DCAF Geneva Centre for Security Sector Governance

Ministry of Foreign Affairs of the Netherlands

Mitigating AI biases in security institutions



Practical tools for inclusive security sector governance

Acknowledgements

This inaugural guidebook of DCAF's Inclusive Approaches to Artificial Intelligence (IA to AI) series was developed as part of DCAF's 'Enhancing SSG/R Policy and Practice 2025' project, funded by the Netherlands' Ministry of Foreign Affairs. It was informed by an expert consultation held in May 2026. DCAF gratefully acknowledges the valuable contributions of all participants, whose insights helped shape the analysis presented in this publication. DCAF also extends its sincere appreciation to the external expert reviewers for their thoughtful feedback and critical reflections, which strengthened the guidebook and its practical relevance.

Authors: Amanda Van Dort (PeacePays.AI); Laylo Merali (DCAF)
Reviewers: Dr Rumman Chowdhury (Humane Intelligence PBC); Dr Darja Schildknecht (Swiss Armed Forces); Nathalie Gendre, Dr Dawn Lui and Matteo Polizzi (DCAF); Jennifer Bradshaw and Katrina Fotovat (PeacePays.AI)
Copy editor: Cherry Ekins
Layout and design: Lea-Maija Laitinen / leamaija.works
Publisher: DCAF – Geneva Centre for Security Sector Governance.

About DCAF

DCAF – Geneva Centre for Security Sector Governance is dedicated to improving the security of states and their people within a framework of democratic governance, the rule of law, respect for human rights, and gender equality. Since its founding in 2000, DCAF has contributed to making peace and development more sustainable by assisting partner states, and international actors supporting these states, to improve the governance of their security sector through inclusive and participatory reforms. It creates innovative knowledge products, promotes norms and good practices, provides legal and policy advice, and supports capacity-building of both state and non-state security sector stakeholders.

© DCAF – Geneva Centre for Security Sector Governance, 2026.

DCAF encourages the use and dissemination of this publication. However, users are requested to acknowledge and cite the material appropriately and not alter its content. All rights reserved.

Disclaimer: While the guidebook offers a foundation on awareness of biases in AI and on non-exhaustive approaches to mitigation, it does not promote the use of AI tools by security institutions.

ISBN: 978-92-9222-845-3

Mitigating AI biases in security institutions

Practical tools for inclusive security sector governance

DCAF Geneva Centre for Security Sector Governance

Ministry of Foreign Affairs of the Netherlands

List of acronyms

AI	artificial intelligence
CoT	chain-of-thought
CSO	civil society organization
DCAF	Geneva Centre for Security Sector Governance
EU	European Union
HITL	human-in-the-loop
ICRC	International Committee of the Red Cross
IHL	international humanitarian law
ISR	intelligence, surveillance, and reconnaissance
SOP	standard operating procedure
SSG	security sector governance
SSG/R	security sector governance and reform
TFGBV	technology-facilitated gender-based violence
UNIDIR UN	Institution for Disarmament Research
WHRD	woman human rights defender
WPS	Women, Peace, and Security

Contents

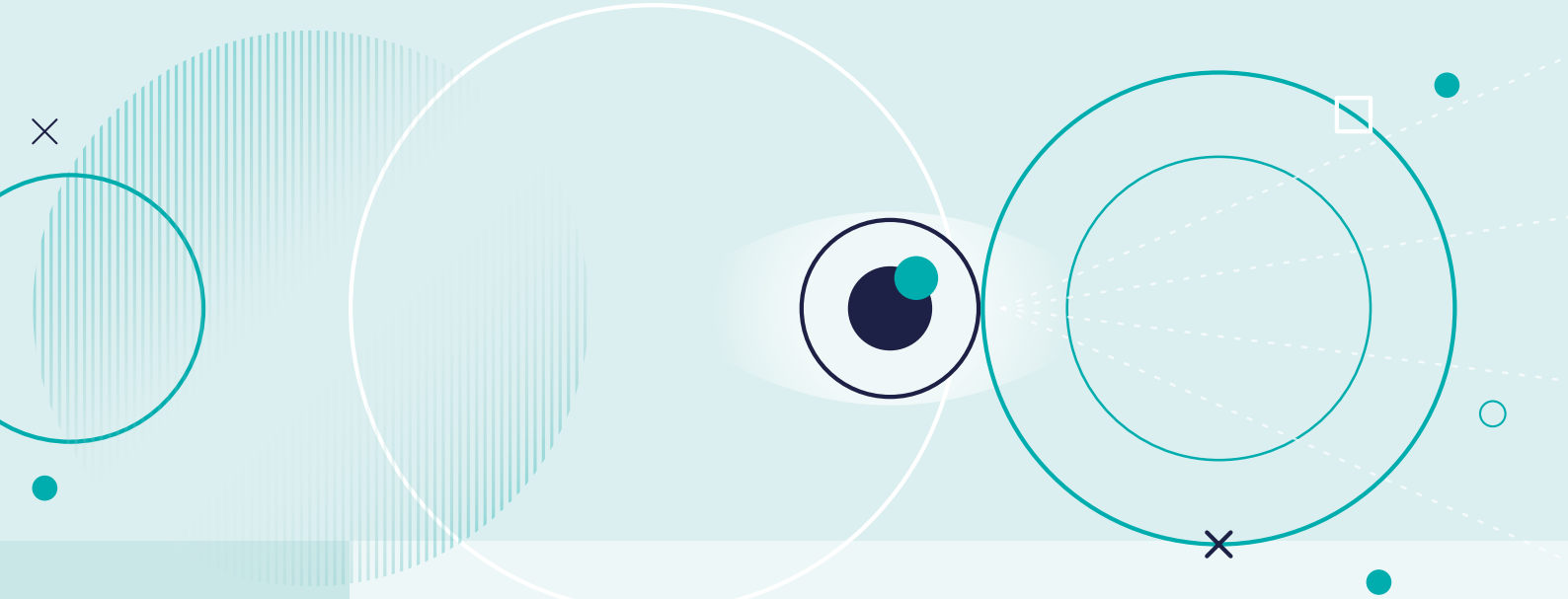
6 **Overview**
Context and purpose 6
Executive summary 8
Key concepts 11
Limitations and considerations 12
Structure of the guidebook 13

14 **Section I: Understanding AI bias in security institutions**
Why AI bias matters for security institutions 15
What bias in AI systems means for the security sector 17
How security institutions use AI systems 18
Where bias enters AI systems in the security sector 22
Governance and oversight 28
Differentiated impact of AI-supported security decision-making 30
Consequences of bias in AI systems 32
Summary: Implications for SSG/R 39

42 **Section II: Mitigating AI bias in security institutions**
Frontline use: Governing AI use in security institutions 44
Legal: Ensure compliance and rights protection 48
Institutional: Establish accountability and control 52
Management: Monitor, adapt, and maintain legitimacy 56
Scenario-specific checklists 61
Conclusion: From risk awareness to institutional practice 69

70 **Section III: Toolkits**
Tool 1: Scenarios, questions, and interpretation frameworks for security institutions 73
Tool 2: AI bias risk assessment 79
Tool 2A: Risk assessment for use of commercial AI tools 82
Tool 2B: Risk assessment for feeding institutional data into a commercial AI system 88
Tool 2C: Risk assessment for institutions building or training their own AI system 95
Tool 3: Inclusive security and gender analysis template 103
Additional tools 118

Overview



Context and purpose

The deployment of AI-enabled security sector tools has become a reality, including in policing, justice, border management, intelligence, and the military, among others institutions. The debate is mainly on how, to what extent, and to what effect the tools are used. In light of their enhancement of speed, ability to process large volumes of information, and real or perceived potential for increases in accuracy, AI-enabled tools are developing and deploying widely and deeply.

However, AI deployment in security institutions often takes place more rapidly than guardrails are enacted, which is unparalleled compared to any other technological development. This in turn shapes high-risk decisions on human security, with algorithms that reflect and amplify human biases around gender, race, ethnicity, and overall structural discrimination.

Consequently, a distinct security sector governance challenge emerges: bias embedded within AI systems may systematically shape security outcomes in ways that undermine effectiveness, accountability, and human rights protections. It is already hard enough to address human bias, and one can imagine the challenge to identify and mitigate this in AI systems. But mitigating bias in AI systems is fundamental to governing the risks associated with the use of these tools.

While some define security institutions quite narrowly as structures that utilize force, the concept is broader: it includes those mandated to control how force is utilized via management and oversight, justice system actors linked to provision of security, and non-state actors in security provision and civil society.

This guidebook focuses on how security institutions can understand and address bias in AI systems. Rather than treating AI bias as only an extension of human or institutional bias, it recognizes that the ways AI systems are designed, deployed, and used can create specific risks for how bias enters the technology and persists in outputs. While institutional practices, data sources, and security sector cultures may influence the development of AI systems, these systems can also automate bias in new ways.

The guidebook is designed broadly for security sector leaders, policymakers, oversight bodies, and practitioners involved in approving, procuring, deploying, using, or reviewing AI systems. Technical partners, civil society – especially tech-focused civil society organizations – and women peacebuilders are among those who can likewise benefit, given their roles around governance, monitoring, and accountability processes. While developed with leadership-level decision-making in mind, the guidebook also provides practical checklists and tools that leaders can apply to training and frontline functions.

Given the potential for AI systems to reproduce and amplify existing inequalities,

this guidebook seeks to locate bias mitigation within the broader context of inclusive security governance and the Women, Peace, and Security (WPS) agenda. Throughout the guidebook, WPS is used as a practical framework to better understand and improve security effectiveness, legitimacy, and responsiveness through meaningful inclusion and recognition of differential impacts.

Executive summary

AI biases are often ascribed to 'algorithms', but the causes are much broader, especially as they pertain to security institutions. They include training data, human feedback cycles, deployment contexts, and usage patterns – and, importantly, institutional governance. They are therefore more than just technical issues.

Bias is not confined to a single failure point. AI bias can emerge during any stage of development – problem definition, data collection, model development, procurement, deployment, or monitoring. Hence, it can shape outputs that affect near-term decisions, such as threat detection, and longer-term security outcomes.

If an AI system's outputs are systematically biased, causing unjustified behaviour changes by decision-makers or frontline personnel, this may continue or worsen the discrimination that certain groups face based on gender, race, or class, among other characteristics.[1] Unlike isolated technical errors, biases are embedded in how systems are designed, trained, and deployed, and can persist across time and contexts. There are three key forms of such biases *(Figure 1)*.

AI systems not only reproduce existing institutional and societal biases, they amplify, scale, and embed bias in ways that fundamentally reshape security decision-making.[2] Their capacity to operate at speed, across large populations, and often without transparency means that biased assumptions can be reproduced and institutionalized in ways that exceed the impact of individual human decision-making.[3] These risks are particularly significant in security institutions, where the outcomes of the use of AI-enabled systems affect liberty, protection, surveillance, and use of force. There are broadly four reasons why these biases matter for the security sector *(Figure 2)*.

Security sector governance (SSG) provides a frame by which to understand the implications of biases. SSG includes national security and human security – which is linked to economic, social, and political stability – ensured through systemically sound institutions.[4] The effects of biased security sector AI systems are broad, from undermining frontline decisions to eroding public trust. There are seven key principles of security sector governance and reform (SSG/R), each of them relevant to the development and deployment of AI-enabled tools.

- **Accountability.** AI-powered systems can reduce auditability and lines of responsibility, challenging the mandates of oversight bodies to ensure the fulfilment of expectations. This is true unless AI tools are deployed towards bias-mitigated checks and balances on security sector actors.

Figure 1. Overlapping forms of biases in AI-enabled systems.

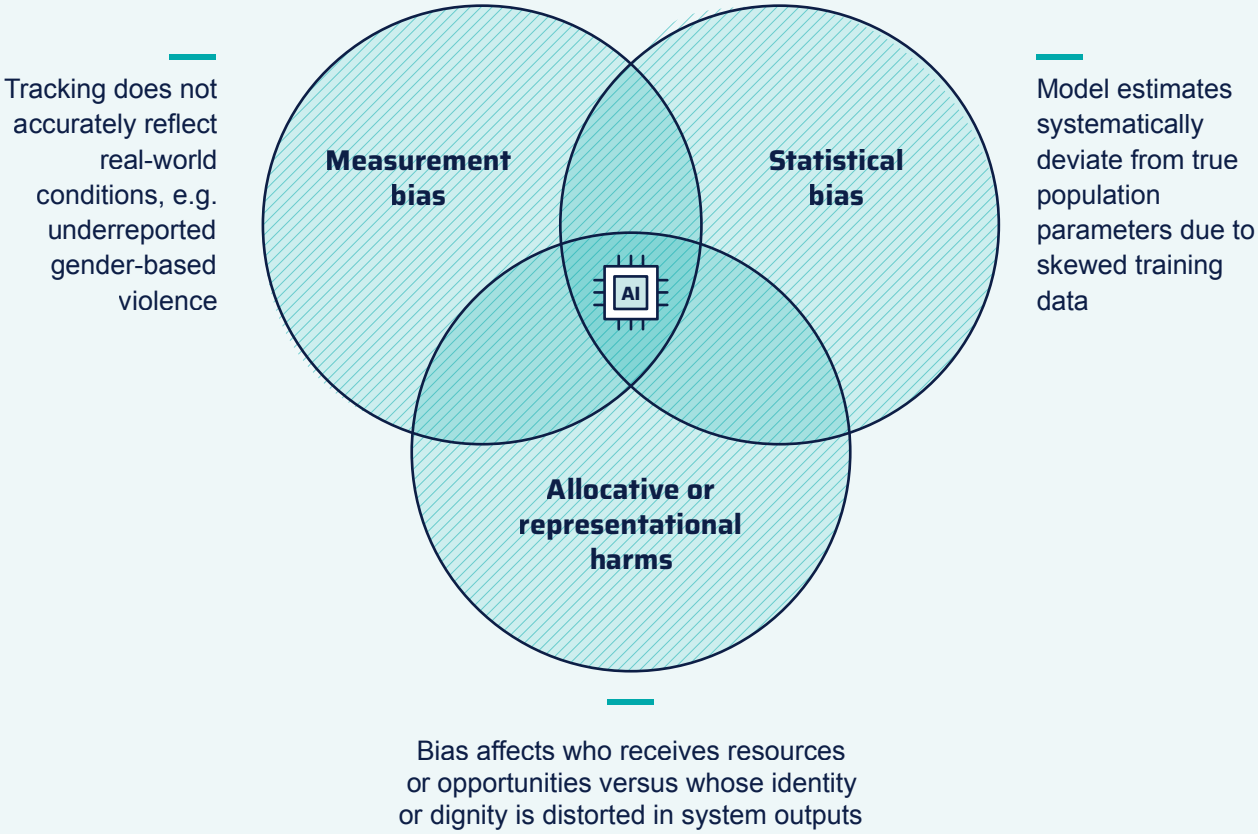
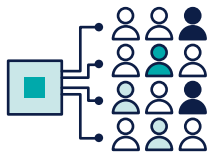


Figure 2. Why bias in AI matters in security sectors.



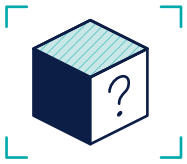
Scalability

Deployed across large populations



Persistence

Embedded across institutional processes



Opacity

Difficult to interpret or contest



Implementation

Translated into real-world decisions

- **Transparency.** The complex algorithms that underpin AI-enabled tools used by security institutions are rarely available to, or understandable by, affected stakeholders – and even the mere use of AI is not always disclosed.

- **Rule of law.** If unchecked, AI-generated outputs can lead to violations of impartiality, due process, equal treatment, and human rights law.

- **Participation.** If decisions are made, or services are provided, based on unknown or insufficiently understood algorithms, participation in effect is limited to those having privileged technical access or insight.

- **Responsiveness.** If aspects of security provision are outsourced de facto to algorithms, and if the algorithms have biases at the design stage, a service-oriented security sector culture cannot inclusively be fulfilled. Certain groups become disproportionately targeted for surveillance or enforcement, or denied protection, whereas other groups could unfairly be advantaged.

- **Effectiveness.** Biases in AI tools cause faulty threat detection, intelligence and data prioritization, and distribution of resources, impeding the alignment to the professional standards of security institutions. Alternatively, if AI is well utilized with biases satisfactorily mitigated, effectiveness can be strengthened.

- **Efficiency.** Typically, AI-enabled tools increase efficiency in use of public resources, and towards the desired outcomes. The issue is whether this increase in efficiency is justified by sacrifices in other principles.

Mitigating AI bias requires ensuring that AI systems are designed, deployed, and used in ways that support the core objectives of SSGR: effectiveness, legality, accountability, and legitimacy.

Ultimately, mitigating AI bias is not simply about improving algorithms. It is about ensuring that security institutions retain responsibility for how AI shapes decisions, exercise effective oversight throughout the system life cycle, and remain accountable for the outcomes that AI-enabled decisions produce.

Mitigation is not a one-time technical exercise undertaken during system development or procurement. It is continuous institutional responsibility that spans the entire AI life cycle, from defining the purpose AI is intended to serve through procurement, deployment, oversight, monitoring, etc. Institutions are responsible for:

- defining why and how AI will be used;
- ensuring that systems are lawfully authorized and technically appropriate;
- governing how outputs are used;
- monitoring, evaluating, and adapting systems and practices based on lessons learned.

