days after the request.
Czech Republic	In 2009, the Czech Republic was adopted an Antidiscrimination Act (ADA), finally protecting citizens' right to equal treatment and non-discrimination.	The Czech Data Processing Act (DPA) is directed at processing personal data to predict crime. It ensures internal security and personal data protection for citizens.	The Freedom of Information Law, approved in 1999, guarantees citizens' access to information held by the government, states, local self-governing bodies, etc.
Denmark	In Denmark, there is no overall legislation that protects citizens from discrimination. However, the Danish Anti-Discrimination Act protects employees from discrimination based on race, gender, color, religion, belief, etc.	The Danish Data Protection Act (DPA) regulates the processing and movement of individuals' personal data.	The Access to Public Administration Files Act ensures citizens the ability to access administrative documents. When this act was put in place, the previous Public Access to Administrative Information Act of 1970 was repealed.

continuation

Country	Anti-Discrimination Law	Data Protection and Privacy Law	Access to Information Law
Egypt	Although there is no enabling legislation, Article 53 of the Egyptian constitution bans discrimination on any basis.	Egypt's recent Data Protection Law (DPL) explicitly protects citizens' personal information and represents the country's focus on the development of AI. In 2021, Egypt launched a National Artificial Intelligence Strategy with two main objectives: (1) to use AI to serve Egypt's developmental goals and benefit every Egyptian; (2) to take part in fostering regional and international cooperation and AI to reinforce "fairness and equality in all AI-related fora."	There is no legislation enabling Article 68 of the Egyptian constitution, which declares citizens' ownership over information, data, statistics, and official documents. It ensures the states' duties to ensure citizens' access to various sources of information.
Fiji	The Human Rights Commission Act of 1999 outlaws unfair discrimination from private employment sectors, the state, and, in some cases, individuals.	In Fiji, there is no specific legislation to protect citizens' personal data. However, Clause 24 of the constitution ensures the right to the confidentiality of personal information and the right to personal privacy.	The 2018 Information Act facilitates citizens' right to access information, and the right to alter harmful misleading information.
Germany	The German General Equal Treatment Act (AGG) outlaws discrimination in the workforce as well as daily affairs. Germany also established a Federal Anti-Discrimination Agency in 2006 to reinforce AGG.	The German Privacy Act (BDSG) outlines general rules and regulations for the processing of personal data in various contexts, and video surveillance. Germany also adopted the EU General Data Protection Regulations (GDPR) which further protects individuals' personal data.	In 2006, Germany established the Freedom of Information Act which grants citizens the ability to access official information held by the federal government.
Hungary	The Hungarian Equal Treatment and Equal Opportunity Act outlaws discrimination on the basis of religion, race, gender, beliefs, age, illness, or sexual orientation.	In 2018, Hungary amended its national laws to align them with the General Data Protection Regulations (GDPR) drafted by the European Union. GDPR aims to protect privacy, security, and data processing. The Hungarian National Authority for Data Protection and Freedom of Information (NAIH) supervises the implementation of GDPR within the country.	The General Data Protection Regulations (GDPR), enforced by the Hungarian National Authority for Data Protection and Freedom of Information (NAIH) ensure the Right to Access. This includes the right to access information, rectify information, and erase information.
India	Although there is no enabling legislation, Article 15 of the Indian Constitution prohibits discrimination based on race, religion, place of birth, caste, or sex.	In 2017, the Indian Supreme Court declared privacy as a fundamental right but does not have any data protection legislation. Currently, the parliament is reviewing the Indian Personal Data Protection Bill (PDPB) which would regulate use of personal data.	In 2005, India passed the Right to Information Act, guaranteeing citizens the right to access information from any public institutions or authorities.

continuation

Country	Anti-Discrimination Law	Data Protection and Privacy Law	Access to Information Law
Japan	Japan has no law preventing discrimination, but in 2016, the Diet (Parliament) passed the Anti-Discriminatory Speech Act (ADSA) which outlaws hate speech against foreigners living in Japan	In 2003, Japan passed the Act on the Protection of Personal Information (APPI). This act regulates the protection of individuals' personal information from organizations and businesses.	Japan's National Information Disclosure Law ensures individuals the right to request information from government entities and institutions.
Kenya	Article 27 in the Constitution of Kenya prohibits discrimination in any scenario but does not have enabling legislation. However, the Employment Act protects employees in Kenya from discrimination on essentially any basis. Similarly, the Persons with Disabilities Act prohibits discrimination against people with disabilities.	In 2019, Kenya adopted the Data Protection Act (DPA) which protects individuals' data from organizations, companies, and/ or data processors. In 2021, Kenya passed the Data Protection (Registration of Data Controllers and Data Processors) Regulations, and The Data Protection (Complaints Handling and Enforcement Procedures) Regulations.	The 2016 Access to Information Act reinforced citizens' right to acquire information from public bodies and private bodies "acting in a public nature."
Luxembourg	The Law of 28 November 2006 prohibits general discrimination within society.	In 2018 Luxembourg introduced the Data Protection Act which established the National Data Protection Commission (NDPC), and the Data Protection in Criminal Matters Act.	Luxembourg has no general freedom of information laws.
Mexico	In 2003, the Congress of the Mexico decreed the Federal Law to Prevent and Eliminate Discrimination to prohibits discrimination on most grounds including race, sexual orientation, health, etc.	In December 2011, Mexico's Data Protection Law (DPL) entered effect. This law protects citizens' personal information and ensures that individuals must authorize the release of any personal information.	Mexico's Freedom of Information Law grants individuals the right to demand information. The law aims to facilitate transparency within the public administration.
Norway	Norway's Anti- Discrimination Act prohibits discrimination based on race, ethnicity, skin color, religion, belief, or descent.	In 2018, Norway passed the Personal Data Act (PDA), which incorporated many aspects of the EU's GDPR law. These laws aim to regulate researchers' and research institutions' processing of individuals' personal data.	The Freedom of Information Act (FIA) in Norway declares that any individual can "apply to an administrative agency for access to case documents, journals, and similar registers…"
Philippines	The Anti-Discrimination Act of 2017 outlawed discrimination on the basis of religion, sexual orientation, gender, etc.	The Data Privacy Act protects all forms of individual data — private, personal, and sensitive, and regulates data processing.	Freedom of Information was put into effect as a law by an executive order in July 2016.

continuation

Country	Anti-Discrimination Law	Data Protection and Privacy Law	Access to Information Law
Rwanda	Article 16 of the Rwandan constitution declares that "all Rwandans are born and remain equal in rights and freedoms." This article implies that discrimination on any basis is against the law.	In October 2021, the Protection of Personal Data and Privacy Law (PPDP) was passed which aligns with international data protection standards and provides a basis for its goals of a "technology-enabled and data-driven economy." Rwanda is already a global leader in the development of data and AI policy, as the main Centre for the 4th IR	In 2013, the Rwandan government passed the Rwanda Access to Information Law (AIL). This law grants all public bodies and some private bodies access to information, with a few exemptions.
Sri Lanka	Article 12 of the Sri Lankan Constitution ensures the right to equality before the law and equal protection of all persons. It also outlaws discrimination on the basis of race, religion, language, caste, sex, political opinion, and birthplace. Article 12 protects citizens from potential prejudices that arise in AI systems.	The Personal Data Protection Act (PDPA), passed in 2022, regulates the processing of personal information and strengthens citizens' protection of personal data. For example, persons can withdraw consent for data processing, or request data erasure, under certain circumstances.	In 2016, Sri Lanka's Right to Information Act (RTI) went into effect. This act ensures citizens the right to request and, most likely, receive information within 14 days of their request.
South Africa	The Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) prohibits unfair discrimination and promotes equality. This law protects from biases that may arise due to racial prejudices or other unfair forms of discrimination.	The Protection of Personal Information Act (POPIA) protects citizens' right to privacy and limits use of their online personal information. It restricts the government and private entities from processing personal information for use that is beyond a limited purpose	The Promotion of Access to Information Act (PAIA) offers citizens access to any information, in private or public sectors, that reinforces the exercising of their rights. In the context of AI, PAIA grants citizens access to online information, including SARS records.
Turkey	In 2016, Turkey passed the Law on Human Rights and Equality Institution of Turkey which aims to promote human rights and equality and eliminate discrimination	The Law on the Protection of Personal Data both established the Turkish Data Protection Authority and enforced regulations on personal data privacy and data processing.	The Turkish Law on the Right to Information guarantees everyone the ability to request information from organizations or institutions.
Uganda	Although there is no specific anti-discrimination legislation in Uganda, Article 21 of the Constitution of Uganda prohibits discrimination on any ground and ensures equality before the law.	The Data Protection and Privacy Law, established in 2019, regulates the processing and extracting of personal data, therefore protecting citizens' personal information.	In 2005, Uganda enacted The Access to Information Act (ATIA) which aimed to facilitate citizens' engagement with public decisions and promote transparency within the government.

National Policy Frameworks for Ethical Use of AI

Below is a table of countries that have developed policies for ethical use of AI. The chart provides a framework for the overall approach and underlying values governing AI use. In the absence of a national policy on AI, however, countries can still refer to the body of international standards and legal obligations to conceptualize how AI-driven technologies will be developed and implemented.