These mitigation steps require engagement across frontline, technical, legal, and oversight functions. This engagement can benefit from systematically applying checklists relevant towards different scenarios, which help prompt the right questions, and a set of detailed tools that provide more comprehensive methods by which to assess and mitigate AI biases in security institutions.

Key concepts

These concepts provide a foundation for understanding how bias operates in AI-supported security systems.

- *Artificial intelligence (AI)* systems are computational technologies that use data to make predictions, classifications, recommendations, or decisions based on patterns in data. In security institutions, AI systems may be used to support activities like intelligence analysis, surveillance and screening, border management, threat detection, targeting support, logistics, cybersecurity, or early warning alerts.

- *Biases in AI systems* are systematic and repeatable discrepancies in system outputs or performance that create unjustified advantage or disadvantage for particular groups, places, languages, behaviours, or risk profiles. In security institutions, AI bias matters because security decisions may affect who is deprived of their liberty, who receives protection or mobility, who is targeted, or who can access services.

- *Error versus bias.* Error is when a system makes an inaccurate decision or prediction. Bias occurs when a system produces systematic errors and unequal outcomes.

- *Discrimination* – In security institutions, biased outcomes are discriminatory when used to treat different people unequally or unlawfully.

Three overlapping forms of bias are relevant to security practitioners.

- *Algorithmic bias* in AI systems can occur at many stages: in the data that is used to train or test the system, how data is labelled, how models are designed, how proxy variables that act as stand-ins for sensitive data like race or gender are used, how performance thresholds are set for different groups, and what the system is optimized to do.

- *Statistical bias* occurs when a model's estimates consistently deviate from true population parameters because the training data is skewed towards certain groups or contexts. Allocative and representational harms occur when bias determines who receives resources, protection, or scrutiny (allocative), or when system outputs distort how identities are described or classified (representational). These distinctions matter because each requires different technical remediation.

- *Institutional bias* in security services can stem from policies, incentives, frontline use assumptions, organizational culture, or historical patterns of behaviour enforcement. AI systems can encode institutional cultures and patterns into technological design and applications.

AI bias also affects populations differently depending on gender and intersecting identities.

- *Gender* is defined as the socially constructed norms, roles, activities, and behaviours associated in a given society at a given point with masculinity or femininity. Often, gender is explained in contrast to sex, which is defined by biological characteristics.

- *The Women, Peace, and Security (WPS)* agenda was established by UN Security Council Resolution 1325 in 2000. This global framework focuses on promoting the safety of women and girls in conflict and crises, and supporting the meaningful participation of women in political and security decision-making processes.

- *Intersectionality* is an analytical framework for understanding how overlapping identities shape experiences of discrimination, exclusion, and power and privilege, including the combination of gender, race, ethnicity, religion, social status, age, disability, etc.[5]

Limitations and considerations

- *AI is evolving rapidly.* Capabilities, deployment practices, and governance frameworks are evolving faster than most traditional research and publication cycles. New tools, risks, and policy debates continue to emerge, meaning that some examples and recommendations may require periodic updating.

- *Many security applications remain difficult to evaluate externally.* Classification restrictions, procurement confidentiality, and proprietary systems limit public visibility into how some technologies are designed and used, particularly in defence and intelligence settings. Much of the available evidence therefore comes from investigative reporting, court cases, publicly documented incidents, or the relatively small number of systems subject to independent review.

- *The evidence base is uneven across regions and institutional contexts.* Research and public reporting on AI bias tend to be concentrated in higher-income countries with stronger oversight institutions and civil society monitoring. There is far less publicly available documentation from many conflict-affected or lower-resource settings that may face some of the highest risks from weak oversight and biased systems. Practitioner interviews partially addressed this gap.

- *Implementation capacity varies significantly, impacting mitigation measures.* Many recommendations in Section II assume the existence of

governance infrastructure, including legal expertise, technical capability, independent oversight, and continuous monitoring, that institutions in fragile, conflict-affected, or lower-resource settings may not have. In these contexts, a minimum viable safeguards approach may be more practical, emphasizing clear prohibitions on high-risk uses, mandatory pre-deployment testing, and contractual rights to audit systems and suspend their use when necessary.

- *Bias is difficult to attribute to a single source in many AI systems.* In practice, algorithmic outputs are often shaped by existing institutional patterns embedded in historical data, frontline reporting, watchlists, or enforcement decisions. Those outputs reinforce future institutional behaviour. Because these feedback loops make causal attribution difficult, this guidebook treats bias primarily as a governance and accountability challenge, emphasizing oversight, transparency, and institutional safeguards rather than attempting to identify a single technical cause in individual cases.

- *This guidebook is not an in-depth analysis on AI use by any specific aspect of the security sector.* While it offers examples and scenarios on policing, intelligence, border management, defence, etc., these are meant to illustrate broader concepts.

These limitations do not diminish the relevance of the findings, but highlight the importance of applying the guidebook alongside ongoing monitoring, local context, and periodic review as AI technologies and governance frameworks continue to evolve.

Structure of the guidebook

This guidebook is structured in three sections:

Section I explains key concepts: what AI bias looks like in security institutions, how AI systems are typically used in security applications, where bias can occur during an AI system's life cycle, who may be affected by biased AI systems, and the consequences for SSG/R. Where and how are security institutions already integrating AI into frontline decision-making? How does bias enter AI systems and become reinforced over time?

Section II outlines practical measures that institutions can adopt to mitigate AI bias across governance, design, deployment, monitoring, and oversight functions. Which governance approaches appear most useful for reducing these risks in practice, particularly in security environments where transparency and oversight may already be limited?

Section III provides tools and templates that security institutions can use and adapt to support mitigating AI bias in their organizations.

SECTION I

Understanding AI bias in security institutions



This section examines how bias emerges in AI systems used by security institutions, and why it matters for effectiveness, legal compliance, accountability, and legitimacy. Specifically, it explores how AI systems are used across the security sector, where bias can enter throughout the AI life cycle, and how these dynamics affect different populations and institutional outcomes. Examples illustrate how bias operates in practice. Key implications for security institutions are articulated. These are linked to mitigation measures and checklist items in Section II, and the risk assessment tools in Section III.

Why AI bias matters for security institutions

AI is already embedded in the security sector, as governments scale across core national security domains such as defence, intelligence, and law enforcement, among others. AI applications include prioritizing cyber alerts, processing surveillance feeds, analysing imagery, identifying persons of interest, forecasting risk or crime trends, supporting border screening processes, remote biometric identification, and enabling military operations.

The uses of AI are broadening. For example, military use of AI has expanded beyond autonomous weapons to a broader range of capabilities. These include, among others, analysis of intelligence, planning operations, logistics, and training.[6] This comes in addition to more common integrations of AI in security institutions, e.g. three-quarters of organizations integrate AI in cybersecurity functions.

AI increasingly shapes risk identification, resources allocation, and decision-making about the inherently high-stakes functions of surveillance, targeting, and protection.

When these functions are biased, their outcomes are distorted in three ways:

1. Institutional effectiveness is undermined, i.e. security provision is diminished.

2. Individuals' rights, safety, and access to services are affected, often impacting women, men, girls, boys, and gender-diverse individuals differently. This can influence whether individuals are subject to surveillance or detention, able to move freely, afforded protection, or targeted for enforcement or use of force.

3. Institutional legitimacy is impacted, from a combination of reduced effectiveness and implications for individuals.

This means AI systems do more than automate tasks; they also shape what information is presented to decision-makers. They influence which risks are escalated, which people or places are flagged, and how responses are prioritized. Therefore bias in AI systems is not only a technical challenge; it is a governance issue that affects the functioning of security institutions *(Figure 3)*. Within an SSG/R framework, bias shapes the ability of institutions to operate effectively, uphold legal obligations, and maintain legitimacy.

The risk is not only that security institutions deploy biased AI systems but that the systems scale and accelerate bias, and make it less detectable or visible. The International AI Safety Report notes that advances in AI capabilities are outpacing understanding of the risks, including reliability and safety, misuse, autonomy, and governance gaps.[7] Furthermore, the opacity of many AI systems makes it difficult to assign accountability when harms occur.[8]

These dynamics extend beyond individual decisions. Over time, biased systems can reshape institutional practices, reinforce patterns of exclusion, and erode trust between security actors and the populations they serve – weakening cooperation and undermining long-term stability. Addressing AI bias is therefore essential not only for fairness, but for the effective functioning of security institutions.

In short, AI systems are not neutral tools. They are decision-shaping systems that should be evaluated as rigorously as doctrine or policy. Understanding why bias matters requires first clarifying what bias means in AI systems and how these patterns emerge.

Figure 3. Manifestations of governance failures of AI in security institutions.



Misallocated resources

Heavier attention on already highly surveilled individuals



Missed threats

Failure to detect threats that fall outside existing data patterns



Disproportionate scrutiny

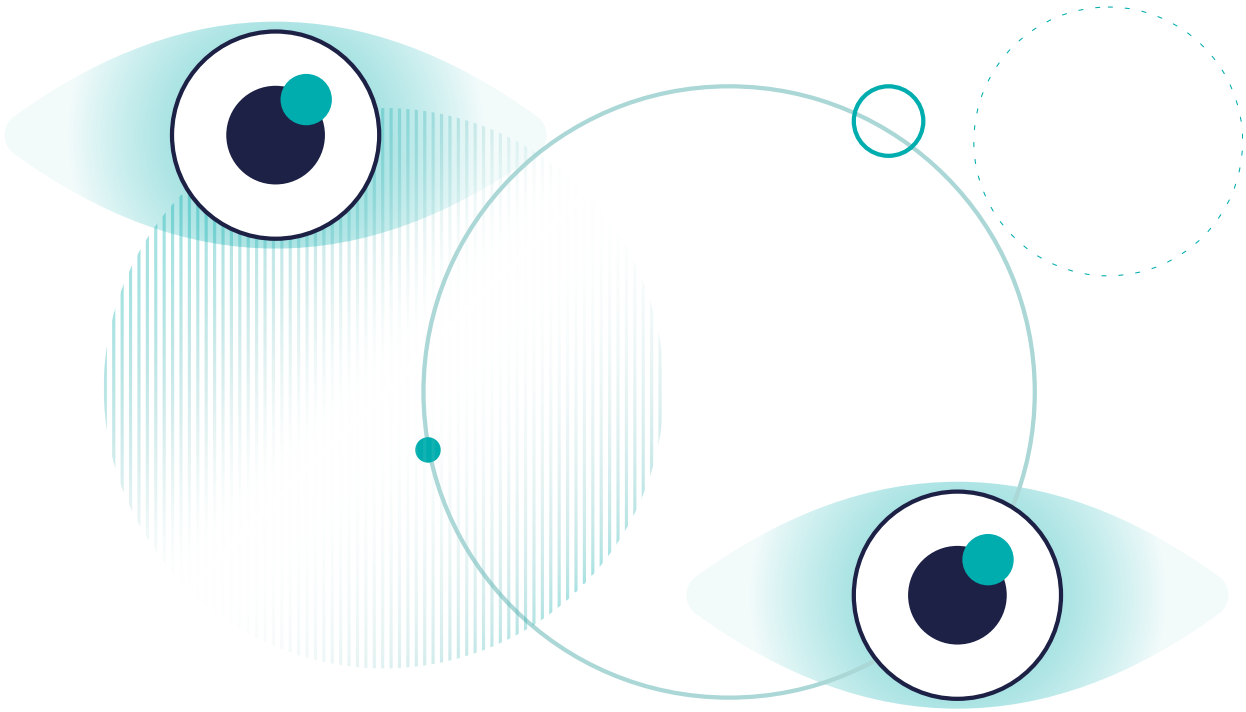
Specific population segments are unfairly targeted

What bias in AI systems means for the security sector

Unlike random errors, which occur unpredictably, biases in AI systems produce consistent patterns that advantage or disadvantage particular populations.[9] Discrimination arises when these patterns result in unequal or unlawful treatment. Bias does not necessarily mean that a system has low overall accuracy, as systems can perform well on average while producing inaccurate or inequitable outcomes for specific populations or contexts. Understanding these distinctions is particularly important in security contexts.

AI shapes consequential decisions across the security sector: who is identified as a threat or placed on watchlists; who is surveilled, stopped, or detained; who is investigated and prosecuted; who is prioritized for protection or allocation of resources, etc. In doing so, these systems also determine which data is surfaced, which risks are elevated or ignored, and how security is ultimately distributed, often reinforcing existing biases and inequities embedded in the underlying data and decision-making processes.

Although AI systems reproduce bias through technical design, the underlying sources of bias are rarely purely technical. Much of the data used in security contexts is derived from past human decisions: arrests, stops, incident reporting, watchlists, intelligence files, battlefield reports, and surveillance records. These sources may embed unequal treatment due to enforcement patterns, underreporting, political priorities, or blind spots. When trained on these records, systems may appear to identify and justify these patterns as objective risk indicators – even though they reflect biased outcomes encoded as technical signals.



EXAMPLE 1

Facial recognition biases leading to wrongful arrests

One study found that facial recognition algorithms are more likely to misidentify Black and Asian individuals with photos of known criminals than to misidentify white individuals with the same criminal photos.[10] Another study found that the commercial facial analysis software tested incorrectly identified darker-skinned women up to 35 per cent of the time, compared to a less than 1 per cent failure rate for light-skinned men.[11]

The stakes increase when systems are used in real-world settings. In the United States, biased-use cases have led to multiple documented incidents in which individuals were arrested for crimes they did not commit based on inaccurate matches produced by an AI system.[12] Authorities treated matches as definitive evidence rather than investigative leads requiring corroboration.

At scale, false positives and false negatives can put entire categories of people at greater risk of being stopped, searched, and detained. This creates both effectiveness risks (matching to the wrong person) and rule-of-law risks (targeting certain demographics for enforcement).

These risks are the product of flawed algorithms from the design of technology, selection of training data, and choices made during model optimization. They are also shaped by institutional policies, practices, or historical trends.

Thus these sources of bias are closely connected: institutional bias informs patterns learned by AI systems, and in turn AI systems can encode and scale these biases across large populations, function at speeds driven by real-time pressures, and obscure decision-making behind technical complexity.[13] Collectively, the features can normalize biased outcomes across various security institutions.

How security institutions use AI systems

Security institutions use AI across a range of functions, including risk identification, information processing, surveillance, operational planning, and decision support. These applications vary across bodies within the security sector, but they share a common purpose: AI systems shape how institutions identify, prioritize, and respond to perceived threats.

These systems can improve efficiency by processing large amounts of information, identifying patterns, and supporting faster operational decisions. However, AI systems do not operate independently of the institutions that deploy them.

Their outputs are shaped by the quality and representativeness of the data on which they are trained, the assumptions embedded in their design, and the policies governing how they are used.

The next subsection examines the principal ways in which security institutions use AI systems, and highlights the key governance challenges associated with each application.

Risk identification and prioritization

Security institutions use risk assessment systems to determine whether a person, place, behaviour, or event should receive additional scrutiny. Specifically, these systems predict crime risk, prioritize border screening, assess programme beneficiaries for fraud, identify networks of concern in intelligence data, and provide early risk warning – including, for example, the potential for violence.

While especially useful when viewed through a lens of efficiency, these systems raise concerns when risk categories are based on historical enforcement patterns, proxy variables, or incomplete data. This can allow emerging actors or less traditional threats to fly under the radar. It can also result in known-pattern bias, where systems assess individuals and groups based on how closely they resemble previous intelligence targets.

Predictive policing systems determine where and when police attention should be focused. They rely on historical incident and crime data, including arrests and reported crimes. Increased policing generates more recorded incidents. Historically overpoliced areas are then flagged by AI systems as high risk, resulting in a self-reinforcing feedback loop that diverts resources from underreported areas, and thus reduces overall patrol effectiveness and further marginalizes certain communities.[14]

These examples illustrate why risk assessment systems require transparent criteria, regular auditing, continuous retraining, data drift detection, and meaningful human oversight.

Information processing and intelligence analysis

Security institutions use AI to process large volumes of information. These systems process satellite imagery, open-source information, communications intercepts, biometric data, incident reports, and intelligence repositories. Information relevant to women and local communities is frequently collected through informal networks, qualitative reporting, and civil society channels that may be less visible within conventional intelligence data sets.

AI systems used for information processing can help analysts identify anomalous patterns based on past examples, but they can also shape which threats are visible. If the underlying data is skewed based on incomplete reports, biased surveillance practices, or other variables, the system may artificially amplify certain risks. Because AI prioritizes certain signals, it also determines which signals are deprioritized.

Governance should therefore ensure that AI complements, rather than replaces, other sources or methods of processing intelligence.

Surveillance, monitoring, and identification

AI systems enhance real-time surveillance capabilities. Facial recognition systems, gait analysis software, object detection systems, behavioural analytics, and biometric matching tools enable institutions to identify or monitor individuals. Patterns of surveillance may also affect women and girls, LGBTQI+ populations, and political actors differently.

Facial recognition tools alone are used by government agencies in at least 75 countries, across both democratic and authoritarian contexts.[15] Deployment is increasing across both democratic and authoritarian contexts through smart city platforms, crime prevention applications, and predictive policing systems. While these tools increase state visibility, they also raise risks of oversurveillance of marginalized communities and political actors.

EXAMPLE 2

AI-enabled surveillance and political control – Smart Cities (Uganda)

AI-enabled surveillance systems – often deployed through 'smart city' initiatives – integrate facial recognition, behavioural analytics, and centralized monitoring platforms. These systems are intended to improve public safety by enhancing situational awareness and enabling faster response to incidents.

However, in some contexts they have been used to monitor political opposition groups, journalists, and marginalized people. For example, reporting by Human Rights Watch indicates that surveillance technologies in Uganda were used to track opposition figures during elections.[16]

Even where systems are technically accurate, bias can arise through decisions about where, how, and against whom they are deployed. In these cases, AI systems may be used in ways that disproportionately target certain populations or reinforce existing power asymmetries.

Beyond improving algorithmic performance, effective governance extends into ensuring that the institutional use of AI systems is lawful, proportionate, and subject to oversight.

Otherwise, institutional legitimacy becomes affected, by way of:

- undermined accountability if surveillance practices are opaque;
- eroded public trust if used disproportionately against specific groups;
- distorted role of security institutions, shifting from protection to control.

Decision support in high-risk settings

AI systems inform decisions that may place individuals at risk. These include decisions that concern detention, border exclusion, targeting, and the use of force. Security risks and protection needs are experienced differently across populations, and AI-supported decisions may reinforce these differences if such perspectives are absent from system design.

Through an SSG/R lens, this links to concerns about international humanitarian law, including the principles of distinction and proportionality; responsibility and accountability for targeting decisions; and legitimacy of military efforts. Hence growing attention is paid to how existing international legal frameworks apply to AI-enabled military systems. The International Committee of the Red Cross (ICRC) outlined procedures for legal reviews of new weapons, means, and methods of warfare. It encourages states to determine whether new weapons or methods of war they possess, or intend to possess, comply with international law obligations, in advance of design, development or procurement.[17]

For example, states should assess where autonomous AI systems may be used for decisions that require human judgement, such as distinction, proportionality, or precaution in attack. These risks underscore the importance of maintaining meaningful human judgement over decisions involving the use of force.

Resource allocation and planning operations

Prediction systems inform where staff and resources are deployed. For example, in military and intelligence contexts, AI-supported systems may prioritize intelligence, surveillance, and reconnaissance (ISR) coverage based on where activity has been observed in the past. This can result in disproportionate monitoring of known areas of operation, while reducing efforts in regions where threats are emerging but underreported.

Similar issues can occur with AI-enabled imagery analysis. These systems may be biased towards identifying activity that matches previous threat signatures, while failing to detect civilian or anomalous activity. Such prioritization challenges also affect systems designed to anticipate future instability or violence.

Regular reviews of deployment patterns can help prevent AI from reinforcing existing operational blind spots. This is both a technical matter and an obligation of governance – at the very least to ensure that such reviews take place.

Early warning and risk prediction

AI systems can assist with identifying potential risks. However, data limitations can exclude entire categories of risk, like gender-based violence. In practice, AI systems shape not only decisions made but also what is seen, counted, and prioritized.Threats affecting women and other at-risk groups – including attacks on women leaders, mobility restrictions, and conflict-related sexual violence – are often underrepresented in conventional data sets.

Where bias enters AI systems in the security sector

Bias can enter AI systems in security institutions at multiple stages, alongside the decisions and oversight – or lack of oversight – undertaken by governance. Understanding how bias manifests at each stage is essential for mitigation.[18] Three technical mechanisms of introducing bias in AI-enabled systems are important for practitioners to understand *(Figure 4)*.

For security practitioners, the important thing is that bias does not arise only after a model has been built. In practice, bias can enter at six key points, each corresponding to a decision that institutions make throughout the security decision cycle.

Problem definition and institutional priorities

Problem definition is the stage at which an institution determines what security problem an AI system will address. *Problem formulation is not neutral.* Rather, it reflects policy priorities, institutional incentives, and underlying power structures. Two institutions using the same data and the same machine-learning techniques could produce very different – and potentially more or less equitable – outcomes simply because they defined the objective differently. How the problem is framed determines the decisions an AI system will make.

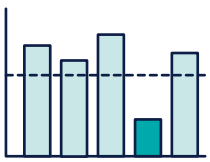
For example, framing a problem as 'identifying high-risk communities' produces a different logic than defining it

Figure 4. Technical mechanisms of introduction of bias in AI-enabled systems.



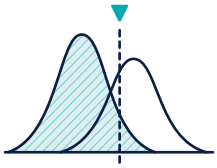
Feature selection

Choice of input variables can introduce proxy discrimination even when sensitive attributes are excluded



Model evaluation

Aggregate accuracy metrics can mask severe performance disparities for specific subgroups



Performance thresholds

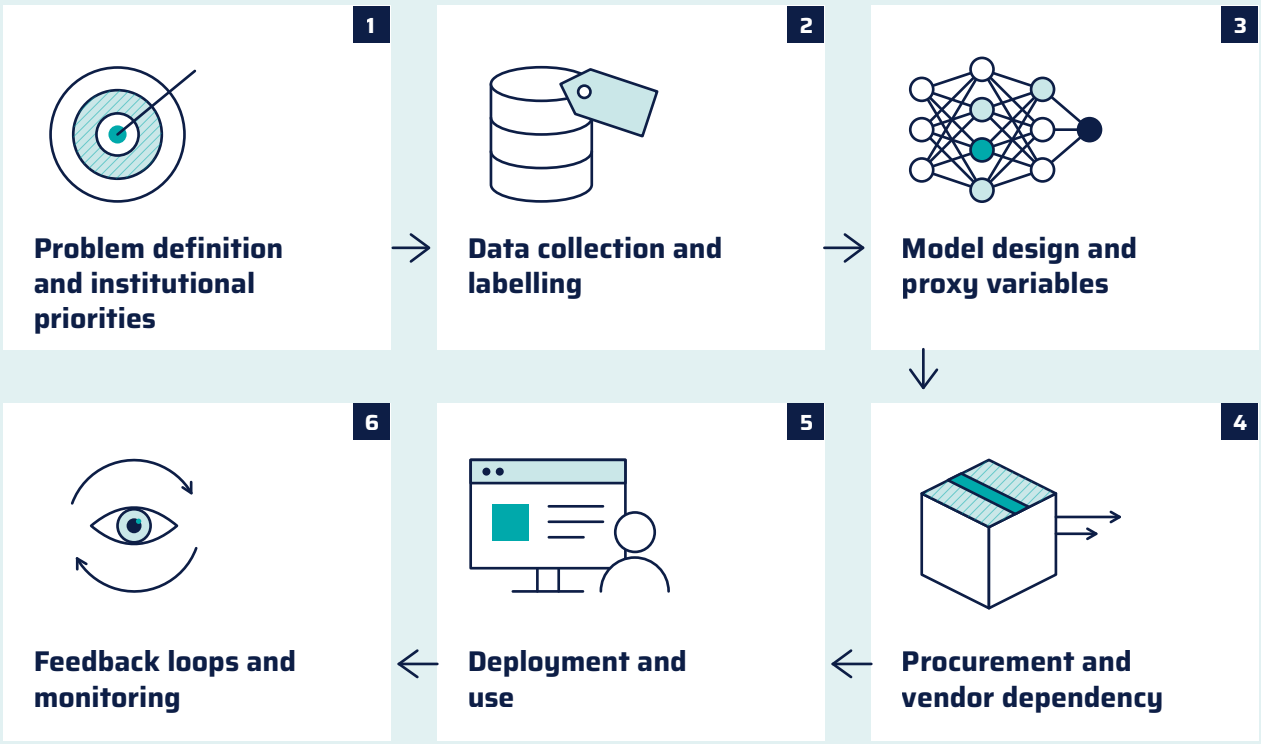
Score or confidence level at which a system triggers an action is set without consideration of differential error rates across demographic groups

as ‘identifying where protection needs are greatest’. Here, the framing of the problem is itself a source of bias because it shapes what the system is designed to optimize. The distinction is that ‘identifying high-risk communities’ frames communities as potential sources of risk or threat, which will lead the AI system to look for indicators associated with danger, crime, or instability. This may result in increased surveillance, policing, or security interventions directed at those communities. In contrast, ‘identifying where protection needs are greatest’ frames communities as people who may be vulnerable and in need of assistance. Following this prompt, the AI system will prioritize indicators of vulnerability, exposure to harm, or lack of resources, leading to humanitarian support or protective measures. Both approaches might analyse similar data, but they answer fundamentally different questions: ‘Who poses the greatest risk? versus ‘Who is the most vulnerable?’, respectively. As a result, bias can be embedded before new data is collected. Once institutions define what problem AI systems are intended to solve, those assumptions shape what new data is collected and how it is interpreted.

Data collection and labelling

Bias is further introduced during data collection, where training data sets often reflect historical inequalities and prior patterns of enforcement. Similarly to security decisions, security data is not neutral. Data sources such as arrest

Figure 5. Points at which biases enter AI-enabled systems.



records, incident reports, intelligence files, border refusals, and watchlists are generated through institutional practices. Bias is introduced because the data itself reflects human and institutional decisions: it records what security institutions chose to observe, not necessarily where actual risk exists. If surveillance is concentrated on certain communities or population segments, they will be overrepresented in data sets and AI may conclude that they are more risk-prone, when in reality they are simply monitored more intensely.

Similarly, if certain forms of violations are underreported, they will not appear as significant trends. The data reflects patterns of enforcement, not necessarily patterns of risk or threat. Because machine learning identifies statistical patterns, it may learn historical patterns as if they represent a genuine risk.

Bias also enters through labelling processes, where human judgement determines what constitutes suspicious behaviour, extremism, threat relevance, or other categories. Security data sets are frequently based on outputs from institutional practices like surveillance, reporting, and enforcement decisions. As a result, data sets may reflect patterns of who and what has been surveilled, rather than underlying risk. These risks become particularly visible when systems are applied more broadly across heterogeneous populations.

EXAMPLE 3

Biometric systems and risks of exclusion – migration in Europe

Biometric systems are widely used in migration and asylum management across Europe, including systems such as Eurodac, a fingerprint database for asylum seekers and irregular migrants.[19] Beyond fraud risks commonly associated with biometric identifiers, there are concerns about error rates when biological characteristics do not align with system design assumptions. Environmental, occupational, or physiological differences can result in poorer quality data from populations in certain regions. This leads to slower processing for asylum applicants, misidentification of individuals, and wrongful denial of services or protection. While not necessarily purposeful or due to malice, AI-enabled systems must be designed for real-world conditions – encompassing assumptions about data integrity made during the technical systems' development.

To avoid an AI system reproducing and even reinforcing existing institutional biases, practitioners must critically examine how the data was generated and labelled, not just how the model is trained.

Model design and proxy variables

Model design introduces another important source of bias. Developers decide which variables the model should use, which are considered important, and what outcomes the system should optimize. These choices are not purely technical, but also reflect assumptions about what constitutes a risk or threat. Even when sensitive attributes (sex, ethnicity, religion, etc.) are excluded, the model may rely on proxy variables – such as location, language, travel history, family network, or prior police contact – that correlate with these attributes and may indirectly shape assumptions on identity or social status. As a consequence, the model may produce discriminatory outcomes despite never explicitly using protected characteristics. In security settings this may influence who is stopped and searched, referred for investigation, detained, or subjected to monitoring.

Because proxy discrimination originates during model development, it should be identified and addressed before deployment rather than only through post-deployment monitoring. Two complementary approaches are commonly used to that end.

- **Disparate impact analysis** can identify whether a nominally neutral variable produces systematic outcome differences across protected groups even in the absence of intentional discrimination.
- **Counterfactual fairness** *analysis* tests whether an individual's predicted outcome would change if their protected characteristics were different while all other relevant attributes remained the same. If the outcome changes solely because of the protected characteristics, the model may be unfair.

Procurement and vendor dependency

Many institutions procure AI tools from external vendors rather than developing them internally. Bias can be introduced through procurement, particularly when organizations purchase AI systems that they cannot fully scrutinize in terms of how data sources were chosen, how the model was trained, what testing has been done, or what limitations exist. Institutional context plays a significant role in shaping technology choices, for example in policing.[20]

> Institutions are often dependent on vendors for technical information about how a system operates.

Meanwhile, vendors may claim commercial confidentiality or intellectual property protections to limit disclosure when asked about system source code, model architecture, or algorithmic logic. This asymmetry of information may create accountability gaps, making it difficult for security institutions to assess system performance, identify potential biases, or determine whether the model performs consistently across groups or contexts.

Even where procurement processes are solid and due diligence is conducted, additional risks emerge during

implementation. As personnel begin incorporating AI outputs into operational decision-making, there is a risk that system recommendations may be afforded undue authority or treated as inherently objective. Without appropriate governance, training, and oversight, practitioners may rely on outputs without fully understanding their limitations, uncertainty, or potential biases, allowing errors or inequitable outcomes to be enforced in practice.

Deployment and use

Deployment is the stage when personnel begin using systems in real-world decision-making. At this point bias may arise not only from the model itself, but also from the way personnel interpret and rely on its outputs. In high-stakes security environments, individuals have a tendency to over-rely on automated outputs while discounting or overlooking contradictory information, a phenomenon known as automation bias. This is especially strong when decisions are made under time pressure and a perceived risk of severe consequences.[21] In these circumstances, personnel may defer to algorithmic results because they are viewed as faster or more objective than human judgement.

The risk is that AI-generated recommendations will be accepted without sufficient scrutiny, even when other evidence points to a different conclusion. Automation bias can amplify underlying biases in the model, embedding them into operational decision-making.

EXAMPLE 4

Automation bias in intelligence fusion centres (NATO)

AI systems filter and prioritize information in multinational security settings, like the intelligence fusion centres that NATO partners operate. These systems help analysts by surfacing signals from the many information streams that such centres must monitor, potentially allowing for quicker decisions. Operators can become reliant on machines, especially when the machine is viewed as an authority figure and the operator is under time pressure.

This can lead operators to place too much weight on signals identified by machines, disregard signals not identified by machines or the context around them, and critically analyse machine output less than human-generated information.

Feedback loops and monitoring

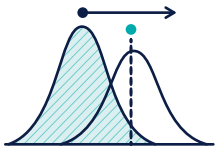
Monitoring and evaluation are the final stage at which bias can emerge or intensify throughout the AI life cycle. AI systems can reinforce their own assumptions over time. Without appropriate oversight, this can create feedback loops that reinforce and amplify bias over time. Decisions informed by AI outputs shape future data collection, which in turn retrains and reinforces the model.

The predictive policing example cited earlier demonstrates the phenomenon, whereby patrols are directed to specific neighbourhoods based on historical data. More police presence leads to more recorded incidents in those areas – not necessarily because crime has increased, but because more activity is being observed and documented. These records are incorporated into training data, reinforcing the AI system's original assumption and creating a self-reinforcing cycle.[22] Similar feedback loops can occur with intelligence prioritization, border screening, and military surveillance systems where AI-assisted decisions influence future data collection.

Detecting feedback loop effects requires specific monitoring techniques *(Figure 6)* beyond standard accuracy testing, to determine whether system performance, fairness, and operational impact are changing over time.

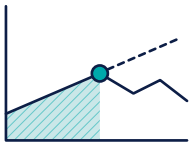
Despite the importance of continuous oversight, monitoring and auditing

Figure 6. Monitoring techniques for feedback loops.



Distribution shift monitoring

Tracks whether the statistical properties of incoming data are changing over time in ways that suggest the model is altering the environment it observes



Concept drift testing

Detects whether the relationship between inputs and correct outputs is changing, which may indicate that historical patterns are no longer valid



Data poisoning detection

Identifies whether outputs are being systematically fed back into training data in ways that entrench existing biases

mechanisms are often underdeveloped. Without continuous evaluation, biases can evolve and persist unnoticed over time. Many AI systems, particularly those based on complex machine-learning models, operate as 'black boxes', making it difficult for practitioners to understand how decisions are made or identify sources of bias.[23] For these reasons, scholars argue that AI governance should be viewed as an ongoing process, and emphasize the importance of ongoing auditing and impact assessment to ensure that systems remain aligned with fairness and accountability objectives.[24]

Governance and oversight

Because bias can emerge at every stage of the AI life cycle, from data collection and model development to deployment and operational use, governance and oversight are essential to ensure that AI systems remain lawful, accountable, and subject to meaningful human review. A growing set of policy frameworks seeks to address these risks, but coverage and implementation remain uneven.

International frameworks

The UNESCO *Recommendation on the Ethics of Artificial Intelligence* (2021) calls for transparency, accountability, and human oversight, including in security applications. It emphasizes structured testing to 'identify failure modes, risks of misuse, and unintended consequences' before deployment.[25] Similarly, the OECD *AI Principles* (2019), adopted by over 40 countries, emphasize fairness, robustness, and accountability. Both rely on voluntary implementation.[26]

Binding regulation

Binding regulatory approaches are emerging at the regional level. The European Union Artificial Intelligence Act (2024) is the most comprehensive regulatory framework to date, covering systems used in law enforcement and border control by classifying them as high risk and imposing requirements for risk management, data governance, transparency, and human oversight.