Table 4. Countries that have National Policy Frameworks for Ethical Use of AI

Country	Description of Framework
Canada	The Pan-Canadian Artificial Intelligence Strategy, a national policy framework for the ethical implementation of Artificial Intelligence, is based on three main pillars: (1) commercialization; (2) standards; and (3) talent and research. In June 2022, Canada launched the second phase of their Pan-Canadian Artificial Intelligence strategy.
China	In 2017, China released their New Generation of Artificial Intelligence Plan, which recognizes the country's goals as an active participant in the 4th IR, and outlines strategies and tasks to reach their national Artificial Intelligence goals.
Egypt	Through its National Artificial Intelligence Strategy, Egypt hopes to use AI to achieve its sustainable development objectives, establish their country as an active participant in the development of AI, and promote collaboration within the African and Arab regions.
France	France's National Artificial Intelligence Research Strategy is largely based in a report produced by Cedric Villani called AI for Humanity. In 2018, President Emmanuel Macron announced the focus of France's national AI strategy — to make France a global leader in Artificial Intelligence research, development, application, etc.
Germany	Germany is making strides with AI through their national strategy, AI Made In Germany. and recently announced a large amount of funding for five national AI competence centers.
India	India's National Strategy for Artificial Intelligence (NSAI), announced by NITI Aayog promotes the ethical implementation of AI, and highlights the potential social and economic benefits of this technology.
Rwanda	In collaboration with a variety of stakeholders, including the Ministry of ICT and Innovation, GIZ/FAIR Forward, and The Future Society, Rwanda is in the process of developing a comprehensive national AI policy that will promote their national goal to "become the premier technology innovation hub and AI leader across Africa."
Singapore	In May 2017, Singapore released their National Artificial Intelligence Strategy, a cohesive outline of the necessary steps to achieve their national goal to transform their economy.
United Arab Emirates	The UAE Strategy for Artificial Intelligence aims to improve government performance, enhance their market and economy, generate an efficient, problem-solving digital system, reach the goals outlined by the UAE Centennial 2071, and ultimately make UAE the number one in the field of AI investments.
United Kingdom	In September 2021, the UK released a new National AI Strategy, one that promotes the nations' current strengths and outlines the future benefits of AI.
United States	The United States' "National Artificial Intelligence Initiative", offers detailed information about the country's approach to the future of Artificial Intelligence. Also, in 2020, the U.S. Government passed the National AI Initiative Act of 2020, which promotes the research and development of AI across the federal government.

C. Emerging best principles, practices, and processes

As AI-driven technologies become more ubiquitous in everyday life, the body of emerging best principles, practices and processes will continue to evolve. That said, it is useful to unpack each of these principles to see in more detail how they are reflected in practice and to learn from the experiences of how they are interpreted and applied in a variety of contexts. The accompanying set of actions, through processes and practices, create a framework that allows for increased levels of transparency and accountability. This is essential not only for ensuring responsible and ethical use of AI, but also for building public trust and legitimacy in the use of crime prediction technologies.

What principles currently support responsible and ethical use of AI?

The normative framework for ethical AI includes a set of six overlapping principles, including: explicability, accountability, fairness, oversight, privacy, and human-centered.

Explicability	• people must be able to understand what it does, how it works, and the risks involved
Accountability	• ensures the proper functioning of systems and takes responsibility when things go wrong
Fairness	• does not perpetuate bias or impose unfair discriminatory outcomes against particular categories of persons
Human Autonomy	• the power to decide whether to act ultimately rests with humans
Privacy	• privacy, security and 'capability caution'
People-Centred	• promoting well-being, preserving dignity, and sustaining the planet

The following section offers a more detailed explanation of what these principles mean for AI-driven technologies and provides concrete examples of what actions could be taken to adhere to these principles in the deployment of prediction technologies.

1. **Explicability:** Practitioners should strive to make readily understandable what AI-driven technologies do, how they work, and the risks involved in their deployment. The principle of explicability requires developers and implementers of AI-driven tools to explain the logic of algorithmic decisions and the data used to generate predictions to potential users and beneficiaries using clear language and non-technical terms. This principle aligns with broader efforts to increase algorithmic transparency by turning 'black-box into glass-box models' to build public trust.

- **Example:** In the context of predictive policing, adherence to the principle of explicability could require extensive training with law enforcement officers prior to deployment. Explanation about how prediction technologies work can influence officer trust in the tool's accuracy and influence whether they act in accordance with what the prediction suggested. Such training can also capacitate officers to explain how the technology works to communities where it is used and other stakeholders.

2. **Accountability:** The principle of accountability requires someone (entity, user, developer, etc.) to ensure the proper functioning of the AI system and take responsibility when things go wrong or if harms result from its use. It also requires institutions to present visible ((viable?)) opportunities for users and stakeholders to learn more about AI-driven technologies as well as compliance with regulatory frameworks and international standards. There should also be mechanisms for filing complaints that are adequately supported by institutional protocols that require due diligence to ensure issues are resolved fairly and thoroughly.

   - **Example:** For the principle of accountability to take root in the deployment of crime prediction technologies, robust performance management systems should be in place to ensure their proper use and deployment by law enforcement officers. A core focus of the performance management system should be implementation of quality assurance measures for data entry to ensure crime prediction tools are relying on information that is timely, accurate, and complete.

3. **Fairness:** The principle of fairness demands that developers and users of AI-driven technologies to avoid perpetuating the risk of bias or use of discriminatory practices against persons based on their race, gender, class, national origin, or other protected grounds. Because fairness is one of the field's biggest concerns, AI implementers must work to mitigate the risk of discriminatory outcomes and unfair practices that may inflict harm on individuals, communities, and society. Institutional mitigation measures should be made publicly available and subject to independent review on an ongoing basis to identify emerging risks and new forms of potential harm.

   - **Example:** Before implementers make the decision to deploy crime prediction technologies, they must conduct an audit of existing datasets to ensure records are accurate, up-to-date, and representative of the local population. Adherence to the principle of fairness could involve practitioners conducting an independent assessment prior to deployment to identify and isolate potential sources of bias. Based on the findings, potential users should implement a series of measures to mitigate the specific risks, such as cleaning the training data or recalibrating the algorithm, improve the tool's accuracy and reduce the risk of it perpetuating existing forms of discrimination or introducing new harms.

4. **Human Autonomy:** While AI-driven technologies offer important advantages when it comes to data synthesis and analysis, the principle of human autonomy places the ultimate power to decide which action to take with humans, not with machines. This means that the autonomy of humans should be promoted, while the autonomy of machines should be restricted and ultimately subjected to human oversight.

- **Example:** The principle of human autonomy should find strong expression in the standard operating procedures for developing and deploying crime prediction tools. For example, under the principle of human autonomy an operator might be required to explain why they chose to ignore or deviate from a prediction or suggested course of action generated by the tool. Although users or operators should not be reprimanded for deviating from a decision, they should still be required to explain their reasons for doing so. Such information could be used to improve the tool by using human inputs, and to counterbalance prediction technologies with officer knowledge and experience.

5. **Privacy:** The principle of privacy requires measures to be put in place to protect the personal information of intended users, beneficiaries, and other stakeholders of AI-driven technologies. This includes compliance with privacy and data governance rules, including retention policies and information sharing protocols, and restricting the use of personal data to a limited and lawful purpose. The principle of privacy also provides additional protections against unfair discrimination given that information about a person's attributes (race, gender, national origin, etc.) may be used as a proxy for automated decision-making, which potentially threaten the right to privacy by how the information has been used.

   - **Example:** One way to adhere to the principle of privacy is to avoid using data generated by facial and voice recognition technologies in crime prediction systems. In addition to escalating risks of false positives, there is growing concern that image recognition technologies violate the right to locational privacy, which includes the right of people to move about freely without having their movements tracked in the absence of reasonable suspicion. This is especially relevant for countries to provide constitutional protections not only for the right to privacy, but also for the right to freedom of movement, such as South Africa.

6. **People-Centred:** The principle that AI be people-centred - in order words, used to promote the well-being of people and the preservation of the planet for future generations. Accordingly, this principle aims to ensure that AI technologies benefit and empower as many people as possible, rather than the commercial interests of big tech or the political agenda of the government. It also extends to the environment and calls for AI-driven technologies to protect the basic preconditions of life on the planet.

   - **Example:** In the context of predictive policing, this could mean using tools that are designed to not only benefit law enforcement operations, but also to enhance the personal safety of ordinary people. For example, CrimeRadar in Brazil developed a public-facing interactive heat map that allows ordinary people to understand threats to safety in their area and plan their routines and commutes accordingly. In this regard, the tool is designed to benefit a larger demographic cohort and expand transparency of information relating to public safety and security.

What best practices currently contribute towards responsible and ethical use of AI?

In addition to ethical principles, an area receiving increased attention is emerging best practices that contribute towards responsible and ethical AI. Below are a few examples of ways to ensure that AI-driven technologies do not impose unintended harms, even when they are likely to evolve. Therefore, it is important to stay informed of ongoing developments, especially ones that would be relevant to crime prediction technologies in your local context.

1. **Independent Fairness Testing:** Independent Fairness Testing (IFT) testing is used to detect forms of algorithmic bias that may create or reinforce disadvantages or discriminatory practices against disadvantaged and underrepresented groups. Considered by experts to be one of the essential components of AI governance, IFT uses different metrics to measure the fairness of the model.

   Below is a sampling of the types of metrics IFT can measure:
   - Statistical Parity Difference: measures the difference in the probability of favorable outcomes between privileged and unprivileged groups
   - Equal Opportunity Difference: measures the difference in positive rates (or scoring) between unprivileged and privileged groups
   - Disparate Impact: measures the ratio in the probability of favorable outcomes between the unprivileged and privileged groups

2. **Third Party Audits:** Third-party auditing should take place at regular intervals following the deployment of AI-driven technologies. Users of AI should invite independent and experienced third parties to understand and review their algorithmic decision systems, which requires disclosing sufficient information to allow accurate testing, monitoring and feedback. Ultimately, the goal is to inform end-users and stakeholders that an algorithmic decision system was audited by a trusted third-party and that it remains open to independent auditing in the future.

   In 2019, the International Systems Audit and Control Association (ISACA) released the Control Objectives for Information and Related Technologies (COIRT) as a thorough and ethical auditing framework for technology that can be applied to the auditing of AI.

3. **Social Impact Assessments:** Social Impact Assessments (SIAs) measure the impact of AI-driven technologies on the social elements of life. SIAs have been traditionally conducted on affected groups of persons against six categories of metrics, including: (1) employment (including labor market standards and rights); (2) income; (3) access to services (including education, social services etc.); (4) respect for fundamental rights (including equality); (5) public health; and (6) safety. It is important to note that this is not an exhaustive list and should be context specific.

   In the context of AI, impact assessments should also measure the technical, legal, and ethical implications of AI, using such tools as the Artificial Intelligence Impact Assessment (AIIA), recently developed by the European Community of Best Practice (ECP).

What best processes currently contribute towards responsible and ethical use of AI?

In addition to a growing set of emerging best principles and practices, best processes are also attracting attention for their potential to foster responsible and ethical use of AI. A few of the more prominent ones are outlined below.

1. **Fairness-by-Design:** This process involves examining different parts of the machine learning process from different vantage points, using an interdisciplinary team of experts with different theoretical lenses. For example, companies can pair data scientists with a social scientist; integrate traditional machine learning metrics with fairness measures; balance representativeness with critical mass constraints when engaging in sampling for training data; and keep de-biasing in mind when building models.

2. **Privacy-by-Design:** This is a methodology to ensure that privacy principles are embedded into the products from their conception through the development process. For example, privacy-by-design entails: (1) using only the data that is needed to achieve a particular purpose; (2) letting people know about the personal data that is stored and giving them the ability to correct or delete information; (3) using anonymized data when possible so it is not possible to connect someone with the data that was involved; and (4) including restrictions at the outset about how data will be used or transferred.

3. **Ethics-by-Design:** This methodology provides guidance for embedding ethical principles in the design, development, and deployment of AI based solutions. Ethics-by-design guidelines may entail specific tasks that should be completed at different stages in the development process of an AI solution. For example, when defining the problem that an AI-driven solution is meant to address, developers must assess whether the objectives align with the six ethical principles. Drawing on expertise from diverse disciplines may be useful in identifying other ethical issues that could be implicated during deployment of the technology.

PART 3: Assessing institutional readiness for AI

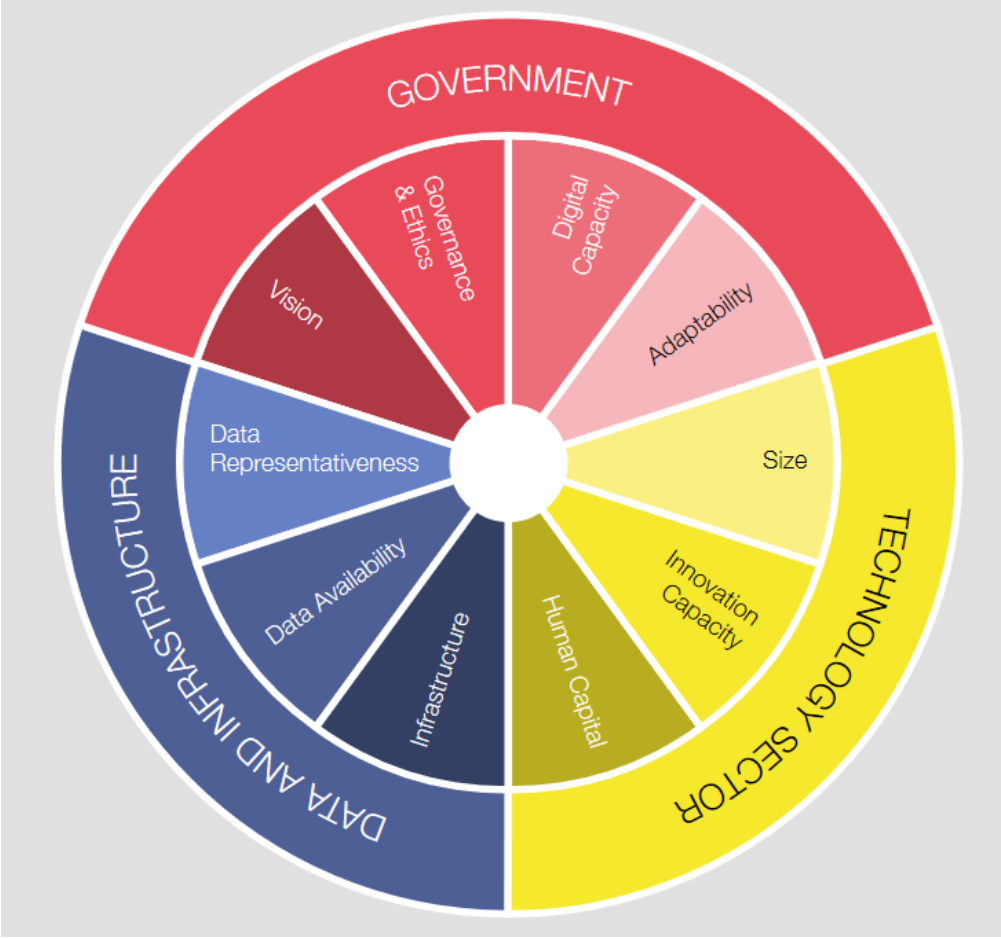
While AI-driven technologies have the power to address some of society's most vexing challenges, it is important to remember that technologies are only tools - not solutions. Their efficacy depends upon the 'readiness' of institutional environments to adopt and use them. Understanding 'readiness' is crucial for leveraging the technology for responsible and ethical implementation.

'Readiness' means more than political will and investment in AI. Readiness considers the strength of existing infrastructure and capacities of institutions to design, develop, deploy, and oversee their use of AI-driven technologies. Such capacities include things like digital literacy and skills of users, data infrastructure and connectivity, quality assurance and performance management systems, as well as cybersecurity protocols and procedural safeguards. Assessing institutional readiness is critical in all sectors, but when institutions are looking to deploy high-risk technologies, such as crime prediction tools, building capacity to improve readiness becomes urgent and critical.

Accordingly, the following section offers guidance for assessing the institutional readiness of implementing institutions for responsible and ethical AI by: (1) presenting Oxford Insight's AI Readiness Index; (2) providing a framework for conducting a preliminary assessment of core capacities; and (3) exploring the implications of procurement processes.

A. Building a responsible and ethical AI ecosystem

In 2020, the International Research Development Centre (IRDC) in Canada released Oxford Insight's Government Artificial Intelligence (AI) Readiness Index. The Index also provides a framework for assessing the readiness of a country's AI 'ecosystem' by identifying the foundational elements of an AI ecosystem. They include government, public policy, academia, private sector, data, infrastructure and skills. Because not all elements will be present in each country, only the foundational ones have been provided below.