However, practitioners in security institutions must understand a critical limitation: Article 2(3) of the EU AI Act explicitly excludes AI systems used exclusively for national security purposes from the entire regulation – not merely from specific provisions. This means that military targeting, intelligence analysis, national security surveillance, and most of the use cases discussed in this guidebook fall entirely outside the EU AI Act's scope for EU member states.

Practitioners should not assume that their institution's AI use is covered by this regulation. The exemption reflects a deliberate political choice, but it creates a significant governance gap in precisely the domains where the risks documented in this guidebook are most acute.

The Council of Europe's *Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law* (2024) is the first legally binding international treaty on AI (pending ratification by signatories), and recognizes that the misuse of AI technologies poses serious threats.[27]

Standards or guidelines

Most countries lack binding regulation on AI in the security sectors, but there are examples of standards or guidelines. In the USA, governance related to AI remains fragmented and sector-specific, without binding legislation. The National Institute of Standards and Technology AI Risk Management Framework (2023) provides guidance on managing risks of bias and lack of accountability, but is voluntary and uptake depends on institutional priorities.

The 2023 Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence introduced requirements for federal agencies to improve AI oversight, including safety testing and reporting. However, as an executive action, it remains politically contingent and subject to revision or repeal. What is more, these measures do not fully extend to classified or national security systems, where oversight is more limited.

Despite their differences in legal status and scope, these frameworks reveal a common pattern: areas where the use of AI carries the highest risks – such as intelligence, counterterrorism, and military operations – are often the least regulated and least transparent. This creates a governance asymmetry, in that oversight is strongest in civilian and commercial domains, but weakest in the security contexts where risks to rights, safety, and accountability are most acute.

Taken together, these gaps show that bias in security sector AI systems remains insufficiently managed. While progress has been made in establishing principles and regulatory frameworks, current approaches are unlikely to address fully the scale and stakes of AI-driven decision-making in security contexts. Until binding governance frameworks become more comprehensive, security institutions cannot rely solely on external regulation to manage AI-related risks. Instead, they should establish internal governance mechanisms, including legal review, documented risk assessments, independent auditing, meaningful human oversight, and regular monitoring throughout the AI life cycle. Without such shifts, AI systems risk formalizing and amplifying existing inequities.

In addition to alignment with regulatory frameworks, security sector institutions should carry out legal reviews at multiple stages: before they procure AI systems, before they deploy them, and whenever AI systems are modified or deployed in new ways or new contexts.

There are several key questions to cover:

- Is there clarity in the legal basis for the security institution to use AI for the intended purpose(s)?
- Is the use of AI necessary for the purposes, and is the risk of impact on individuals proportionate? Are fundamental risks safeguarded when using a lens of domestic and international law (including human rights and humanitarian law)?
- Are personal data collection, analysis, and retention by the AI-enabled tools compliant with the laws on data protection?
- Is accountability for compliance and outcomes clearly defined?
- Is required human oversight maintained?
- Is the use of AI documented and traceable?

Differentiated impact of AI-supported security decision-making

AI-supported security decisions affect populations differently. Groups experience distinct benefits and risks based on factors including sex, age, race, ethnicity, language, disability, migration status, socio-economic status, geography, political affiliation, and perceived association with armed actors.

> These differences are especially important in conflict and crisis settings, where security risks are shaped by gender norms and unequal access to protection.

Gender

The WPS agenda highlights how gender norms shape patterns of violence, access to security, and institutional responses.[28] However, these dynamics are often insufficiently reflected in AI systems used to support security decision-making. Many security data sets prioritize visible or formally reported incidents, such as fatalities, arrests, or police reports, while overlooking forms of harm that disproportionately affect women, including sexual violence, harassment, movement restrictions, coercion, and attacks on women community leaders.

As a result, AI systems trained on these data sets may underestimate or fail to detect threats that fall outside conventional reporting channels. This is particularly problematic for early warning systems, where gender-sensitive indicators may emerge well before violence becomes visible through traditional security metrics.

EXAMPLE 5

Gender bias in AI-enabled early warning systems (Columbia)

Colombian early warning systems have traditionally focused on monitoring kinetic violence and high-casualty attacks. While useful, many of these systems have missed threats to women community leaders, who often experience intimidation, coercion, and high-risk targeting by armed actors prior to physical attack.[29]

This results in systems indicating high-risk periods only immediately before or during visible violence. Early warning signals are often detected too late. However, many forms of violence – particularly those affecting women and marginalized groups – can present early warning signals through escalating threats, restricted movement, or localized intimidation before becoming fatal. The result of missing such signals is delayed or misdirected response.

This example illustrates that bias arises not only from inaccurate data but also from what is absent from security data sets. AI systems can unintentionally exclude risks that are not well represented in their underlying data. This in turn challenges how institutions define effective data and risk. If acts of violence or types of survivors are not captured in data sets, they will be left out of critical decision-making processes. For SSG/R this has two key implications.

- **Effectiveness:** AI systems may miss early warning indicators, delaying preventive action and reducing the ability of security institutions to identify emerging threats.

- **Responsiveness:** Failure to recognize gender-specific risks can result in unequal protection for women and other marginalized groups whose security concerns are less visible in conventional data.

Intersectionality

While gender is a critical dimension of AI bias, it rarely operates in isolation. Individuals often exhibit multiple identity factors simultaneously (sex, ethnicity, migration status, disability, and language). The interaction of these factors can shape both their exposure to security risks and the likelihood that AI systems recognize those risks. For example, women community leaders, displaced young men at a border crossing, persons with disabilities at checkpoints, and linguistic minorities within intelligence data sets each face distinct risks and varying levels of visibility or exclusion.

AI systems that fail to account for these dynamics risk producing incomplete or distorted analyses, particularly where certain groups are underrepresented, misclassified, or not captured in underlying data. For security practitioners testing systems for bias, outputs should be evaluated across intersecting identity factors, rather than in isolation. Intersectionality is therefore critical for AI systems used to monitor, assess, and identify emerging threats.

Intersectionality creates a distinct bias gap: systems might work well for certain forms of abuse but fail where forms of harm intersect across multiple identities, including sex, ethnicity, religion, political affiliation, or other factors. In security contexts, this can mean that the people who face the most serious combined risks online will have the least security protection from AI-enabled monitoring tools.

This has direct implications for security institutions. Online threats can translate into real-world violence, especially against women politicians, activists, and community leaders. Failure to detect these threats can undermine SSG/R principles, specifically:

- *effectiveness:* they miss critical early warning signals;

- *responsiveness:* they fail to protect those most at risk;

- *accountability:* they obscure blind spots.

The implication for governance is ensuring that bias testing evaluates AI system performance across intersecting identity characteristics rather than single demographic variables alone. Procurement, monitoring, and auditing processes should require evidence that AI systems performed consistently across relevant population groups.

EXAMPLE 6

Intersectional bias in AI detection of online threats

Machine-learning tools intended to detect harmful online content have become more integrated into early warning systems and security monitoring. However, they often cannot identify threats across intersecting factors. This is true especially if abuse is specific to certain contexts or uses disguised language.

Automated moderation systems are less likely to detect and remove harassment and threatening content directed against women who are also part of marginalized racial, ethnic, or religious groups. Harassment targeting women politicians or activists, for instance, may use language that is both gendered and racialized or tied to religious or cultural norms that are not present in moderation systems' training data.

Consequences of bias in AI systems

Bias in AI-supported security systems can have cascading consequences for effectiveness and decision-making, legal aspects, institutional aspects, and matters related to conflict and stability. These four aspects are described in the following subsections. The lessons for governance, with a lens of SSG/R, are noted for each one.

While these four dimensions can individually help better analyse the relevant impacts, understanding the connections is equally important *(Figure 7)*.

Effectiveness and decision-making consequences

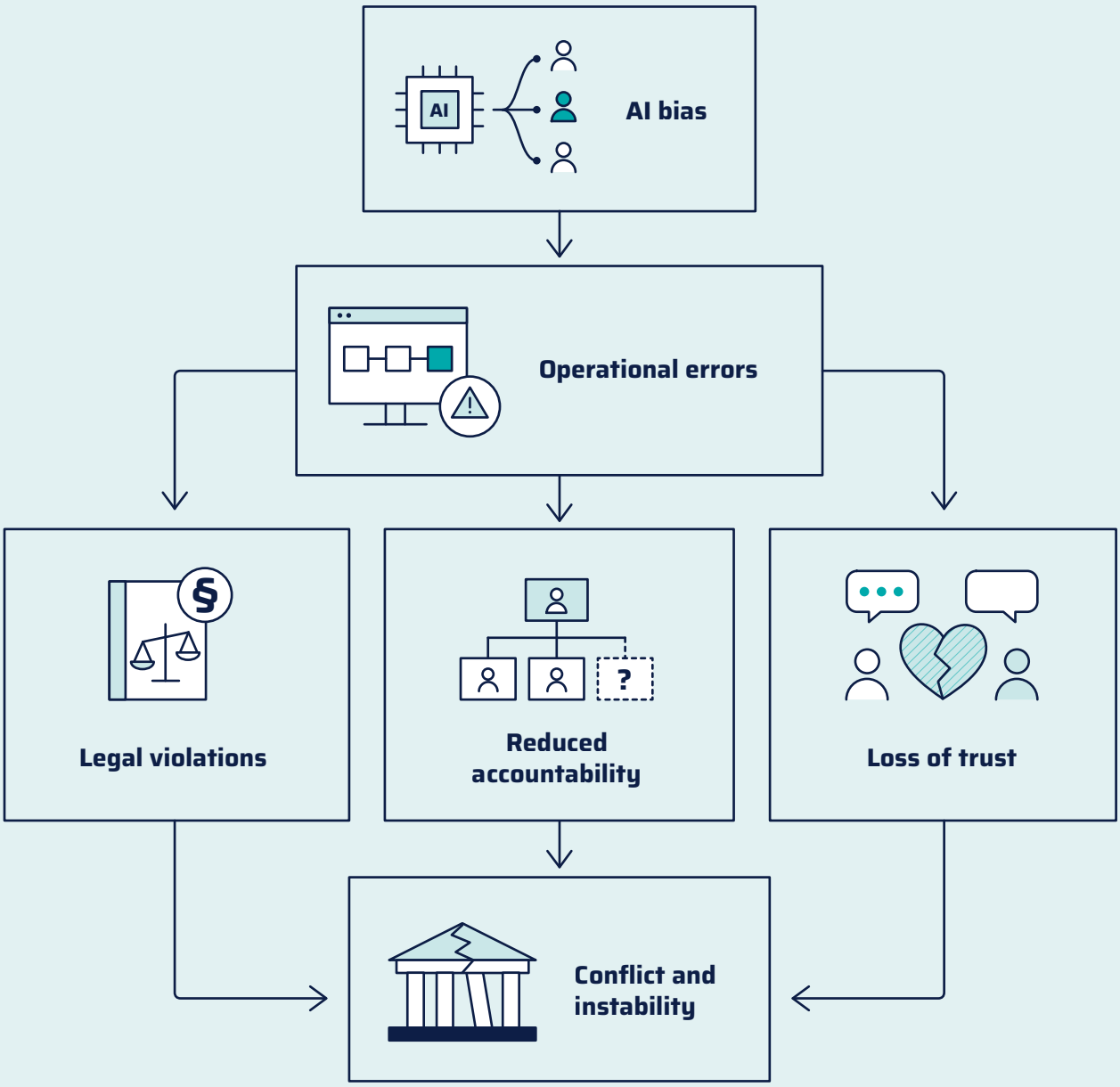
Bias shapes how security actors perceive threats, allocate resources, and make day-to-day decisions. This can manifest as:

- misidentifications and false positives;
- inaccurate risk scores;
- resource allocation based on historically over-observed individuals, locations, or networks;
- failure to detect where protection is most needed;
- false negatives that allow threats to go undetected;
- distorted signal prioritization in intelligence or investigations.

In law enforcement, these dynamics can lead to overpolicing of certain populations or areas. In intelligence analysis, they can distort threat prioritization. In border security, they can lead to unequal levels of scrutiny. In military operations, they can influence targeting decisions and situational awareness.

Wrong labelling, inaccurate risk-based searches, overpolicing, or activities based on incomplete information are not only fairness concerns. They are also effectiveness failures. To mitigate these risks, institutions should assess whether

Figure 7. Connections among consequences of AI biases in the security sector.



AI-supported decisions meet effectiveness standards. Does the system identify the correct individuals? Does it protect intended populations? Performance testing should include not only rates of false positives, but also false negatives, variations across demographics, geographic variation, and effects downstream of initial decisions.

Governance must therefore be attentive towards misused resources, overlooked risks, and potential civilian harm that results from bias in AI systems, e.g. skewed threat identification. Given that these failures also become legal risks when biased outputs affect rights, liberty, or use of force, governance attention should be early and continuous.

EXAMPLE 7

AI-supported ISR tasking bias and blind spots (military surveillance)

A military intelligence unit adopted an AI-supported intelligence, surveillance, and reconnaissance (ISR) tasking tool to help analysts identify priority areas for aerial surveillance coverage within a contested environment. The tool used previous incident reporting, past ISR collection trends, signals intelligence (SIGINT), and threat reporting to suggest areas of interest for analysts and drone operators to focus surveillance collection efforts.[30] Personnel adopted the tool to cope with high tempos and information flows, and to determine where to task limited ISR resources.

Users began to focus more of their ISR collection on areas that were repeatedly identified by the system as priority locations, rather than where adversaries were changing their behaviour. This created blind spots. Adversaries changed travel routes to take place in areas where there was less surveillance coverage, and became aware that analysts were relying heavily on machine-generated prioritization. This is not a technological failure: it occurred because insufficient safeguards were in place to govern how users validated and relied upon automation-supported analytics.

Legal consequences: Compliance, visibility, and contestability

AI bias creates legal and compliance risks when outputs lead to unequal treatment that violates international human rights law. For example, biased outcomes in law enforcement or border contexts may violate rights related to non-discrimination, equality before the law, liberty, privacy, and due process. In armed conflict, biased AI-enabled targeting or decision-support tools may violate international humanitarian law (IHL) principles of distinction, proportionality, and precaution in attack.[31]

These risks increase where AI-informed decisions cannot be explained or challenged by affected people. If a person is denied entry at a border, placed on a watchlist, detained, or investigated based partly or entirely on an AI output, there needs to be a way to document how that output was used and to contest the decision. If affected individuals cannot realistically exercise these rights, they remain theoretical.

Biased outcomes expose security providers to risks of breaching international human rights and humanitarian law, particularly where liberty, movement, or use of force is affected. These risks become especially visible in systems that rely on predictive risk scoring. There are several impacts on SSG/R principles.

- **Rule of law:** Failure to provide transparency compromises due process.
- **Accountability:** Ambiguity around responsibility between government agencies hinders accountability.
- **Responsiveness:** Overpolicing of specific groups perpetuates systemic injustice.
- **Effectiveness:** Public perception of unfairness delegitimizes institutions.

Governance has accountability to ensure that legal dimensions are reviewed, and that compliance is assured, when an AI tool is deployed by security sector institutions. Effectiveness alone is insufficient justification to deploy a tool.

EXAMPLE 8

Real-time facial recognition and legal safeguards (United Kingdom)

Police departments in the UK utilized live facial recognition technology at public gatherings and other crowded public venues to attempt to match individuals – efficiently – to people on police watchlists. Facial recognition technology was found to have technical accuracy issues through third-party audits and legal analysis. False positives needed to be assessed. Additional critiques related to privacy, proportionality, transparency, and bias.

Litigation was brought arguing that the implementation guidelines for facial recognition technology did not adequately outline specific legal safeguards: circumstances under which facial recognition technology would be used; how watchlists would be developed; what oversight would be available to prevent unreasonable or disproportionate use; and what mechanisms were available for citizens to dispute or review use of the technology. The UK Court of Appeal in Bridges v South Wales Police ruled in favour of the litigants.[32]

Rules and regulations had not kept up with technology use, without clear guidelines around what standards needed to be met before using the technology; what level of inaccuracies were appropriate; how proportionality was determined; independent oversight; and documentation, auditing, or appealing use of the technology.

Institutional consequences: Accountability and trust

Biased outcomes also affect security providers' internal governance and relationships with civilian populations. When bias in these systems causes harm, determining responsibility for harm becomes difficult – for example, whether it was the state or the vendor that purchased the system – raising challenges for oversight.[33]

Responsibility is distributed across developers, vendors, agencies, analysts, and frontline personnel. When failures occur, responsibility can be diffused across actors, complicating accountability. Vendors may blame the agency for misuse; agencies may blame operators for relying on technology too much; and operators may blame developers for poorly designed outputs.

AI bias also damages trust. Biased outcomes against certain populations affect security providers' relationships with civilians. When security providers disproportionately target or exclude certain groups, these providers may be perceived as biased or discriminatory. Over time, this can cause civilians to become unwilling to cooperate with security providers. This can harm public information-sharing, community policing, early warning, and stabilization and conflict prevention.

The lesson for governance is not to think only about the near-term frontline impacts of AI, but also the long-term and broader impacts on legitimacy. When legitimacy is weakened, effectiveness is put in jeopardy.

EXAMPLE 9

Automated border screening and accountability gaps (European migration)

An international border agency implemented an AI-enabled screening and risk-assessment platform to support migration processing and traveller prioritization at border checkpoints and processing centres. The platform scored and ranked individuals using biometric matches, travel records, identity documents, and behavioural scores to identify persons for secondary screening, intensified interviews, or detention review.