Data & Infrastructure Requirements:

- Data Representativeness
- Data Availability
- Infrastructure

Government Requirements:

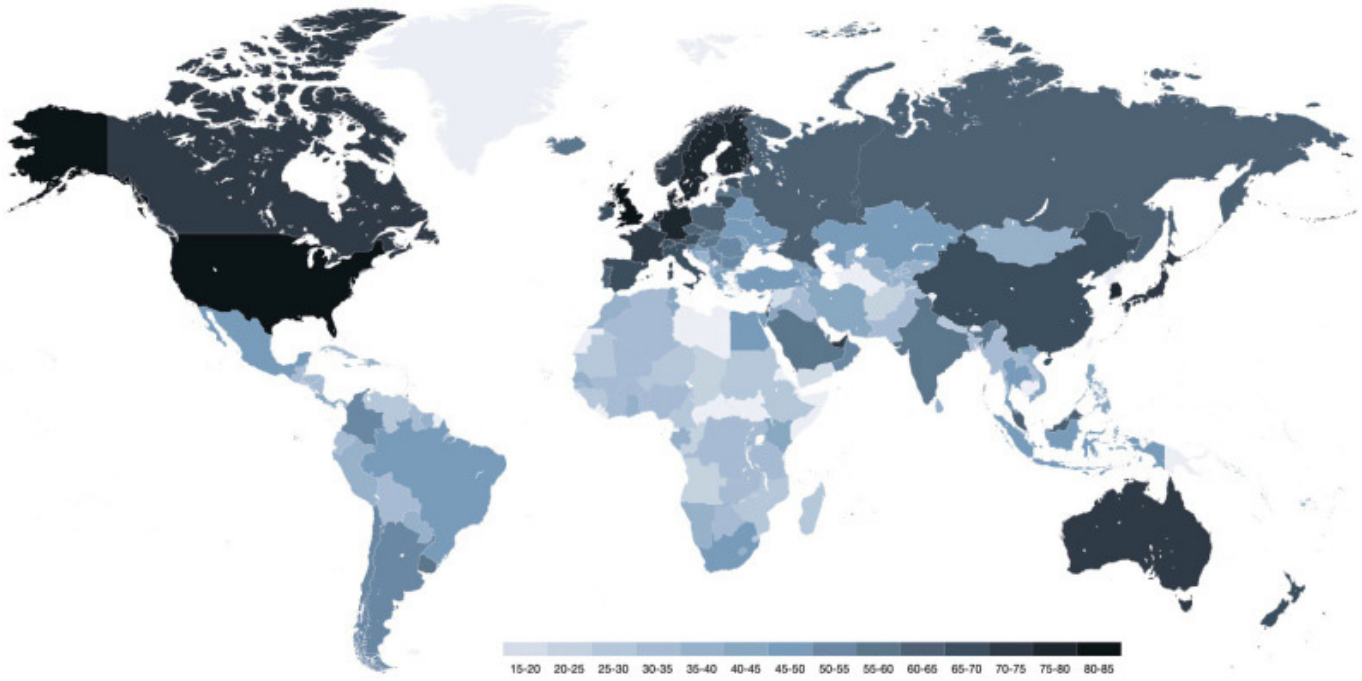
- Vision
- Governance & Ethics
- Digital Capacity
- Adaptability

Technology Sector Requirements:

- Size
- Innovation & Capacity
- Human Capital

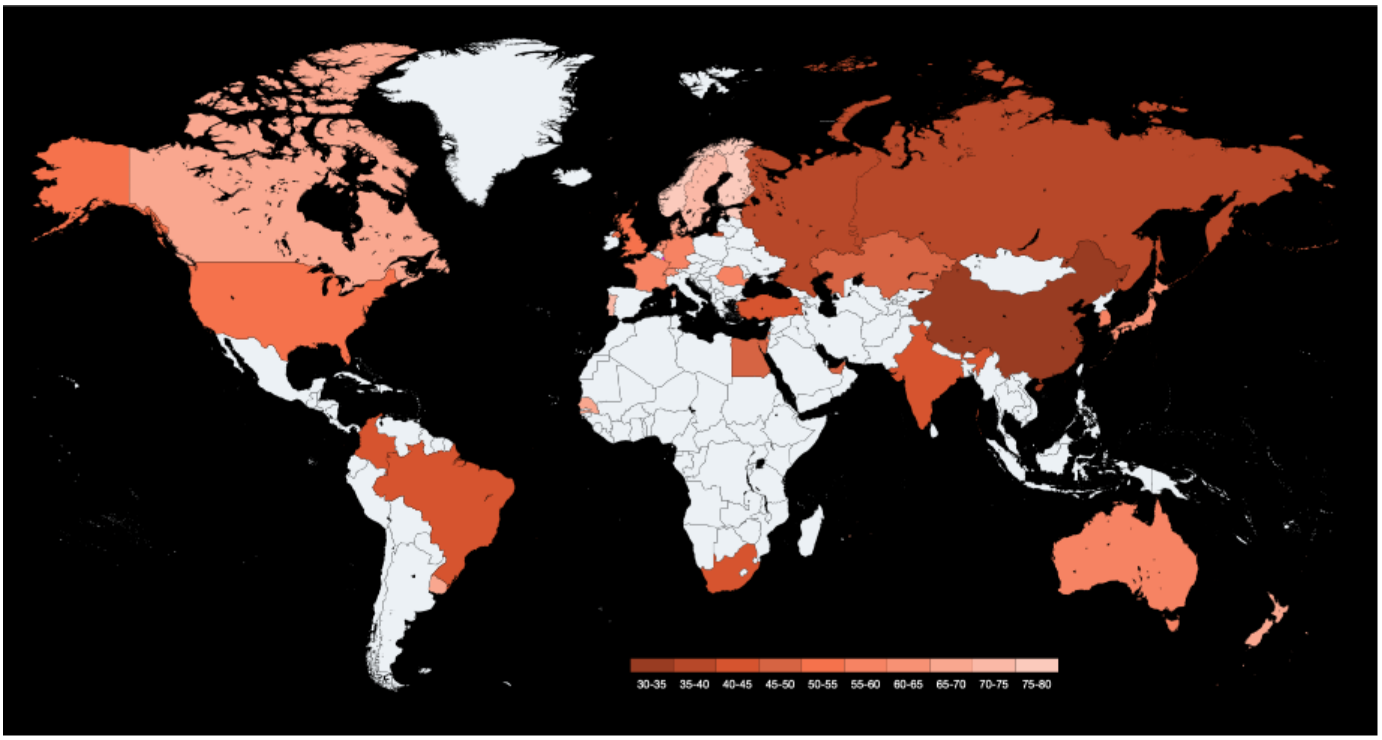
Source: The Government Artificial Intelligence (AI) Readiness Index (2020) from Oxford Insights and the International Research Development Centre https://www.oxfordinsights.com/government-ai-readiness-index-2020

Map of AI Readiness Index 2020



It is worth noting that findings from Oxford's AI Readiness Index in 2020 reveal that countries in Europe, North America and Asia are more 'ready' for AI than countries in Africa, Latin America, Middle East, and Southeast Asia. However, when those findings were compared with the Responsible Use Index, some of the world's wealthiest countries scored significantly lower on the metrics measuring responsible use. These insights are important for gauging what AI-driven technologies should be used to address and who they are designed to benefit.

Map of Responsible AI 2020

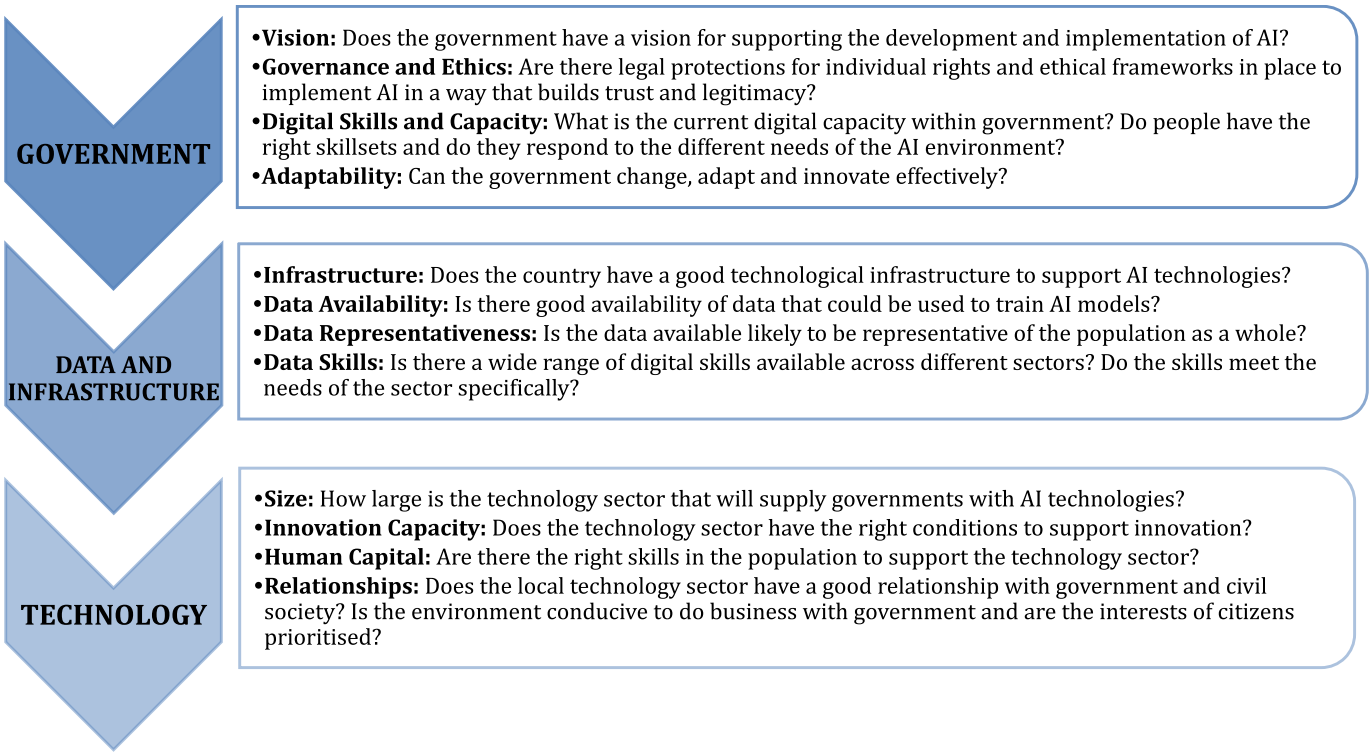


Source: Government AI Readiness Index 2020 https://www.oxfordinsights.com/government-ai-readiness-index-2020