Authorities hoped the system would improve throughput, effectiveness, and efficiency. The organization delegated day-to-day responsibility for the platform to teams including border management officials, migration and asylum services, external technology vendors, and data management and analytics personnel. Over time, allegations surfaced that migrants from specific countries or regions were being disproportionately selected for additional screening, intensive questioning, or unwarranted detention. Individuals faced delays or additional screening

multiple times without justification. During times of high-volume arrivals, operators increasingly deferred to automated risk scores – with limited visibility into decision-making and no guidance on when to challenge or escalate inconsistent outputs.

Border management officials used the system daily but were not responsible for its design. Procurement offices oversaw vendor contracts but could not directly monitor performance. Vendors controlled elements of the model framework and training data used to generate scores, while regulatory teams lacked technical documentation to understand system outputs. Eventually, organizations struggled to answer basic questions about who was responsible for reviewing discriminatory impacts, authorizing the suspension or modification of the tool, and taking accountability for promoting transparency, fairness, and due process.

Frequent failures were only possible because responsibility for procurement, deployment, and oversight was segregated across a fragmented system – including a failure to provide officers with systematic and ongoing training about tools being used at the border. Clear lines of authority, accountability, and escalation were never formally established to define organizational responsibility for various phases of the system life cycle. Accountability gaps formed where institutions could identify problems but not easily remedy them.

Stability and conflict dynamics consequences

Biased outcomes can fuel grievances, create new insecurities, enable recruitment and misinformation by armed groups, and ultimately undermine stability. Over time, systems that treat people unfairly can reinforce drivers of conflict.

Bias can distort decision-making. If AI systems consistently prioritize the risks that are easiest to measure, while missing threats that are harder to quantify, decision-makers may come to see this output as an objective depiction of their operating environment. This concern applies to conflict prevention, where anticipating risks often requires reading weak signals before they escalate into violence. Over time, bias can become entrenched, distorting threat perceptions and fuelling instability. Adversaries can also exploit these dynamics.

Bias can play a role in how security providers think about and prioritize risks. If AI systems skew or oversimplify understandings of conflict or security risks, they could make institutions more or less likely to focus resources on mitigating them. Taken together, these consequences show that bias has the potential to undermine

security providers' longer-term objectives. AI outputs should inform – but not define – analysis. Technology cannot replace local knowledge, civil society engagement, gender analysis, or community validation.

Further to the impact on perceived legitimacy, the implications for stability and conflict dynamics are broader and longer-term areas of concern for governance. Indirect and unintended effects are equally important to operational objectives when deploying an AI-enabled tool in fragile or conflict settings.

EXAMPLE 10

AI-supported stabilization analysis and institutional legitimacy (Horn of Africa)

A multilateral stabilization force active in the Horn of Africa employed AI-supported analysis to inform conflict prevention and stabilization planning. The analysis tool used incident reporting, population displacement data, open-source reporting, and operation reporting to map where portions of the country were trending at-risk for instability and extremist violence. It aimed to help security institutions better allocate resources, prioritize engagement efforts, and improve early warning reporting in information-rich environments.

The AI tool worked as intended across several technical performance metrics. However, civil-military operations teams and implementing partner civil society organizations (CSOs) began hearing from local communities that security operations were starting to feel targeted or discriminated against. Specifically, certain communities felt that locations seeing an increase in surveillance and security forces were being labelled unsafe or extremist.

Eventually, these attitudes began to affect security actors' relationship with local communities; these communities were increasingly unwilling to share information, participate in civil-military engagement activities, or support stabilization efforts. This created a poor operating environment for security forces, as well as decreasing the quantity of information available to analysts to provide actionable information.

It resulted in a feedback loop. As local communities were less willing to share information with security actors, the information fed into the AI tool decreased in quality. This then decreased the quality of future analysis produced for planning. Eventually, the tool continued to meet analytic usefulness standards, but legitimacy and effectiveness were undermined.

EXAMPLE 11

AI bias and conflict dynamics (Myanmar)

Disinformation, hate speech, and harmful content targeting Myanmar's Rohingya population have been scaled by digital communication platforms and AI-enabled content amplification systems. While not created by security institutions, these dynamics shape the environments in which they operate. Algorithmic amplification of harmful signals worsened local dynamics, contributing to ethnic violence and the displacement of hundreds of thousands of civilians.

Detection failures – due to limited training data, language coverage, and contextual understanding – allowed harmful content to spread. These dynamics directly affect conflict and instability. When hateful or biased signals are amplified, they can contribute to violence, erode social cohesion, harden security dynamics, and complicate response efforts.

Summary: Implications for SSG/R

The objective of SSG/R is to ensure that security sector institutions are effective, accountable, transparent, responsive, and governed according to the rule of law.AI-supported decision-making can strengthen these objectives, but only when systems are designed, deployed, and governed in ways that prevent bias and preserve institutional accountability.

AI systems are embedded in core security functions, reshaping how risks are defined, decisions are made, and authority is exercised. While these technologies offer efficiency gains, they also introduce systemic bias that can undermine the foundational objectives of SSG/R. Bias is not an isolated technical issue, but a cross-cutting risk embedded throughout the AI life cycle, shaped by institutional priorities, data practices, model design, and patterns of use.

The implications span all pillars of SSG/R. Biased AI systems can:

- weaken effectiveness by distorting threat assessments and resource allocation;
- create legal risks and compliance failures by enabling discriminatory or unlawful outcomes;
- erode institutional accountability where responsibility for decisions is unclear or obscured;
- undermine responsiveness and effectiveness from reduced legitimacy.

Over time, these effects erode public trust, reduce cooperation with security providers, and exacerbate instability and conflict. Perceived unfairness can be as damaging as technical inaccuracy. Even where AI systems perform well according to conventional performance metrics, communities may lose confidence in security institutions if AI-supported decisions appear discriminatory, opaque, or inconsistent. Reduced trust can discourage communities from sharing information, reporting security incidents, participating in community policing initiatives, or cooperating with stabilization and conflict-prevention efforts.

Addressing AI bias is central to effective security governance. This requires embedding safeguards in each stage of use, aligning system design and deployment with rule-of-law principles, and reinforcing governance frameworks that ensure transparency, oversight, and accountability. Within an SSG/R approach, this means not only improving technical systems, but also strengthening institutional capacities, clarifying roles and responsibilities, and enabling meaningful oversight by civilian authorities and affected communities.

This institutional reform challenge spans governance arrangements, procurement processes, legal review, internal oversight, auditing, training, and civilian accountability mechanisms. While technical improvements remain important, the greatest determinants of fair, lawful, and effective outcomes are the institutional frameworks that govern how AI systems are procured, deployed, monitored, and used.

Recognizing where and how bias emerges provides the basis for identifying practical mitigation measures. The next section builds on this analysis by identifying practical measures that security institutions can adopt to mitigate these risks in practice. Table 1 recaps AI biases in relation to principles of SSG/R.

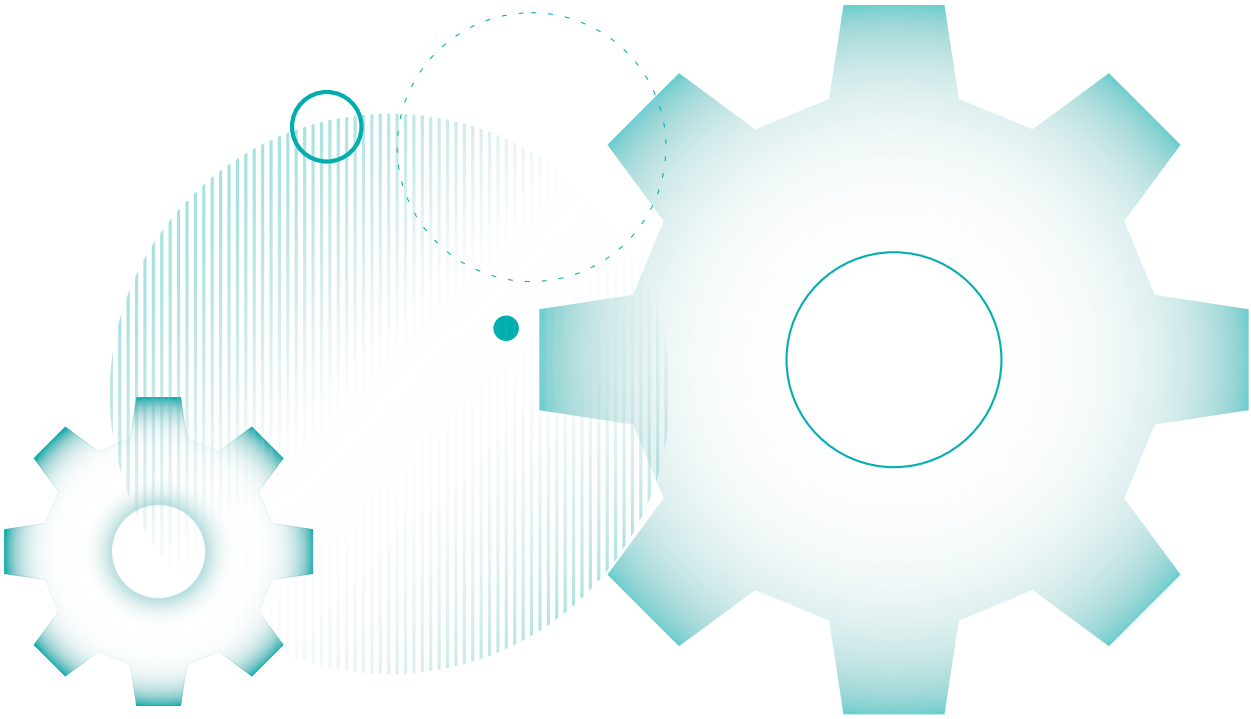


Table 1. AI biases in relation to principles of SSG/R.

SSG/R principle	Potential AI impact	Example	Governance action
Accountability	Responsibility for AI systems become fragmented (security institutions, vendors, developers, etc.)	Facial recognition mistakes cannot be attributed to the source or cause	Define key roles and metrics against which results are measured
Transparency	Automated outputs (recommendations, etc.) cannot be audited or challenged	Police receive risk scores but officers and/or affected population do not have the ability to understand them	Document AIenabled systems, and conduct and publish audits
Rule of law	Discriminatory outcomes could be unlawful by infringing on rights	Border control selects travellers for screening influenced by race, etc.	Conduct legal review before and during an AIenabled system's deployment
Participation	Impacted communities unable to contribute meaningfully towards an initiative's design	Surveillance system development excludes relevant CSOs	Purposefully include community stakeholders in procurement and design stages
Responsiveness	Needs of certain communities overlooked as data poorly represents their risks	Early warning systems in conflict or fragile settings	Include reporting by communities and diverse data sources (including gendersensitive)
Effectiveness	Threats misidentified	Military targeting	Test AIenabled systems for false positives and negatives, including human review
Efficiency	Resource misallocated (connected to ineffectiveness)	Predictive policing	Request justification beyond AI outputs for resource allocation decisions

SECTION II

Mitigating AI bias in security institutions



Section I introduced the risks that bias in AI systems poses to frontline decision-making, legal compliance, institutional accountability, public legitimacy, and the long-term effectiveness of security institutions. This section translates those risks into practical guidance for how security institutions can, and should, mitigate bias.

The section proceeds from a central premise: AI bias in security sector institutions is fundamentally a governance challenge rather than solely a technical one. Therefore, AI bias mitigation is not a one-time technical task; rather, it is an ongoing governance responsibility that spans system planning, procurement, deployment, use, monitoring, and review.

No single safeguard is sufficient to prevent AI bias. Effective mitigation depends on combining technical, procedural, legal, and institutional measures throughout institutional decision-making across the entire AI life cycle. Additionally, because security sector institutions make decisions affecting liberty, privacy, the use of force, and national security, governance measures that may be considered good practice in commercial settings become essential safeguards in security contexts.

The contents draw upon emerging international AI governance frameworks discussed in Section I, which emphasize human rights, accountability, transparency, human oversight, and risk management. These frameworks consistently highlight bias, and call for explicit safeguards to ensure compatibility with human rights and democratic governance.

AI model performance alone does not ensure safe outcomes. Lessons from AI development communities increasingly point to the importance of iterative testing, red-teaming, human feedback mechanisms, monitoring, and context-specific evaluation. Bias in AI is fundamentally a governance challenge with direct consequences for effectiveness, accountability, inclusivity, and adherence to the rule of law. Meeting this challenge requires more than high-level commitments. It requires integrating governance principles into day-to-day institutional practice.

The following subsections provide guidance on how to meet this challenge, organized around four core security sector functions: frontline use, legal compliance, institutional accountability, and management. Together, these domains provide a practical framework for embedding safeguards before AI systems are deployed, during operational use, and throughout ongoing oversight and institutional learning.

Consistent with the principles of SSG/R, the measures presented here aim to ensure that AI systems support – not undermine – effective, accountable, transparent, inclusive institutions governed by the rule of law.

Frontline use: Governing AI use in security institutions

AI systems should support, not replace, human professional judgement. Frontline personnel remain responsible for decisions taken using AI-supported recommendations. Effective governance therefore requires clear rules governing when AI may be used, how its outputs should be interpreted, and when human judgement must prevail.

The desired way of working is that AI-enabled systems are not necessarily avoided, but are governed and put into use with a view to strengthening effectiveness, accountability, legitimacy, and trust. Illustrative examples include the following.

- A police officer does not take action based solely on AI-generated risk scores, but sees the risk scores as inputs to prompt or reinforce human judgement.
- A border officer documents reasons for taking a decision based mainly on their judgement, and outlines the specific role played by the AI-generated flag either in alignment or in conflict with such judgement.
- A military analyst triangulates an AI-generated target list through other sources, particularly for lower-confidence-level recommendations.
- A military commander suspends or alters an AI tool's use after a review finds an escalation in false negatives or positives given a change in the context of the conflict.

Safeguards for high-risk uses

Where AI informs decisions that may lead to the deprivation of liberty (such as detention or surveillance), targeting, or other uses that restrain individuals' movement, access, or rights, additional safeguards should be adopted to account for the greater risks.

These may include:

- heightened thresholds for evidence;
- training personnel on how to recognize and raise suspected bias or system errors through formal channels;
- formal procedures for escalating these reports when they are received;
- supervisory approval requirements;
- additional documentation and reporting requirements.

SOPs, corroboration, and human judgement

Mitigating this risk requires organizations to put in place specific controls that define how outputs will be used. Outputs should inform decisions, rather than being treated as decisions. Thus institutions should clearly define how outputs will be interpreted, when corroboration is needed, and where human judgement needs to override automation.

This typically starts with clear standard operating procedures (SOPs) that outline:

- authorized and unauthorized uses;
- decision thresholds;
- requirements for independent corroboration;
- escalation and override policy;
- rationale for documentation and retention;
- procedures for reporting perceived bias or system malfunction (particularly on high-risk uses).

SOPs must be practical and enforceable. Human-in-the-loop (HITL) safeguards are ineffective when personnel do not have the time, authority, training, or access to knowledge to question outputs; in this situation it becomes human-in-name-only Controls are only effective if personnel understand both the capabilities and the limitations of the systems that they oversee.

Meaningful human review requires differentiated standards calibrated to tempo in practice. In real-time environments – where decisions are made under time pressure – minimum safeguards must be predefined and built into system design before deployment.

These include:

- mandatory uncertainty indicators;
- automatic escalation triggers when confidence thresholds are not met;
- hard stops that prevent the system from producing actionable outputs without a documented human authorization step;
- annual review without AI assistance, whereby security institutions compare unassisted analysis with AI-supported assessments to identify automation bias or dependence on AI recommendations.

In deliberate planning environments – where analysts have time to evaluate multiple information sources – more comprehensive review protocols are appropriate, including independent corroboration requirements and structured challenge mechanisms.

> Applying a single HITL standard across both environments produces meaningful review in deliberate contexts, but procedural formality in real-time ones.

Human decision-making and visibility

Given that even well-designed AI systems generate foreseeable failures when deployed without sufficient guardrails, training is needed. This is true especially in security contexts, where decisions have serious consequences, need to be made rapidly, are produced with imperfect information, and are enacted under specific rules of engagement and command authority.

Training can overcome the issues that arise from AI systems that determine what operators see, pay attention to, and interpret – particularly if AI outputs are treated as deterministic conclusions rather than probabilistic judgements.

Training should ensure personnel understand the proper use of AI systems and their strengths and limitations. End

users should be trained to understand that each system output has specific confidence levels, known error rates, and potential blind spots.[34] Otherwise, the time and effort saved by AI could be offset by reduced analytical scrutiny and dangerous dependence on system outputs.

Training should include specific instructions on automation bias, ensuring that personnel are able to do the following.

- Question machine-produced information independently, including a review of signals that fall outside of, or are assigned a lower priority by, the tool.
- Check system outputs against corroborating information.
- Identify system conclusions that may be skewed by gaps in the underlying data or systemic bias.
- Recognize instances where high frontline tempos may increase reliance on automation as a shortcut.

Tools are only as good as the rules governing their use. Effective automation use requires that personnel know when and how to trust system-generated analytics, and understand their tool's limitations. Countermeasures should be built into the technology itself, and also incorporated into SOPs and oversight.



IMPLEMENTATION CHECKLIST

Frontline use

Governance

- ☑ AI systems should be used only for clearly defined purposes and within authorized use parameters.
- ☑ SOPs should exist for each AI-enabled system, and should clearly define authorized uses, prohibited uses, escalation pathways, and verification requirements.
- ☑ Additional safeguards, approvals, or supervisory review mechanisms must exist for high-risk uses involving surveillance, detention, targeting, movement restrictions, or use of force.

Human oversight

- ☑ AI-supported outputs should be treated as advisory inputs rather than definitive conclusions.
- ☑ Independent corroboration must be required before AI-supported outputs are used in high-risk decisions.
- ☑ Users should be empowered to question, challenge, or override AI outputs when judgement requires.
- ☑ Uncertainty indicators, confidence scores, or known error rates should be visible to users where feasible.

Training

- ☑ Personnel must understand the purpose, limitations, and known risks associated with each AI system they use.
- ☑ Frontline personnel should receive regular training on automation bias, system limitations, and appropriate use of AI-supported recommendations.
- ☑ Personnel should know how to report suspected bias, anomalies, or system failures.

Monitoring

- ☑ Systems must be regularly tested for reliability, including false positives, false negatives, and performance across different contexts and demographic subgroups.
- ☑ Monitoring should explicitly distinguish between static performance testing (checking current accuracy) and dynamic feedback loop detection (checking whether the model is altering the data environment it depends on). Specific techniques include distribution shift monitoring, concept drift testing, and disaggregated subgroup performance evaluation. Thresholds for remediation – model retraining, suspension, or discontinuation – should be defined in advance and documented.
- ☑ Use of AI-supported decisions must be documented in a manner that supports supervisory review and accountability.

Legal: Ensure compliance and rights protection

Legal compliance is not a one-time review conducted before an AI system is deployed. It is a continuous governance function that ensures AI-supported activities remain lawful, proportionate, accountable, and compatible with human rights throughout the system life cycle. Because AI systems increasingly shape security decisions, legal safeguards must apply to both the technology itself and the institutional processes governing its use.

Legal review and human rights due diligence

AI systems should be subject to legal review throughout their lifetime. When deployed in security contexts, e.g. law enforcement, border management, intelligence, or military uses, these systems may implicate domestic law, international human rights law, data protection obligations, and in some cases IHL.

AI systems deployed for security purposes often raise legal obligations, and therefore require legal scrutiny despite rarely operating as the final decision-maker. Instead, they shape human decision-making processes by influencing what information is seen, which risks are prioritized, or which options appear available. For example, AI generates a risk score that informs the discretion of a border officer. An algorithm processes watchlists that determine who is subject to additional screening or search. An automated triage tool prioritizes intelligence reports and determines which signals analysts prioritize for review. A targeting support system feeds into staff planning decisions about future military operations.

In each of these examples, the AI system can determine the information considered, how questions are framed, and what choices appear available to decision-makers. This means legal accountability applies to both direct and indirect uses of AI.

Institutions should undertake human rights due diligence both before deploying a system and throughout its use. This includes analysing how the system may affect:

- privacy and data protection;
- liberty and due process;
- equality and non-discrimination, including freedom from algorithmic bias and disparate impact;
- freedom of movement, expression, and association;
- access to protection, documentation, and asylum.

Where AI systems are deployed for military purposes, legal review should examine how the intended or foreseeable use complies with IHL. This includes the principles of distinction, proportionality, and precautions in attack. States are required to comply with these obligations not only when developing and deploying weapons; humanitarian principles must also guide the use of AI-enabled decision-support systems that inform decisions about who to target, what areas to surveil, or how to deploy military force.[35]

States already have this obligation under international law. Article 36 of Additional Protocol I to the Geneva Conventions mandates that states assess how new weapons, means, and methods of warfare would be used and whether those uses would be lawful. These reviews must take place before weapons are adopted or deployed, not be introduced to justify

a weapon's use after it causes harms.[36] The same logic also applies to AI-enabled systems used for military purposes. The risk that AI-enabled technologies may be biased is particularly important here. If a machine-learning system is used to make targeting decisions or choose between different levels of civilian risk, bias could have lethal consequences.

Nonetheless, the applicability of Article 36 to AI-enabled decision-support systems – systems that recommend or inform targeting decisions but do not autonomously execute them – is currently contested. Irrespective, security sector institutions should adopt a precautionary approach whereby AI systems undergo a structured legal and operational review before use and periodically after use. States and legal scholars disagree on whether such systems constitute weapons or methods pertaining to warfare for the purposes of Article 36 review, with this ambiguity a gap in the current legal framework. UNIDIR's 2025 work on AI in military decision-making similarly highlights that review obligations for non-autonomous AI systems remain unresolved in international law.[37] Security practitioners should not assume that Article 36 reviews are automatically required for all AI military tools, nor that the absence of a formal review obligation means a system is lawfully appropriate for use.

Beyond questions of lawfulness, legal review should be used to help define acceptable limits of use. In some cases, planned uses of AI may be unlawful by design, or systematically incompatible with rights-based safeguards. For example, using an AI system that is opaque, not subject to independent testing, and lacks any mechanism for redress to make decisions about who can be detained, targeted, or granted access to protection would raise serious legal questions. In other cases, technical or frontline use safeguards may be required to ensure use is lawful throughout the system. These include human approval gates, auditability, and independent oversight. These risks become particularly visible in the deployment of AI-enabled surveillance systems.

While AI products can be effective and useful, high-performing systems may still create risk if governments do not provide a framework for appropriate use. Laws apply to how institutions use technology, not solely the technology itself. Organizations must also have policies and procedures that can ensure technology use is lawful, accountable, and proportionate.

Contestability, transparency, and remedy

Legal compliance depends not only on whether a system was lawfully approved before deployment, but also on whether its decisions remain reviewable, explicable, and contestable during operation. When institutions lack this ability, technical accuracy takes precedence over legitimacy, and the accountability demanded by the rule of law disappears.

Institutions must be able to review, explain, and contest AI-informed decisions. At a minimum, this requires the ability:

- to review decisions after the fact;
- to explain the logic supporting a decision to a sufficient degree;
- to challenge decisions and seek remedy for affected populations.

These requirements are critical for any system that affects a person's liberty of movement, placement under surveillance, personal data and privacy, placement on a watchlist, or ability to access protection from violence. Failing to ensure decisions meet these criteria could lead to systems that entrench rights violations which cannot be reviewed, appealed, or overturned.

The implementation of contestable rights differs significantly between contexts:

- *In civilian law enforcement and administrative settings*, contestability is relatively well established through due process protections, administrative law, and data protection regulation. Individuals can request reasons for decisions, challenge outcomes through administrative appeals, and seek judicial review.

- *In military targeting and intelligence contexts*, several structural constraints limit these mechanisms. Evidence bases may be classified and cannot be disclosed to affected individuals; security concerns may prevent disclosure of the criteria or data used in a decision; and affected individuals in conflict-affected environments may lack access to legal processes altogether. Institutions operating in these contexts should acknowledge these constraints explicitly rather than applying civilian contestability standards that cannot be implemented.

Where full contestability is not possible, institutions should:

- document the constraints;
- apply maximum available transparency to affected individuals;
- ensure that an independent internal review is conducted even where an external challenge is not feasible.

Strong legal accountability measures will be difficult to achieve where institutions purchase 'black box' technology from vendors without knowing how results are produced.[38] Contracts that limit an institution's ability to access source code, data inputs, or audit logs also limit the ability of that institution to meet its own legal obligations. Institutions should require vendors to build transparency and contestability into the systems they design. Institutions should avoid procuring AI systems whose contractual terms prevent them from fulfilling their own legal obligations. If an institution cannot independently evaluate how a system performs or investigate alleged bias, it may also be unable to demonstrate compliance with domestic or international law.

In certain circumstances, the environment in which AI systems will operate may be sensitive or classified. Complete public transparency will not always be possible. System explicability is difficult to guarantee, even to system developers, due to the underlying technical complexity of AI models. When sensitive operating environments are also not publicly explicable, there is an increased risk that these systems will be used in ways that violate rights. Confidentiality does not absolve institutions from the responsibility to identify and remedy rights violations.

Agencies should be prepared to demonstrate to independent overseers that robust internal review processes are in place to ensure questions of legality, proportionality, and broader impact are considered. Requirements for review, transparency, and remedy should be integrated throughout the AI system development life cycle. This can be accomplished by implementing these principles through formal mechanisms of legal review, documentation, oversight, and redress.

IMPLEMENTATION CHECKLIST

Legal

Legal basis

- ☑ Each AI system and its proposed use should have an identified legal basis.
- ☑ The activity supported by the AI system must comply with domestic law, international human rights law, data protection obligations, and, where applicable, IHL.
- ☑ Human rights due diligence must have been completed prior to system use and updated continuously throughout use.

Operational safeguards

- ☑ Necessary legal review is heightened for systems that may impact on individuals' liberty, freedom of movement, surveillance, or access to rights, services, or benefits.
- ☑ Data flows, uses, and storage practices must meet privacy, retention, consent, and data protection standards.
- ☑ Procedures must be in place for unofficial or shadow use of third-party AI tools by agents, such as prohibiting agents from entering personal or sensitive/classified information into third-party systems outside the agency's control.

Procurement

- ☑ Procurement documents should require vendors to meet minimum disclosure standards before a system can be used, e.g. model cards documenting training data sources, evaluation methodology, known failure modes, and performance limits; data set documentation specifying the demographic composition of training data and any known gaps or biases; pre-deployment test results disaggregated by relevant demographic groups, with performance figures for false positive and false negative rates across subgroups; incident reporting obligations requiring vendors to notify the institution of identified failures, bias events, or material changes to system behaviour; and right-to-audit clauses giving the institution or an independent body the right to inspect system inputs, outputs, and model behaviour.

IMPLEMENTATION CHECKLIST

Accountability

- ☑ It must be possible to review, explain, or legally challenge decisions supported by AI, where appropriate.
- ☑ Individuals or communities affected by systems must have the means to ask for review, explanation, remedy, or redress for adverse decisions.
- ☑ Audit logs, decision records, and other documentation must help enable legal review, accountability, and contestation.
- ☑ Institutions conducting independent legal, judicial, parliamentary, or regulatory oversight must have access to enable system use and impacts review. Systems deployed by military actors or in conflict zones must be reviewed for compliance with IHL principles of distinction, proportionality, precautions in attack, and command responsibility.

Institutional: Establish accountability and control

Effective AI governance depends on clear institutional accountability. Security institutions, not technology vendors or AI systems themselves, are responsible for decisions supported by AI. Accountability therefore requires clearly defined roles, decision-making authority, oversight arrangements, and mechanisms for identifying and correcting harms throughout the AI life cycle.

Responsibility cannot be outsourced. Security institutions remain accountable for AI-supported decisions even when systems are developed or operated by external vendors. Contracting out the development, maintenance, or operation of AI systems does not transfer responsibility for ensuring lawful, fair, and accountable use. Procurement contracts should therefore reinforce, not replace, institutional accountability.

This is a fundamental SSG/R principle, and it reinforces the theme that vendors can provide technology but they cannot assume responsibility for how public authorities exercise their legal powers.

Accountability across the system life cycle

AI systems involve multiple actors, including leadership, procurement officials, vendors, engineers, developers, system operators, legal counsel, and oversight bodies.[39] While these actors perform different functions, security institutions remain ultimately responsible for ensuring that AI systems are used lawfully, fairly, and effectively. Bias and other harms are more likely to arise when responsibilities are fragmented, poorly defined, or assumed to

rest solely with technical teams or external vendors.

Organizations should establish clear accountability across the entire AI life cycle – from problem definition and procurement through deployment, operational use, monitoring, and retirement.[40] Responsibilities should be explicitly assigned for approving system use, monitoring performance, identifying and reporting bias, investigating incidents, implementing corrective actions, and periodically reviewing whether the system remains appropriate for its intended purpose. Accountability should extend beyond technical teams to include operational, legal, procurement, governance, and executive leadership functions.

Institutions should designate a governance body or senior official with responsibility for coordinating AI governance across the organization. Depending on the institution's size and mandate, this function may be performed by an AI governance committee, an ethics or digital governance board, or a designated senior executive. This body or person should oversee high-risk AI deployments, coordinate legal, operational, procurement, and technical functions, review significant incidents, and ensure that lessons learned are incorporated into future policy and practice.

Effective governance also requires coordination across these actors. Procurement personnel need sufficient technical and legal understanding to negotiate appropriate contractual safeguards; operators must be empowered to question or override AI outputs where appropriate; legal advisers should assess compliance throughout the system life cycle; and oversight bodies require access to documentation and performance information. Institutions should therefore establish clear mechanisms for information sharing, escalation, and decision-making so that emerging risks can be identified and addressed before they become systemic failures.

Clear accountability structures are particularly important when institutions must respond to harmful outcomes in real time. Effective governance depends not only on assigning responsibility, but also on ensuring that individuals and teams have the authority, resources, and procedures needed to investigate incidents, implement corrective actions, and suspend or withdraw systems when necessary.

Incident response and organizational control

Clear governance structures are only meaningful if institutions can respond effectively when harms occur. Institutional responsibility requires having the ability to adapt or respond to unexpected impacts. When systems cause or contribute to harmful or biased outcomes, effective remediation depends on incident response processes capable of assessing and mitigating harm. Incident response processes help institutions respond to AI harms in a timely and transparent way.[41]

AI incidents should include both technical and governance failures. AI incidents are not limited to misidentification, discriminatory impacts, illegal uses, data breaches, adversarial attacks, or deployment failures; incidents can also result from technical defects, data limitations, or failures during use.

Institutions should have processes in place to determine:

- how incidents are detected, documented, and reported;
- who is responsible for investigation;
- how findings are communicated;
- what remedies are available, which can include system updates, suspension, or retirement.

They should also maintain authority to act on these incidents. In the case of harmful outcomes, vendors and leadership must retain authority to suspend or discontinue a system. Without this level of control, accountability measures may be able to uncover potential harms but unable to correct them.

Incident response also requires coordination among actors. Vendors may need to facilitate technical investigations or provide data access; operators may need to explain how system outputs are used; and oversight entities may need to validate findings. Institutions can ensure these actors are able to coordinate through contracts and internal operating procedures. Security institutions should establish clear authority across institutional actors throughout the full AI life cycle.

Institutions should use incident reviews to improve future procurement, training, governance procedures, and operations practices rather than treating each incident as an isolated event. There is therefore an important continuous learning component to mitigating AI biases in security institutions.

These actions need to come together coherently in specific use contexts. An example is border screening, where frontline officers should be able to report issues around discriminatory false positives and negatives; an internal structure should review reports (with full data available as necessary); decision-makers should be able to suspend a system during formal review; and vendors should be required to correct the system for biases.

IMPLEMENTATION CHECKLIST

Institutional

Continuous review and organizational learning

- ☑ Institutions must retain authority to pause, modify, restrict, or remove systems from use if risks outweigh benefits.
- ☑ Technology should be periodically reviewed to ensure that it continues to meet intended needs.

Incident reporting and oversight

- ☑ Processes must be in place to internally review recurring failures or patterns of harmful outcomes.
- ☑ Institutions should assign authority to oversee, suspend, modify, or terminate use of AI systems where necessary.
- ☑ Mechanisms must exist for personnel to report suspected bias, harmful outcomes, misuse, or failures.

Coordination and capacity

- ☑ Frontline, legal, technical, procurement, and leadership personnel must have mechanisms to coordinate throughout the AI life cycle.
- ☑ Procurement personnel should understand key frontline, technical, legal, and human rights risks associated with AI systems.
- ☑ Vendors and external partners should be contractually required to cooperate with audits, investigations, and external oversight processes.

Governance and accountability

- ☑ Leadership must be made consistently aware of key performance indicators, risks, and legal concerns.
- ☑ Accountability matrices must exist that clearly outline roles, decision-making authority, escalation procedures, and oversight responsibilities.
- ☑ Oversight bodies require clear procedures to investigate, document, escalate, and remediate AI-related incidents or harmful outcomes.

Management: Monitor, adapt, and maintain legitimacy

Monitoring performance and real-world impact

AI bias does not always present until after a system is deployed and in use.[42] Therefore mitigation strategies must continue past procurement to include monitoring of technical performance and outcomes.

Beyond monitoring model accuracy, false positives and negatives, subgroup performance, and model drift, institutions should track how AI impacts downstream outcomes.[43]

- Does the system make better decisions?
- Are resources being allocated more fairly or efficiently?
- Are impacts being felt disproportionately by certain groups?

Monitoring should also include thresholds for review or remediation, e.g. model retraining, data source re-evaluation, and legal compliance review, or discontinuing use of the system entirely, in order to avoid institutional normalization of harmful performance. Importantly, institutions should monitor both:

- system performance, i.e. model behaviour;
- system impact, i.e. the effects of use on people, operations, and institutions.

Monitoring should not simply identify problems, it should also improve future practice. Institutions should use monitoring results, audits, incident reviews, user feedback, and community engagement to revise procurement requirements, update SOPs, improve training, and strengthen governance arrangements for future AI deployments.