B. Conducting a preliminary assessment of institutional capacities

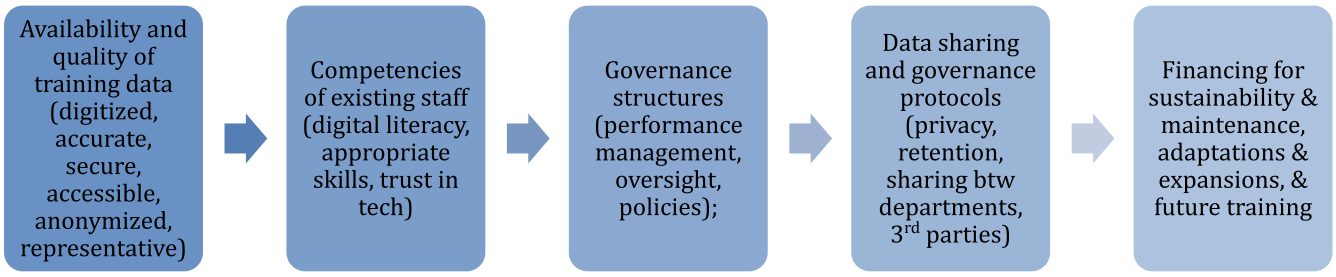
Assessing the foundational elements of an AI ecosystem, even at a preliminary level, is an essential step when deciding whether to use crime prediction technologies. The results can provide critical insights into gaps that may need to be addressed to create an environment conducive to responsible and ethical use of crime prediction tools and to mitigate the risk of potential harms arising at an individual, communal, and societal level.

While there are several tools and frameworks available for assessing AI readiness, the graphic should help get you started in understanding some of the preliminary questions to ask at the outset. There may not be enough time or data to do a comprehensive assessment of the foundational elements listed above, but these questions should still inform the process going forward.



After conducting a preliminary assessment on the foundational requirements of the AI ecosystem, the next step is to conduct a comprehensive assessment of the institutional capacities of the implementing department. This may include multiple departments across different sectors and spheres of governance, so it will be essential to assess each one separately. This is especially important in the rollout of crime prediction technologies since it would be ideal to rely on various sources of data inputs to mitigate certain forms of algorithmic bias, specifically ones pertaining to lack of representativity in the datasets and potentially harmful feedback loops that reinforce existing forms of bias and discriminatory practices.

The areas outlined below are not exhaustive but should be treated as a matter of priority during the preliminary assessment of each implementing department.



Results from the preliminary assessment should inform the strategy for building capacities to improve institutional readiness for AI. If some competencies are stronger than others, begin engaging with role-players at the outset to gauge their support for, and level of trust in, AI-driven technologies. By bringing them into the process at an early stage, it will be possible to build a multidisciplinary team and leverage additional support, as there will likely be other issues that need to be considered. For areas where there are gaps or weaknesses discovered, rectifying them should be treated as a matter of priority and considered a prerequisite for the procurement of crime prediction tools.

C. Understanding the implications of procurement processes

Public procurement processes can become key drivers for the adoption of responsible AI so long as measures are taken to uphold ethical principles. Emerging technologies will continue to develop and deliver new opportunities to advance public services and efficiency in government. That said, it is neither fair nor reasonable to expect procurement officers to understand the evolving complexities of AI and the risks they invariably pose. Without understanding how to ensure explicability, accountability, and privacy in the procurement of emerging technologies, governments may create new risks in the procurement of AI-driven technologies and introduce new forms of harm at an individual, communal, and societal level.

Developing procurement guidelines and ethical frameworks can be a useful way to mitigate some of the risks of AI, especially with high-risk technologies like crime prediction tools, to ensure AI is used in a responsible and ethical manner. AI Procurement in a Box is a tool that was developed in 2018 by the World Economic Forum to guide governments in adapting procurement frameworks to focus on innovation, efficiency, and ethics.

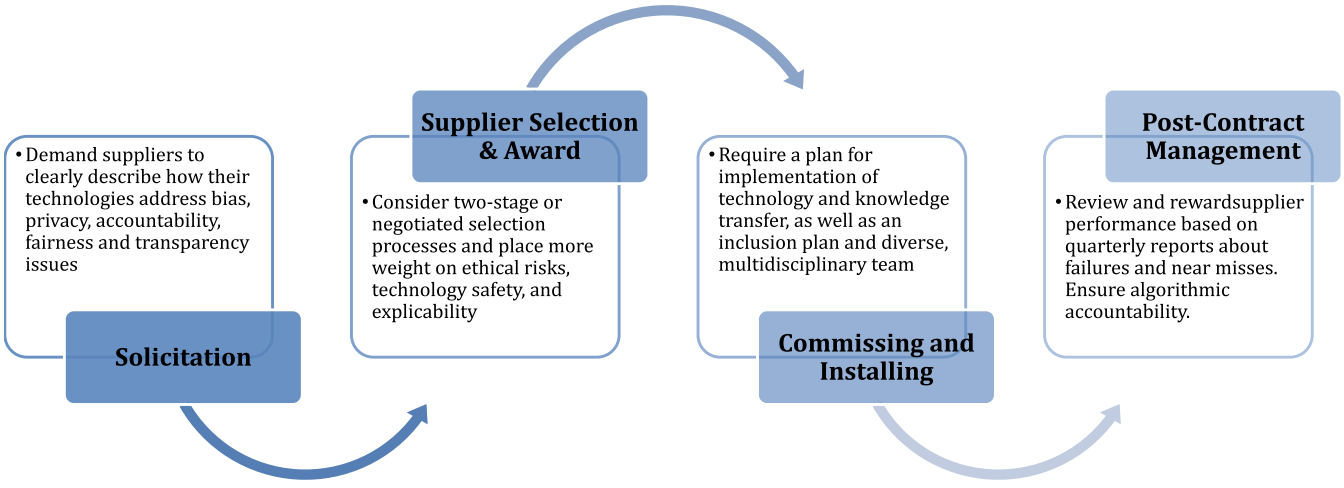
The tool includes a set of ten principle-based guidelines for procurement which addresses the central issues that need to be considered when procuring AI-driven technologies. Referring to this tool for guidance in developing some of the key questions to include in procurement strategies as well as a detailed explanation of the principles is highly recommended.

Ten Principles for AI Procurement

Use procurement process to articulate a problem, rather than prescribe a specific AI-solution, to allow room for innovation	Define the specific public benefit that will be achieved using AI while also acknowledging the risks.	Align procurement processes with existing governmental strategies and policies	Incorporate legislation relating to privacy rights and non-discrimination as well as codes of practice in the RFP.
Ensure proper data governance mechanisms are in place prior to procurement and safeguard the sharing of government data to service providers	Highlight known limitations in the quality of training data and require tenderers to describe their strategies to address these shortcomings	Work with a diverse, multidisciplinary team during procurement and require successful bidder to do so too	Focus on mechanisms of algorithmic accountability and explicability by successful bidder
	Implement a process for the continued engagement with service provider for knowledge transfer and long-term risk assessment	Create the conditions for a competitive market amongst service providers and require open licensing terms to avoid vendor lock-in	

It may take time to integrate the above principles into procurement practices, but efforts should be made to embed principles for responsible and ethical use of AI throughout the procurement process.

Below is an example of how to integrate responsible and ethical criteria into each stage of the procurement process, which extends through all stages of implementation, including post-contract management. These expectations should be articulated upfront in the Request for Proposals (RFP).



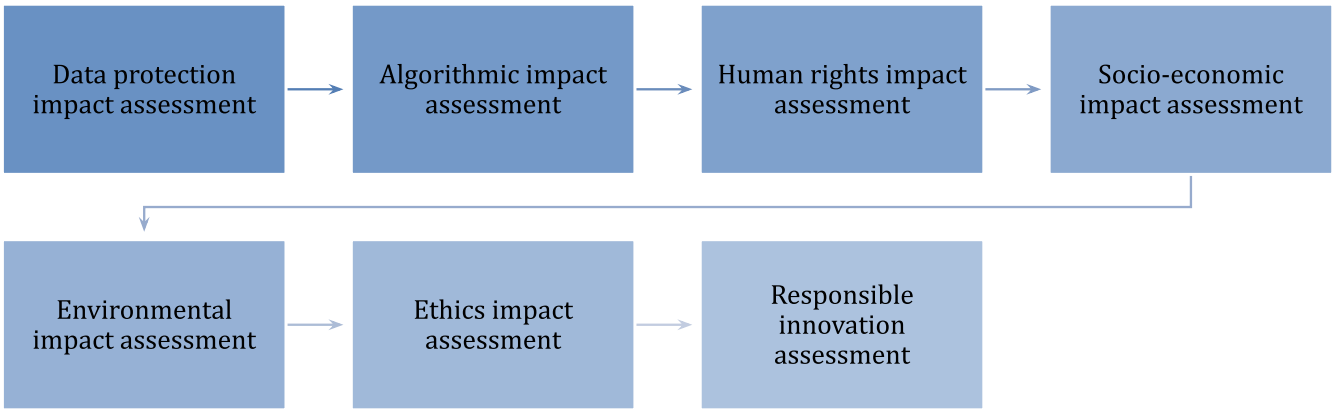
Source: Nagitta, Pross Oluka; Mugurusi, Godfrey; Obicci, Peter Adoko; Awuor, Emmanuel (2021). Human-centered artificial intelligence for the public sector: The gatekeeping role of the public procurement professional. Forthcoming in Procedia Computer Science

PART 4: Social impact assessments

By now, there is almost widespread agreement amongst most democratic countries that a risk-based approach is both necessary and appropriate in the development and deployment of AI-driven technologies. Far less attention has been given to the problem of how to identify, measure, monitor, and mitigate the risks these technologies pose. Nor is there sufficient clarity on defining the scope of risk, distinguishing between them, and knowing when and under what conditions each type of risk demands assessment. This is concerning given the rates at which technologies are being developed and positioned as 'cures' for some of the world's most complex social ills. Having access to a set of tools that can be used to monitor their impact – in terms of both positive and negative outcomes – can be a powerful way to mitigate risks while also leveraging the power and potential of these tools.

Impact assessment, simply defined, is a process for anticipating the future consequences of a proposed plan or set of actions. The purpose of an impact assessment is two-fold. First, it allows for predicting the anticipated impact of both benefits and harms; and second, for assessing the actual impact in terms of both benefits and harms. Given the ubiquitous nature of AI, some experts have suggested that impact assessments for AI should build on existing frameworks that have been developed for other fields, including global development projects.

Examples of impact assessments that may be relevant to AI include the following:
In the context of crime prediction, much attention is focused on the technology's social impact,

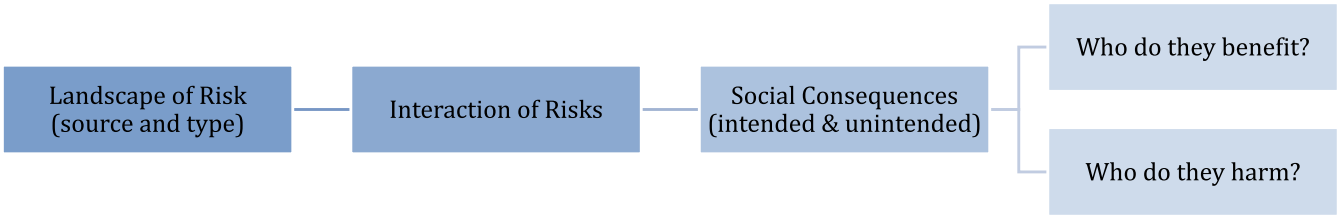


specifically the potential types of harm these tools pose at an individual, communal, and societal level. Accordingly, this next section will focus specifically on Social Impact Assessments (SIAs) as a framework for identifying, measuring, monitoring, and mitigating the risk of harm. It will provide a description of what they measure as well as the recommended principles underlying their design and implementation. It is essential to note that there are multiple ways to do this and that it will be essential to adapt SIAs to address the specific set of risks that arise in the local context.

A. Overview of social impact assessments

SIAs are a type of impact assessment that measures the social consequences of a planned intervention or action, including the deployment of crime prediction tools or other AI-driven technologies. In this regard, SIAs are a systematic process of identifying, analysing, monitoring, and managing the intended and unintended consequences, as well as both the positive and negative social changes, arising from the use of AI.

Accordingly, the primary purpose of an SIA is to develop a better understanding of: (1) the landscape of risk in a given context; (2) how those risks interact with one another; (3) the social consequences they produce (both intended and unintended); and (4) and who is most likely to benefit and who is most likely to be harmed.



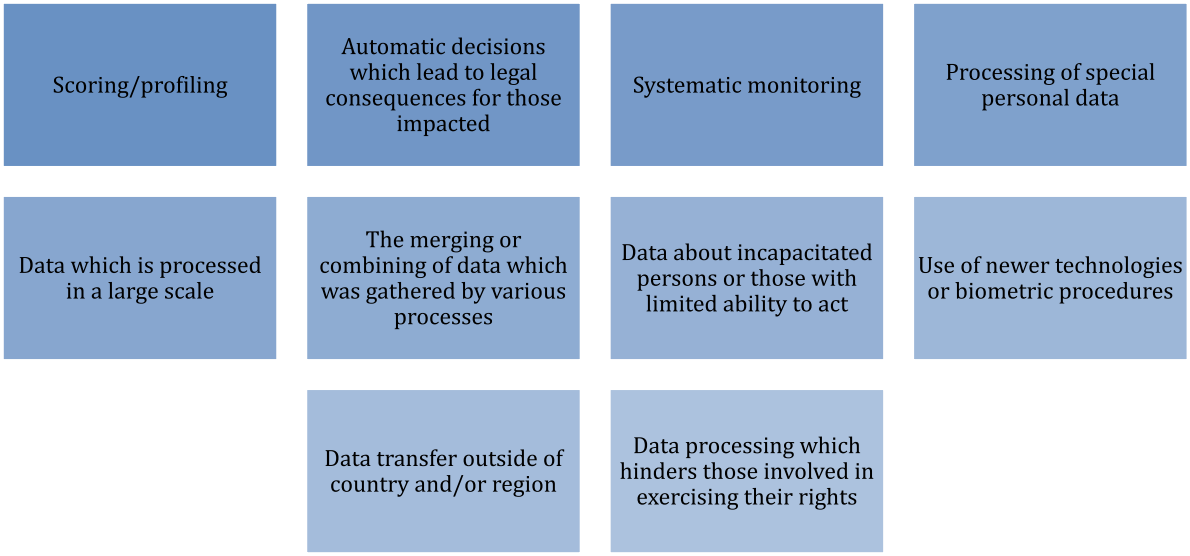
By developing a better understanding of the landscape of risk and how these risks interact to produce social consequences - both positive and negative - it is then possible to develop enhanced targeted risk mitigation strategies. In this way, practitioners could better assess the source and type of risk present, and so optimize use of AI-driven technologies to expand the number of people they are designed to benefit.

That said, it is essential to remember that risks will be context-specific – even at a departmental level – and the interaction of those risks will depend on the environment in which crime prediction tools are being used (institutional) as well as deployed (community). Even though some risks will cut across different departments and communities, they will still be unique to the specific location where they are used and deployed, a strong indicator that SIAs should be as localized as possible.

When is it necessary to conduct a Social Impact Assessment prior to using AI?

Like some other types of impact assessments, SIAs aim to predict and assess the consequences of a proposed action or initiative before a decision to implement is made. This is critical for deploying crime-prediction tools, classified as a 'high risk' technology with significant social consequences.

SIAs that are conducted prior to use are called ex-ante impact assessments. Article 35 of the European Union's General Data Protection Regulation (GDPR) calls for ex-ante assessments (prior to deployment) in instances where high-risk technologies are involved. These include, but are not limited to, ones that involve scoring and/or profiling persons; automatic decisions which lead to legal consequences for those impacted; processing of specialized personal data; and the use of newer technologies involving biometric procedures, amongst others.



Source: Article 35 of the European Union's General Data Protection Regulation: https://gdprhub.eu/Article_35_GDPR#:~:text=Article%2035%20requires%20the%20controller,and%20freedoms%20of%20natural%20persons

Ex-ante impact assessments should follow a similar process to an SIA and the results generated from the assessment should inform the decision whether to deploy crime prediction tools. The decision to deploy these technologies may be affected if the anticipated risks arising in a particular institutional environment are assessed to be too too high, the quality of training data is not representative enough, or if the communities designated for crime prediction tools require more information, greater levels of explicability or more opportunities to engage with stakeholders.

For example, the results of an SIA could be used to generate a risk assessment framework that analysis that identifies: the type of risk, potential sources, existing controls/procedural safeguards, probability of occurrence, consequence of occurrence, and then classifies them according to severity.

Table 6. Example of Risk Assessment Framework for Crime Prediction Technologies

Potential Risk	Source	Existing Controls	Probability	Consequence	Classification
Release of confidential crime intelligence	Department's server is hacked by an 3rd party	Cyber-security software; post-breach investigation	Medium	Inaccurate predictions; destruction of criminal records; cybercrimes	Severe
Outdated training datasets	Training data that meets standards is only available from 3 years ago	Ongoing process to clean data through digitizing and quality-control	High	Inaccurate predictions; over-policing certain areas; under-policing other areas	Moderate
Poor levels of digital literacy in officers	Senior ranking officials do not have digital skills	In-service training to modernize and professionalize police	Medium	Officer resistance to technology; inaccurate data inputs; misuse	Moderate

What are the components of a Social Impact Assessment?

The step-by-step nature of a SIA requires that the process be broken down into separate phases, each providing its own contribution towards assessing the social consequences of AI. The format and structure of the SIA should adapt to the needs and capacities of the context but establishing a step-by-step process that can be conducted on an ongoing and continuous basis is critical for institutionalizing SIAs as part of the process of deciding whether or not to use crime prediction technologies.

The process below emanates from the six steps for a human rights impact assessment based on the UN's Guiding Principles. It sets out a comprehensive, yet manageable approach for conducting an SIA and gathering the type of information necessary to mitigate risks prior to deployment as well as throughout the implementation process.