Feedback loops, trust, and frontline use effectiveness

AI systems operate within broader institutional and social systems. They rely on people for inputs, organizational behaviour, and third-party data sources. As a result, bias can create feedback loops that undermine institutional effectiveness over time. AI can meet a wide range of performance benchmarks and still be ineffective if it contributes towards adverse effects on factors like trust and legitimacy *(Figure 8)*.

If communities do not feel that security practices are impartial or equitable, they are less likely to cooperate with security institutions. This can impact on reporting, intelligence sharing, and community awareness. Reduced reporting, information sharing, and community awareness can exacerbate blind spots in AI systems.

Mitigation requires actively seeking and responding to input from users, communities, and other stakeholders affected by these systems. This can include quantitative analysis of reporting rates or participation but should include qualitative assessment of whether people feel these practices are fair, improve safety, or strengthen legitimacy. These concerns will be especially important for conflict prevention, stabilization, and rule-of-law contexts where effectiveness is heavily tied to legitimacy and trust.

AI-enabled security interventions do not exist within a vacuum. They are used by security actors operating in complex social and political environments. AI tools

Figure 8. AI bias feedback loops.

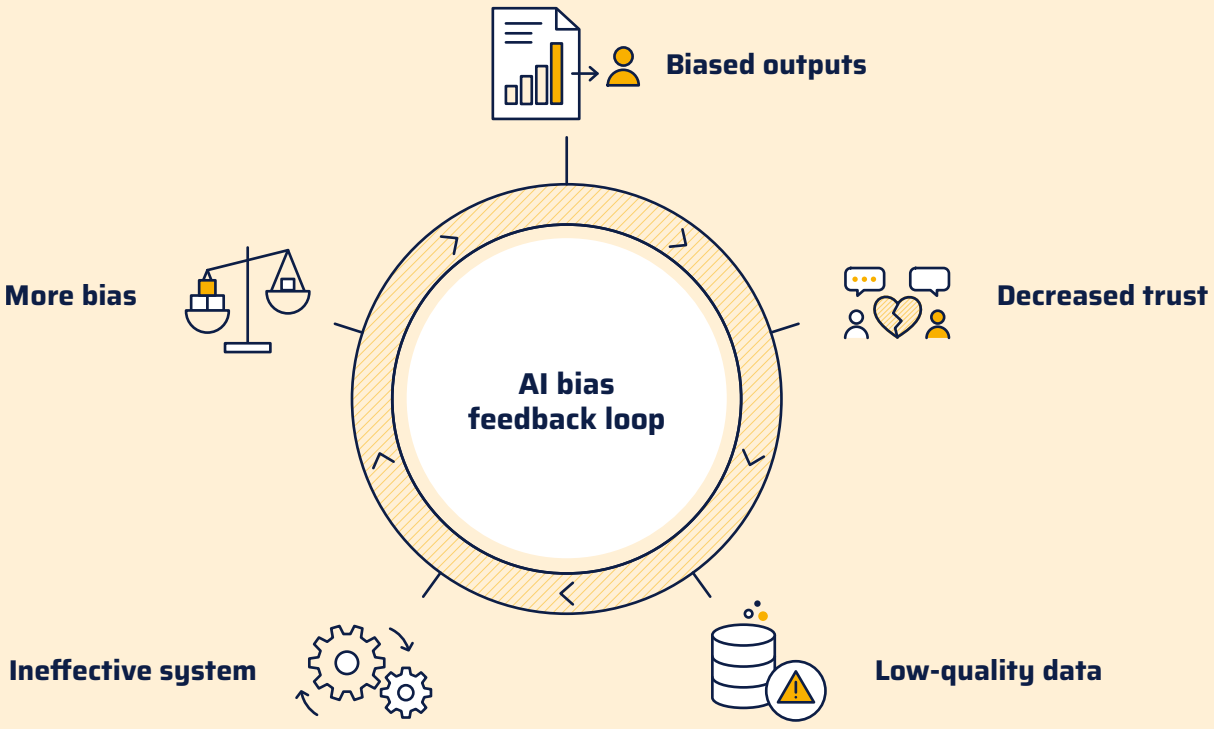


Figure 9. Forms of oversight of AI in security institutions.



that are biased, poorly governed, or fail to understand the social environments they are deployed in can cause local populations to be less cooperative. Decreased cooperation can corrupt future data collection efforts and cause real-world harm to broader mission success.

Oversight, transparency, and participation

Mitigation is impossible if AI tools are deployed in ways that concentrate power, obscure accountability, or allow harms to go unchecked. Efforts to address AI bias must include multilayered oversight both within and outside the organization.[44]

Oversight entities require both access and capacity. They must be able to examine system documentation, underlying data, testing methodologies, audit logs, and vendor contracts. If confidentiality is necessary for security reasons, there should be confidential oversight mechanisms in place. Oversight mechanisms should be operationally and institutionally independent from the teams responsible for developing or using AI systems. Independent oversight helps ensure that concerns about bias or misuse can be investigated objectively.

Transparency is closely linked to participation. Stakeholders affected by AI-supported decisions can provide context

that institutions lack. Consultation with civil society, women peacebuilders, and local communities can reveal risks not evident in technical audits, such as patterns of recurring harm or that certain groups are disproportionately affected.

Consultations must be conducted safely. In certain settings, stakeholders may be placing themselves at risk by interacting with security institutions. Consideration should be given to remote or anonymized reporting tools, secure in-person consultation, and working through intermediary organizations. Oversight must also account for the fact that AI systems affect different populations differently.

Meaningful consultations can improve the performance of AI-enabled systems in security institutions. An example is ensuring a police force's AI-enabled early warning system includes underreported metrics around violence against women, by using community-provided information – including that obtained from victim-focused CSOs.

Gender-responsive and intersectional mitigation

Mitigation should be gender-responsive and intersectional. The WPS agenda is grounded in a core premise directly relevant to mitigating AI bias in security institutions: no population experiences security risks in the same way, and security institutions cannot operate effectively or legitimately if their approaches fail to account for the needs and experiences of all populations.[45]

Risk indicators built around traditional concepts of security such as deaths or arrests may miss gendered harm like sexual violence, coercion, threats, mobility limitations, or attacks on women leaders. These risks may emerge earlier and are often missing from traditional data sets.[46]

Security actors should ask three questions

1. Are systems missing gendered and non-kinetic threats?
2. Do outputs have disproportionate effects on different groups?
3. Is data missing forms of harm that are less likely to be reported?

Systems may appear effective when evaluated against aggregate data sets but miss how they are harming or not serving certain groups. This is particularly relevant for women and girls with intersecting identity factors, such as ethnicity, migration status, age, or disability, which can compound risks of exclusion and discrimination.

Engaging subject-matter experts on gender and inclusion throughout governance, procurement, data analysis, and monitoring and evaluation efforts can help surface these risks early on. The following implementation checklists can help institutions monitor and adapt AI systems over time:

IMPLEMENTATION CHECKLIST

Management

Monitoring

- ☑ Systems should be evaluated on technical performance, as well as broader frontline, institutional, and societal impacts.
- ☑ AI-supported security practices should be evaluated for unfairly targeting certain populations or reinforcing broader patterns of exclusion or discrimination.
- ☑ Subgroup and population-level impacts should be regularly assessed across sex, ethnicity, age, disability, migration status, geography, and other relevant identity factors.
- ☑ Systems should be monitored for feedback loops that could distort future data collection, threat labelling, or prioritization by frontline personnel.

Legitimacy

- ☑ Monitoring must include indicators related to changes in reporting behaviour, information-sharing practices, willingness to cooperate with security institutions, and access to protective services.
- ☑ Monitoring processes should assess how the use of AI impacts trust, legitimacy, community cooperation, and perceptions of fairness.

Oversight

- ☑ External stakeholders, parliamentary oversight bodies, CSOs, and affected communities are consulted in system monitoring and evaluation where possible.
- ☑ Qualitative feedback, community engagement, and local knowledge should be incorporated into system monitoring, evaluation, and revision processes.
- ☑ AI-supported security practices should be verified for adaptability to changing environments, conflict dynamics, and evolving social risks.

Organizational learning

- ☑ Clear criteria and thresholds are established for intervention, retraining, restriction, suspension, or decommissioning AI systems if they produce harmful or ineffective results.
- ☑ Lessons learned from the use of technology audits should be integrated into future procurement, technology governance, and deployment decisions, such as through records of use, incidents, or community feedback.
- ☑ Regular reassessments of decisions to use AI should be conducted against security sector goals related to effectiveness, accountability, inclusion, and rule of law.

Scenario-specific checklists

This subsection presents checklists specific to typologies of scenarios, partially linked to examples provided in Section I. Generally, these include actions that relate to multiple aspects of mitigation presented in Section II (frontline use, legal, institutional, management) and complement the implementation checklists presented under these headings in Section II. These checklists are not intended to be exhaustive compliance requirements, but rather are context-dependent safeguards.

There are 13 checklists:

Identification and analysis

1. AI-enabled identity verification
2. Facial recognition biases
3. AI-enabled surveillance
4. Intelligence fusion centres

Recommendation development

5. AI-enabled risk scoring
6. Risk-identification models that inform resource allocation
7. Machine-generated recommendations

Context-specific deployments

8. AI bias and conflict dynamics
9. Intersectional bias in AI detection of online threats
10. Gender bias in AI-enabled early warning systems

Security institution legitimacy

11. Accountability gaps in border screening
12. Bias and security provider legitimacy
13. Feedback loops and institutional legitimacy

SCENARIO CHECKLIST 1

AI-enabled identity verification

When is this useful?

Use this checklist when AI systems are used to verify or authenticate a person's identity for access, security screening, or service delivery.

- ☑ Test technology on as broad a sample pool as possible.
- ☑ Create fallback plans if technology fails to confirm identity.
- ☑ Consider using more than one modality to verify identity.
- ☑ Track error rates for different populations.
- ☑ Create legal accommodations for people facing adverse outcomes.
- ☑ Re-evaluate periodically after deployment.

SCENARIO CHECKLIST 2

Facial recognition biases

When is this useful? Use this checklist when facial recognition or other biometric matching technologies are used to identify, verify, or investigate individuals.

Technical actions

- ☑ Do not use biometric outputs as definitive evidence of identity.
- ☑ Test systems across demographic groups prior to deployment using established bias detection methods, e.g. disparate impact testing by applying the 80 per cent rule to compare selection rates across groups, evaluation of subgroup performance separately rather than relying on aggregate accuracy alone, or definition of minimum sample sizes per demographic group sufficient for statistically meaningful results.
- ☑ Document performance thresholds – the acceptable maximum disparity in false positive or false negative rates across groups – before deployment.
- ☑ Define thresholds for when outputs can be used.
- ☑ Require independent corroboration as a prerequisite for enforcement.
- ☑ Maintain audit logs documenting how outputs inform decisions.
- ☑ Monitor performance periodically after deployment.

Legal safeguards

- ☑ Have clear legal authority to use a technology.
- ☑ Know and understand how use of a technology may impact human rights.
- ☑ Set parameters around when, where, and how a technology should be used.
- ☑ Keep detailed documentation and logging to facilitate outside review and audits.
- ☑ Give individuals the opportunity to challenge or seek review of decisions that negatively affect them.
- ☑ Implement independent oversight when using high-risk AI-enabled technologies.
- ☑ Know who is responsible for decisions supported by AI systems.
- ☑ Regularly evaluate whether use of a technology continues to provide more benefit than harm.

SCENARIO CHECKLIST 3

AI-enabled surveillance

When is this useful? Use this checklist before deploying or expanding AI systems that monitor people, locations, communications, or public spaces.

- ☑ Deployment decisions should be subject to legal authorization and oversight.
- ☑ Clear use-case limitations should define acceptable versus prohibited uses.
- ☑ Systems should be evaluated not only for accuracy but also for patterns of use.
- ☑ Independent oversight bodies should review surveillance practices.
- ☑ Community impact assessments should be conducted prior to deployment.

SCENARIO CHECKLIST 4

Intelligence fusion centres

When is this useful? Use this checklist when AI tools support intelligence analysis by combining information from multiple sources or security institutions.

- ☑ Train operators to understand AI capabilities and limitations.
- ☑ Build in safeguards that ensure analysts must provide reasoning for AI-informed decisions.
- ☑ Design AI systems with interfaces that make uncertainty and alternative explanations known.
- ☑ Encourage red-teaming and adversarial thinking.
- ☑ Ensure that AI does not replace human judgement.

SCENARIO CHECKLIST 5

AI-enabled risk scoring

When is this useful? Use this checklist when AI assigns or supports assessments of an individual's or group's level of risk that may influence operational or administrative decisions.

- ☑ Provide transparency and explicability regarding AI-enabled risk scoring decisions.
- ☑ Review domestic and international laws to identify any potential legal risk prior to use.
- ☑ Ensure that AI systems are audited to identify proxy discrimination if sensitive attributes are not used.
- ☑ Implement opportunities for individuals to provide input or contest decisions made about them.
- ☑ Restrict use of AI-enabled risk scoring where use of predictive outcomes can affect rights without appropriate safeguards.
- ☑ Apply a proportionality test, i.e. whether the risks to privacy or discrimination are justified by a specific, well-defined public benefit, and whether less intrusive alternatives were considered and rejected.

SCENARIO CHECKLIST 6

Risk-identification models that inform resource allocation

When is this useful? Use this checklist when AI-generated risk assessments influence decisions on where personnel, patrols, funding, or other security resources are deployed.

- ☑ Differentiate between incidents and risk.
- ☑ Identify alternative data for model training, such as victimization surveys and community-generated data.
- ☑ Perform geographic audits on model outputs to identify concentration.
- ☑ Avoid using predictive model outputs as sole justification for deployment of resources.

SCENARIO CHECKLIST 7

Machine-generated recommendations

When is this useful? Use this checklist when AI provides recommendations or decision support that could influence frontline decisions.

- ☑ Define rules of engagement for how automation-supported tools are used.
- ☑ Include gender advisers and affected communities in system design and evaluation processes.
- ☑ Insist on independent source validation before allowing machine-generated recommendations to dictate frontline decisions.
- ☑ Train users on automation bias.
- ☑ Ensure analysts maintain the ability to challenge machine-generated decisions.
- ☑ Slow down decision timelines as necessary to allow for review, investigation, and challenges to machine-generated recommendations.
- ☑ Regularly assess whether surveillance coverage is becoming too concentrated in some areas.
- ☑ Use randomization, alternative analytic techniques, or red-team exercises to discover where blind spots may exist.

SCENARIO CHECKLIST 8

AI bias and conflict dynamics

When is this useful? Use this checklist when assessing how AI-enabled technologies may influence conflict dynamics and broader security conditions, including by contributing to misinformation or polarization.

- ☑ Monitor shifts in AI-enabled information signals as part of threat analyses.
- ☑ Partner with technology companies to help identify and mitigate systemic risks.
- ☑ Include digital vulnerability indicators in early warning monitoring.
- ☑ Understand connections between AI-enabled bias online and security outcomes.
- ☑ Plan for how amplification of harmful signals could impact the security environment.

SCENARIO CHECKLIST 9

Intersectional bias in AI detection of online threats

When is this useful? Use this checklist when AI systems detect or help to prioritize the mitigation or removal of online threats and harmful content, and provide indicators of violence, across different communities.

- ☑ Train moderation tools on intersectional data that is localized and context-specific, including in different languages and regions.
- ☑ Include review steps to ensure human moderation and civil society input can identify gaps.
- ☑ Consider absence of threats detected by AI systems to be indicative of data or model limitations.
- ☑ Expand early warning analysis to include gender and intersectional analysis.
- ☑ Provide supplemental security assistance for those facing compounded threats.

SCENARIO CHECKLIST 10

Gender bias in AI-enabled early warning systems

When is this useful? Use this checklist when AI supports early warning or conflict prediction systems that are meant to detect emerging security risks, especially those affecting diverse populations.

- ☑ Early warning indicators should be context-specific and go beyond fatality data alone.
- ☑ Early warning systems should triangulate data inputs with community reporting and other qualitative data.
- ☑ Analysts should consider whether and how certain types of violence may be underreported.
- ☑ Early warning systems should be field tested to ensure they are able to detect violence as experienced by diverse populations.
- ☑ Women's organizations and other local stakeholders should play a role in the modelling of the systems.
- ☑ Model performance should be evaluated separately for male and female subgroups – and for intersectional subgroups, e.g. women by age group, ethnicity, or displacement status – rather than relying on aggregate accuracy

metrics that may conceal severe disparities for specific populations.

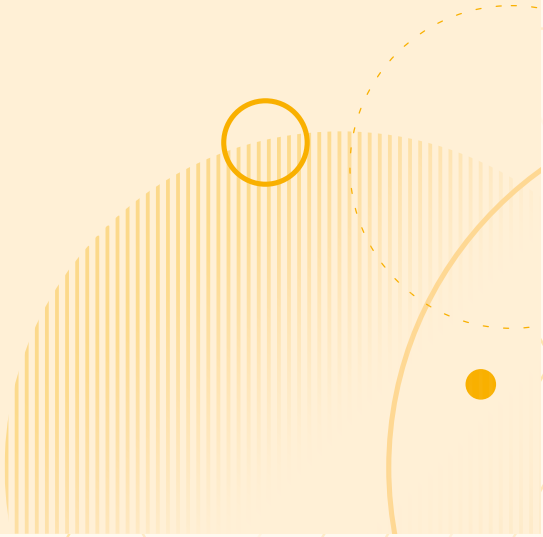
- ☑ Disparate impact testing should be applied using the four-fifths rule or equivalent threshold, i.e. if the selection or adverse outcome rate for one gender or intersectional group is less than 80 per cent of the rate for the most-favoured group, the disparity warrants review and potential remediation before deployment.