It is possible to modify as necessary but following the steps outlined below on an ongoing and continuous basis is useful.

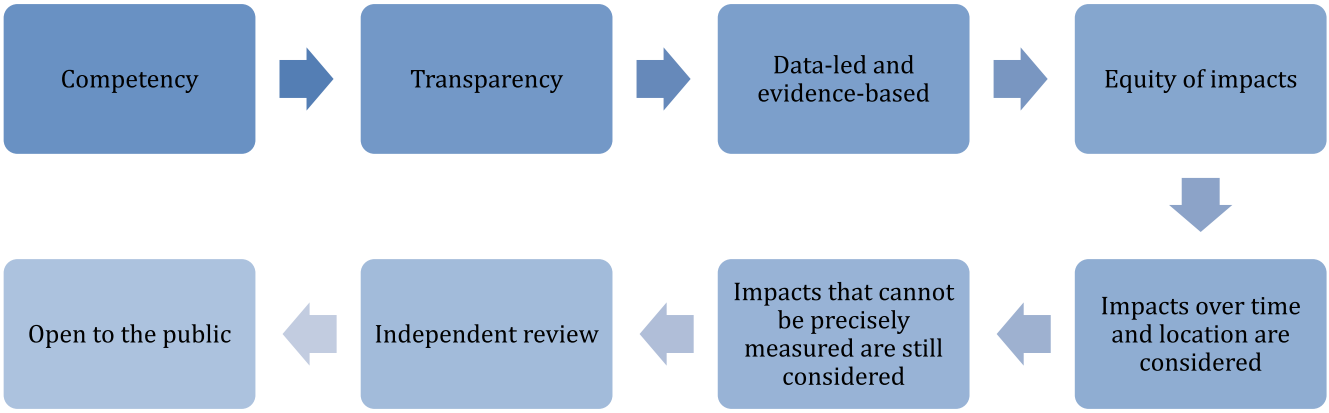
Screening	• Identification of risks and harms that might occur and who will be affected according to protected grounds (age, gender, race, nationality, location, etc.).
Scoping	• Detemine questions to be addressed, methodology and indicators to be used in all areas identified during screening.
Evidence Gathering	• Qualitative methods (interviews, consultations with affected groups), quantitative methods (modelling, regression analysis).
Analysis	• Assessment of evidence against structural factors and other considerations (e.g. laws that may constrain or enable certain behavior).
Conclusions & Recommendations	• Presentation to an entity that can hold the institution in question to account.
Evaluation Mechanisms	• Follow-ups with affected groups and implementing institutions about how the recommendations will be implemented.

Source: Rachel Adams, et al. (2021). Human Rights and the Fourth Industrial Revolution in South Africa, Chapter 7. Cape Town, HSRC Press.

Principles for Conducting Social Impact Assessments

As with other aspects of AI, there is a set of underlying core values that informs the development and implementation of SIAs. These values also delineate how these assessments should be conducted. It is useful to refer to these sets of principles when it comes to executing SIAs in the design and deployment of crime prediction technologies in your context.

These values acknowledge that universal rights are shared equally among cultures, must be respected by the rule of law, and applied equally and fairly to everyone. Because of this, people have a right to be involved in the decisions that are made on their behalf, especially the interventions that are designed to benefit them. This is essential to building trust and legitimacy, but also to ensure that local knowledge and expertise is used to enhance planned interventions.



Source: International Principles for Social Impact Assessment prepared by Frank Vanclay for the International Association for Impact Assessment, May 2003: https://www.socialimpactassessment.com/documents/0303%20Vanclay%20IAPA%20V21N1%20SIA%20International%20Principles_1.pdf

Conclusion

Governments across the world are keen to leverage the opportunities presented by the 4th Industrial Revolution by modernising their systems of government to remain competitive and efficient. However, the desire to modernize and optimize efficiencies must be embraced within the broader societal context to guarantee that innovations for progress do not infringe upon the rights of others by introducing new risks. Remaining cognizant of those risks while committed to evidence-based solutions - and also open to reassessment when the evidence does not always fit the realities of the Global South – is essential to designing and implementing AI-driven technologies that are ethical, socially responsible, and serve the interests of a government and its people.

Appendices

A. Bibliography of references

1. Abassi, A., Li, J., Clifford, G., Taylor, H. (2018, Aug 1). Make "Fairness by Design" Part of Machine Learning. *Harvard Business Review*.
2. Adams, R., et al. (2021). Human Rights and the Fourth Industrial Revolution in South Africa, Chapter 7. Cape Town, HSRC Press.
3. Aguirre, C., Badran, E., Muggah, R. (2019, July). FUTURE CRIME: Assessing twenty first century crime prediction. Strategic Note 33. *Instituto Igarapé*.
4. Australian Government, Department of Industry, Science and Resources. (2019). Australia's Artificial Intelligence Action Plan.
5. Brantingham, P.J., Brantingham, P.L., Andresen, M.A. (2017, Jan 1). The geometry of crime and crime pattern theory. *CrimRxiv*.
6. Brantingham, P.J., Brantingham, P.L. (2008) Crime Pattern Theory. In R. Wortley (Ed.). Environmental Criminology and Crime Analysis. 1st Edition, pp 1-18. London. Taylor & Francis Publishing Group.
7. Brayne, S. Rosenblat, A., Boyd, D. (2015, Oct 27). Predictive Policing. *DATA & CIVIL RIGHTS: A NEW ERA OF POLICING AND JUSTICE.*
8. Brown, A. (2020, Feb 7). Biased Algorithms Learn from Biased Data: 3 Kinds Biases Found in AI Datasets. *Forbes Online*.
9. Brown, S. (2021, Apr 12). Machine learning, explained. *MIT Sloan*.
10. Brunson, R.K. (2020, June 12). Protests focus on over-policing. But under-policing is also deadly. *The Washington Post*.
11. Burgess, M. (2020, Aug 6). 'Police built an AI to predict violent crime. It was seriously flawed'. *WIRED UK.*
12. Data for Development (D4D). (2021). Global Index on Responsible AI.
13. Deahl, D. (2018, Apr 12). Suspect caught in China at music concert after being detected by facial recognition technology. *The Verge*.
14. ECP | Platform for the Information Society. (2018) Artificial Intelligence Impact Assessment.
15. European Commission: Directorate-General for Research and Innovation. (2021, Nov 25). Ethics By Design and Ethics of Use Approaches for Artificial Intelligence.
16. European Commission: Directorate-General for Research and Innovation. (2018). Ethics of Artificial Intelligence.
17. European Commission Futurium. (2019). Ethics Guidelines for Trustworthy AI.
18. European Union General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. Article 35.
19. Floridi, L., Cowls, J. (2019, July 2). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, Edition 1.1. DOI 10.1162/99608f92.8cd550d1.
20. Foody, K. (2020, Jan 24). Chicago police end effort to predict gun offenders, victims. *Associated Press News*.
21. Green, B., Horel, T., Papachristos, A.V. (2017). Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *JAMA Intern Med.* 177(3):326–333. doi:10.1001/jamainternmed.2016.8245
22. Guariglia, M. (2020, Sept 3). Technology Can't Predict Crime, It Can Only Weaponize Proximity to Policing. *Electronic Frontier Foundation*.
23. Hardesti, L. (2018, Feb 11). Study finds gender and skin-type bias in commercial artificial-intelligence systems. *MIT News*.
24. Heaven, W. (2020, July 17). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*.
25. House of Lords. (2017). AI in the UK: ready, willing and able?. Artificial Intelligence Committee.
26. Hvistendahl, M. (2016, Sept 28). Can 'predictive policing' prevent crime before it happens? *Science*.

27. Ibarra, N. (2020, June 24). Santa Cruz, Calif., Bans Predictive Policing Technology. *Government Technology.*
28. IBM Corp. (2022). Everyday Ethics for Artificial Intelligence.
29. Institute of Electrical and Electronics Engineers. (2019). IEEE's Ethically Aligned Design.
30. International Association for Impact Assessments. (2009). Social Impact Assessment.
31. Kerry, C. (2020, Feb 10). Protecting privacy in an AI-driven world. *Brookings Institute*.
32. Lau, T. (2020, April 1). Predictive Policing Explained. *Brennan Center for Justice*.
33. Law Insider. "surveillance technology, n. 1." LawInsider, Inc. Sept 2022. https://www.lawinsider.com/dictionary/surveillance-technology
34. Miriam Webster. 'fairness, n.1.". Online Dictionary. Sept 2022. https://www.dictionary.com/browse/fairness
35. Muggah, R. (2018, June 15). How smart tech helps cities fight terrorism and crime. *World Economic Forum*.
36. Muggah, R. (2017, Feb 2). What happens when we can predict crimes before they happen? *World Economic Forum.*
37. Nagitta, P.O., Mugurusi, G., Obicci, P. A., Awuor, E. (2021). Human-centered artificial intelligence for the public sector: The gatekeeping role of the public procurement professional. *Procedia Computer Science*
38. Organisation for Economic Cooperation and Development (OECD). (2020). OECD AI Policy Observatory.
39. Organisation for Economic Cooperation and Development (OECD). (2019) OECD Artificial Intelligence (AI) Principles.
40. Oxford English Dictionary. "artificial intelligence, n. 1." OED Online. Oxford University Press, Sept 2022. https://en.oxforddictionaries.com/definition/artificial_intelligence.
41. Oxford Insights and the International Research Development Centre. (2020). The Government Artificial Intelligence (AI) Readiness Index (2020).
42. Perelman, L. (n.d.) "Babel Generator": Automated Essay Scoring (AES). *Wordpress*.
43. Project Sherpa. (2020). Recommendation: Develop baseline model for AI impact assessments.
44. Rai, A. (2019, Dec 17). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*. Vol 48. 137–141.
45. Research ICT Africa. (2019). Recommendations on the inclusion sub- Saharan Africa in Global AI Ethics. RANITP Policy Brief 2.
46. Spielkamp, M. (2017, June 12). Inspecting Algorithms for Bias. *MIT Technology Review*.
47. Stirling, R., Pasquarelli, W., Shearer, E. (2021). An Assessment Framework for Measuring Government AI Readiness. *Oxford Insights.*
48. Stone, K. (2022, Aug 2). Are South Africa's police jumping the gun on new technologies?. *ISS Today Africa*.
49. Stone, K. (2020, March). RESPONSIBLE USE OF ARTIFICIAL INTELLIGENCE FOR CRIME PREVENTION IN SOUTH AFRICA. Policy Action Network. *Human Sciences Research Council*.
50. Suresh, H., Guttag, J. (2021, Aug). Understanding potential sources of harm throughout the machine learning lifecycle. *MIT Schwartz College of Computing.*
51. "Understanding Discrimination and Bias". JED Foundation. <Accessed on 15 Sept 2022>
52. United Nations Educational, Scientific and Cultural Organisation (UNESCO). (2022). Recommendation on the Ethics of Artificial Intelligence
53. University of Montreal. (2017). Montreal Declaration for the Responsible Development of AI.
54. US Legal Definitions. 'public safety, n. 1." USLegal, Inc. Sept 2022. https://definitions.uslegal.com/p/public-safety/
55. Vanclay, F. (2003). International Principles for Social Impact Assessment. International Association for Impact Assessment.
56. Wiggers, K. (2021, May 29). Adversarial attacks in machine learning: what they are and how to stop them. *Venture Beat.*
57. World Economic Forum. (2020, June). AI Procurement in a Box: AI Government Procurement Guidelines.
58. Yeung, D. (2018, Oct 22). Intentional Bias Is Another Way Artificial Intelligence Could Hurt Us. *The RAND Blog.*

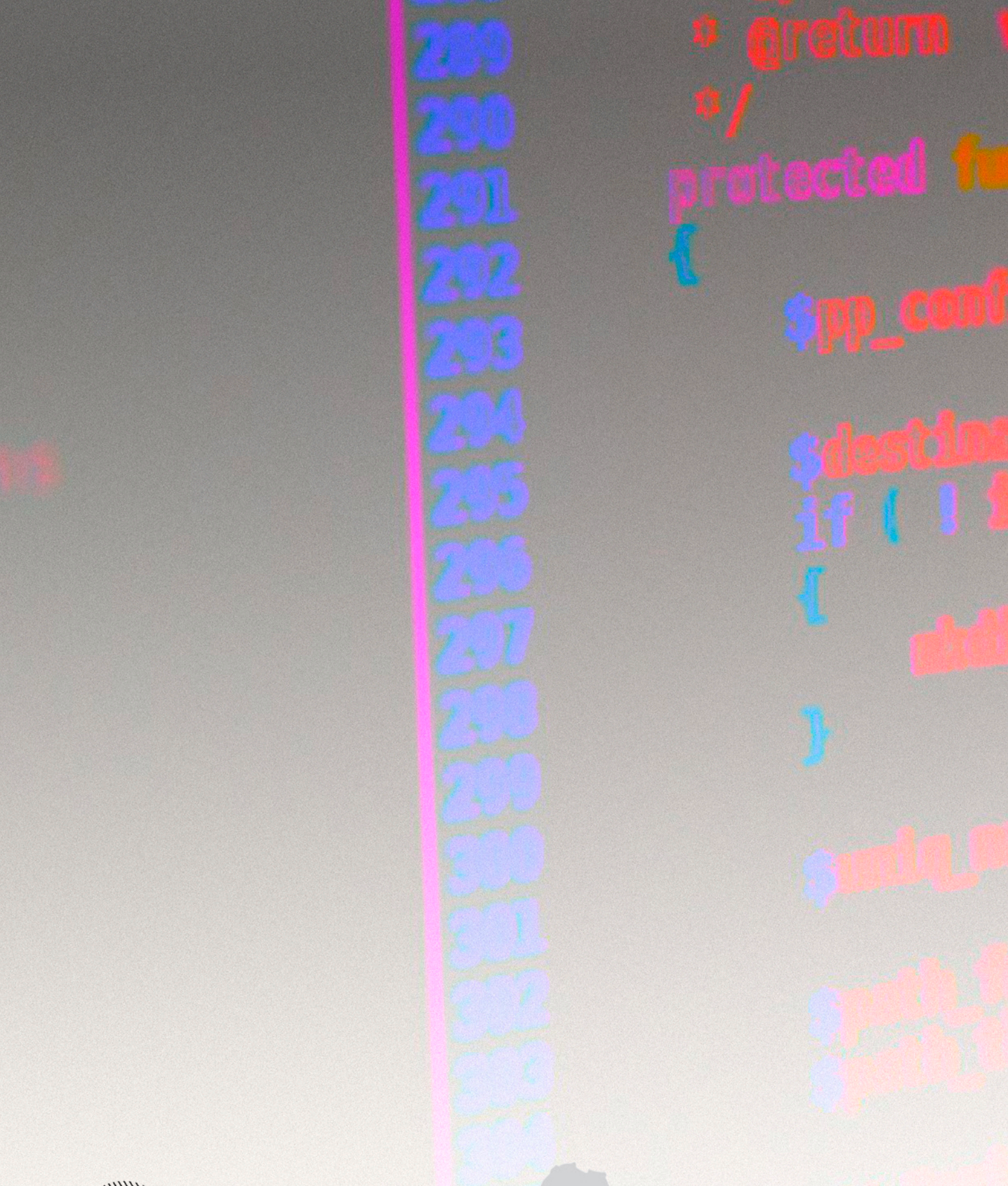
B. Implementation tools

The table below provides access to a series of implementation frameworks and tools that may assist you in the ethical implementation and use of crime prediction technologies. In addition, Data for Development has also put together a comprehensive repository of tools and frameworks from across the globe which can be found here.

TYPE	DESCRIPTION
Process Map	a. This map outlines five steps to follow before implementing an AI solution, which includes: (1) Establishing clear goals and project guidelines; (2) Conducting thorough amounts of data; (3) Defining the algorithmic learning model; (4) Identifying the right AI partners; and (5) Planning for human oversight and decision-making power.
Social Impact Matrices	This matrix presents a risk assessment framework for bias in predictive policing technologies in a structured and easy-to-read format. Each column presents a different Level of Bias, starting with socio-spatial factors, then moving towards resource allocation, down to targeting practices, and ending with potential sources of bias in the algorithm itself. The Level of Bias is meant to correspond with how risks can be structured at an individual, relational, community and societal level, but is by no means exhaustive and may be edited to suit the needs of a particular context.
Score Cards	
Fairness Checklist for Implementers	The items included in this checklist should be treated as a starting point for implementation teams to customize, as most teams will likely need to add, revise, or remove items to fit their specific circumstances. It is important to note that undertaking the items in this checklist will not guarantee fairness, but rather intended to prompt reflection and further discussion on how fairness can be strengthened.
Procurement Guidelines	These guidelines outline key considerations that should be addressed by the government *before* it acquires and deploys AI solutions and services. They should not be treated as a silver bullet for resolving all AI-related challenges in the public sector, but as a tool to help the government in the ethical use and implementation of AI systems.
Readiness Assessment Framework	This tool allows governments to assess their readiness for AI by conducting an assessment of each element of the AI ecosystem (government, policy, private sector, academia, and data) as well as of the strength of the relationships between those elements. Time constraints and limits on the availability of data may not make it possible to conduct a full assessment, so this framework should be used to identify priority areas for focus and intervention.

C. Proposed methodology for designing and implementing SIAs

- Building a conceptual framework for the SIA - individual, community and societal harms (and benefits)
- Pre + post-implementation stage, on an iterative and ongoing basis
- Potential sources of data + diversification of datasets + testing for fairness
- Building a multidisciplinary, diverse team
  - Composition of SIA Team - combination of internal/external personnel with the requisite competencies and capacities + not a once-off
  - Integrated into job description and performance management systems
  - Oversight and governance structures
- Qualitative data collection
  - Outreach to communities
  - Proposed persons to interview
  - Script for semi-structured interviews
  - Confidentiality and Anonymity
  - Feedback and validation using co-creative methodologies
- Social Impact matrices, score cards
- Responding to the results and modifying/amending/terminating deployment of crime prediction technologies accordingly



IGARAPÉ INSTITUTE
a think and do tank

igarape.org.br

ISS INSTITUTE FOR SECURITY STUDIES

issafrica.org