SCENARIO CHECKLIST 11

Accountability gaps in border screening

When is this useful? Use this checklist when AI systems support immigration, customs, border management, or traveller screening decision-making processes.

- ☑ Identify responsible teams for specific roles related to procurement, deployment, and operation of AI systems.
- ☑ Designate specific teams or individuals for reviewing, escalating, suspending, or modifying AI system use.
- ☑ Require vendors to provide technical support for transparency, auditing, documentation requests, and incident remediation.
- ☑ Ensure that operators are trained on processes for challenging AI outputs.
- ☑ Document key decisions related to technology use, including system design changes, use cases, and delegation of responsibilities.
- ☑ Create incident response and review processes that span organizational teams.
- ☑ Establish redress and compliance mechanisms to facilitate explanation, review, and remedy for affected individuals.



SCENARIO CHECKLIST 12

Bias and security provider legitimacy

When is this useful? Use this checklist when evaluating whether AI-enabled practices may affect fairness, and therefore public confidence in and legitimacy, of security institutions.

- ☑ Community perceptions of injustice should be included in monitoring and evaluation.
- ☑ Bias should be measured not just through technical evaluations, but through social and institutional lenses.
- ☑ Oversight mechanisms should involve civilian and community representation.
- ☑ AI systems should be evaluated based on how they will affect legitimacy and trust.
- ☑ Bias should be addressed with a lens of ensuring long-term effectiveness and stability.

SCENARIO CHECKLIST 13

Feedback loops and institutional legitimacy

When is this useful? Use this checklist when monitoring the longer-term impacts of AI-enabled frontline decision-making on public trust and institutional legitimacy.

- ☑ Include sociological analysis and community feedback in information monitoring and evaluation processes.
- ☑ Continuously measure how AI-enabled operations impact public trust, legitimacy, and cooperation.
- ☑ Be aware of how AI tools may create feedback loops that skew future data collection efforts or create flawed threat assessments.
- ☑ Involve local stakeholders and civil society in oversight and evaluation efforts when possible.
- ☑ Use a mixture of technical and sociological metrics to evaluate systems.
- ☑ Restrain or exclude AI-enabled practices that do not align with institutional goals, such as legitimacy, inclusion, and long-term stability.

Conclusion: From risk awareness to institutional practice

Bias cannot be entirely predicted or eliminated in complex security environments. The objective is therefore not to eliminate all risk, but to identify bias early, reduce its likelihood and impact, and respond effectively when it occurs. This requires security institutions to move beyond high-level principles and embed concrete safeguards into governance arrangements, procurement, contracting, data governance, SOPs, legal review, auditing, incident response, and continuous monitoring.

In practice, there are four areas where mitigation measures should be included:

1. **Frontline use**, supported by procedures that ensure AI-informed decisions remain contextual, proportionate, and free of systematic bias.

2. **Legal compliance**, supported by clear legal authority, human rights safeguards, mechanisms for review, and access to remedy.

3. **Institutional governance**, supported by clearly assigned roles, oversight responsibilities, and accountability throughout the AI life cycle.

4. **Continuous management**, supported by monitoring, community feedback, evaluation, and adaptation to maintain legitimacy and operational effectiveness.

AI systems should be treated as institutional capacities rather than standalone technologies. Like doctrine, operational procedures, or new equipment, they require governance, oversight, training, evaluation, and periodic reassessment throughout their use. Institutions that understand where bias can emerge – and establish mechanisms to detect, review, and correct it – will be better positioned to protect rights, strengthen public trust, improve operational effectiveness, and achieve sustainable security outcomes.

Translating these principles into practice requires practical tools and repeatable institutional processes. The next section translates these governance principles into practical implementation tools that security institutions can adapt to their own operational environments.

There are broadly three tools:

- **Tool 1:** Scenarios, questions, and interpretation frameworks for security institutions

- **Tool 2:** AI bias risk assessment, including three more specific tools: (2a) Use of commercial AI tools, (2b) Feeding institutional data into a commercial AI system, (2c) Institutions building or training their own AI system

- **Tool 3:** Inclusive security and gender analysis template

SECTION III

Toolkits



The tools in this section are designed to help security institutions implement the mitigation strategies outlined in Section II. They are designed to support security practitioners in recognizing, mitigating, and responding to AI-related risks while enhancing effectiveness, accountability, inclusivity, and adherence to the rule of law.

- **Tool 1:** Scenarios, questions, and interpretation frameworks for security institutions is designed to improve the quality of prompts, and encourage critical reasoning on AI-generated information in order to avoid overreliance on AI-generated outputs and flag key blind spots and assumptions.

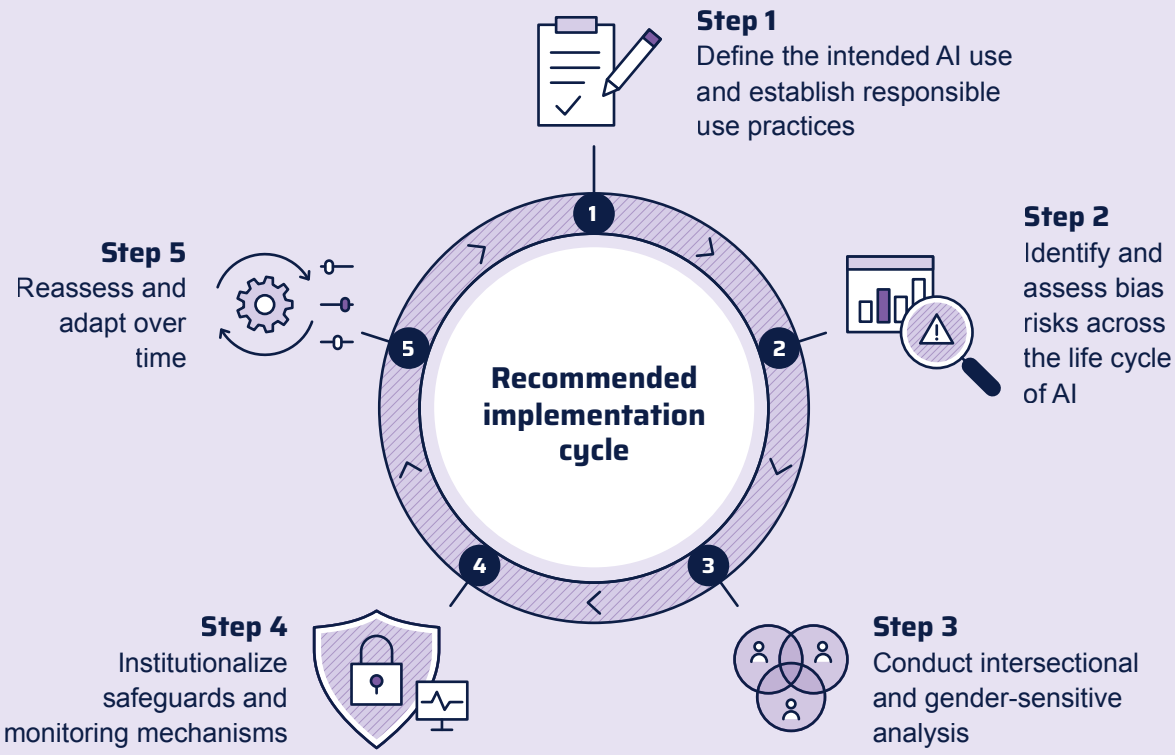
- **Tool 2:** AI bias risk assessment includes three more specific tools for use after an assessment of which tool is best utilized: (2a) Use of commercial AI tools, (2b) feeding institutional data into a commercial AI system, and (2c) Institutions building or training their own AI system.

- **Tool 3:** Inclusive security and gender analysis template better enables relevant analysis regarding dimensions not fully taken into account in standard gender analysis: technology-facilitated gender-based violence (TFGBV), feedback loops compounding in adaptive AI systems, effects on civil participation, and intersectional analysis, i.e. beyond single-category checklists.

These tools are intended to complement, not replace, existing legal obligations, institutional policies, human rights standards, and frontline procedures. Institutions may use the tools individually or together as an integrated process. While described sequentially, implementation will often be iterative, and assessments and adaptations should take place across the AI life cycle. Many of these tools reflect new approaches based on AI safety communities and technical environments – including structured testing, iterative assessments, and human oversight, all of which are tailored to security institutions.

In alignment with inclusive security principles and SSG/R, institutions need to ensure that affected communities, including women, marginalized populations, civil society actors, and relevant technical experts, are meaningfully engaged throughout implementation.

Figure 10. Recommended implementation cycle.



Tool 1: Scenarios, questions, and interpretation frameworks for security institutions

AI use cases and prompting guide

Purpose

Important limitation: prompt engineering is a tool for improving the clarity, consistency, and surface quality of AI-generated outputs – it is not a bias mitigation tool in the technical sense. Prompting can shift the framing or register of outputs, but does not address bias embedded in model weights, training data distributions, or fine-tuning choices. For high-stakes security decisions – targeting, detention, border screening, watchlist management – a system that has received a fairness instruction in its prompt is not necessarily a less biased system; it may produce outputs with a more balanced surface appearance while underlying inferential patterns remain unchanged. This tool should be used to improve output quality and communication. It must not be presented to oversight bodies, procurement authorities, or affected communities as evidence that bias has been mitigated at the model or data level.

Poorly formulated prompts, overreliance on outputs, or insufficient contextual understanding can raise the potential for automation bias, with implications for security decision-making. Prompt quality directly shapes AI bias.

The following are examples of prompts that increase bias:

- Vague or unspecified, e.g., 'Identify dangerous places."
- Demographic information (names, ethnicity, etc.), e.g., 'Visitors from Country X showing …'
- Implicit clues (without labels), e.g., 'Somebody who prays five times daily …'
- Overly polite phrasing, e.g., 'Could you kindly help identify suspects …'

Conversely, the following are examples of prompts that decrease bias:

- Step by step chain of thought, e.g., 'Explain the reasons phase by phase.'
- Fairness introduction with chain of thought, e.g., 'Play the role of fair analyst and consider the matter step by step.'
- Iterative reprompting, e.g., 'Revise the output after reviewing for bias.'
- Counter-stereotype phrasing, e.g., 'Provide factors beyond typical profiles.'

Given these effects of well-framed or poor prompts, this tool enables security practitioners to: improve the quality of prompts input into AI systems; reduce overreliance on automated outputs; encourage critical reasoning and reading of AI-generated information; and identify potential blind spots and assumptions.

Who should use this tool	When should this tool be used?
• Intelligence analysts • Security planners • Border and migration personnel • Early warning practitioners • Human rights advisers • Cybersecurity analysts • Programme managers • Frontline leadership	• In day-to-day AI-supported analysis • Before making decisions based on AI outputs • During analyst training • When developing SOPs

⚠ **Caution:** It is impossible fully to eliminate biases through prompting strategies, because stereotypes in training data remain. Prompt engineering is therefore only a safeguard layer.

Instructions

For each AI-supported task:

1. Provide a clear definition of the question.
2. Create question types (prompts) that help identify assumptions and uncertainty. Use the enhanced prompt templates in each scenario below.
3. Review outputs critically using the interpretation questions provided.
4. Use additional sources of information to validate findings.
5. Determine whether outputs support action or require further investigation.

Scenario 1: Gender-sensitive early warning

Identifying emerging risks of targeted violence against women political leaders

Poor prompt	Enhanced prompt
✗ *'Identify dangerous areas/ regions.'* **Why:** No population specificity, no fairness framing. Likely to surface general hotspots reflecting existing geographic or ethnic biases in training data, not risks specific to women leaders.	✓ *'Using available reporting, identify possible signs that violence against women political leaders in Region X is increasing. Explain confidence levels, underlying assumptions, and key information gaps. Think step by step.'* **Why:** Specifies the geography and group at-risk, requests reasoning and that the model highlight its own uncertainty.
✗ *'Who is most at risk of violence in Region X?'* **Why:** Unanchored demographic framing. Models produce different risk profiles given demographic features suggested by names of regions or population descriptors.	✓ *'Based on available incident data, which categories of political actors in Region X face documented threats, and what patterns of escalation precede such threats? Note confidence levels and data gaps.'* **Why:** Anchors risk to documented incident categories rather than demographic inference, reducing the chance that the model projects stereotyped profiles.
✗ *'Summarize recent violence trends.'* **Why:** No gender lens, no geographic scope, no confidence framing. Likely to aggregate dominant narratives and miss gender-specific patterns that are underreported in mainstream sources.	✓ *'Summarize documented trends in politically motivated violence in Region X over the past 12 months, with specific attention to patterns affecting women in leadership roles. Distinguish between confirmed incidents and unverified reports. Identify which populations or regions may be underrepresented in the data.'* **Why:** Introduces an explicit gender lens, time frame, and source quality distinction, in order to improve transparency and limit amplification bias.

Interpretation questions

- Is the system making assumptions about which groups are at risk – and are those assumptions stated?
- What data or information sources might be missing (e.g. rural areas, minority languages, informal networks)?
- Are women from different ethnic, political, or socioeconomic backgrounds represented equally in the output?
- What empirical evidence supports the identified risk indicators?
- Do the findings warrant action or further investigation, or should uncertainty be flagged to decision-makers?

Scenario 2: Border risk screening

Identifying travellers who may require additional review

Poor prompt	Enhanced prompt
✗ *'Identify suspicious travellers.'* **Why:** Vague. Undefined criteria invites the model to consider demographic stereotypes. Characteristics like ethnicity and nationality may influence AI behaviour.	✓ *'Identify the documented procedural factors used to determine which travellers require enhanced screening. For each factor, explain the confidence level and flag any variables that may produce disproportionate results for specific nationalities, genders, or socioeconomic groups. Think step by step.'* **Why:** Ensures that procedural criteria underpin the request, instead of relying on impressions. Requesting the flagging of biases activities self-correction features in models.
✗ *'Identify patterns among travellers flagged last month.'* **Why:** Without explicit fairness framing, pattern analysis on flagged populations risks reinforcing existing selection biases – the model learns from a set already skewed by prior screening decisions.	✓ *'Analyse documented screening outcomes from last month. Identify whether flagging rates vary significantly across nationality, gender, or age groups. Where disparities exist, assess whether they reflect documented risk criteria or potential proxy bias. Note confidence levels.'* **Why:** Requests analysis of disparities and differentiates flagging based on criteria or proxies, in order to counter stereotypes.
✗ *'Which travel routes carry the highest risk?'* **Why:** Route risk without definitional context defaults to patterns in training data, which may conflate high-traffic or high-poverty corridors with risk – amplifying socioeconomic bias.	✓ *'Based on documented incident and enforcement data, which travel routes are associated with elevated rates of [specify: smuggling/document fraud/etc.]? Distinguish between routes with strong evidentiary support and those based on limited reporting. Flag any geographic or demographic groups that may be disproportionately represented.'* **Why:** Specifies the type of risk, requires source quality differentiation, and flags proportionality concerns, in order to limit amplification bias.

Interpretation questions

- Could nationality, ethnicity, or geography be functioning as a proxy variable for risk?
- Are any groups disproportionately flagged compared to their share of the traveller population?
- Are there alternative, non-demographic explanations for any identified patterns?
- Would this output hold up to independent legal or human rights review?
- Is this output sufficient for action, or does it require corroboration?

Scenario 3: Community sentiment analysis

Tracking online narratives associated with instability

Poor prompt	Enhanced prompt
✗ *'Identify dangerous online activity.'* **Why:** Undefined scope and no fairness instruction. Without specificity, the model can amplify content from majority demographic groups and downplay minority voices.	✓ *'You are a fair and impartial analyst. Identify online narratives associated with a potential for violence in [context], distinguishing between legitimate political expression, misinformation, and direct threats. Explain confidence levels, note whose voices may be missing, and flag assumptions. Think step by step.'* **Why:** Combines role framing with chain of thought, which can achieve larger reductions of bias. Explicitly distinguishes dissent from threat, reducing risk of conflating the two.
✗ *'What are communities saying about [group/event]?'* **Why:** Framing around a named group without a fairness instruction can cause the model to aggregate stereotyped narratives associated with that group in training data, presenting them as community sentiment.	✓ *'Across available online sources, what range of perspectives are expressed about [group/event] in [region/platform]? Identify dominant narratives, dissenting voices, and communities whose views may be underrepresented. Note the confidence level for each narrative and the types of sources it draws from.'* **Why:** Requests numerous perspectives and names underrepresentation as a important to surface, which helps counter stereotypes.
✗ *'Is unrest increasing in this community?'* **Why:** Binary framing (increasing/ not increasing) suppresses nuance and uncertainty. Prompts that are ambiguous or binary are more prone to confident-sounding yet stereotyped outputs.	✓ *'Based on available reporting, what indicators of social tension or unrest are documented in [community/region]? For each indicator, describe the strength of evidence, note which populations are represented in the sources, and identify what information would be needed to assess trajectory with confidence.'* **Why:** Replaces binary framing with an evidence-mapping structure, requires source population transparency, and explicitly names what is unknown – all reducing automation bias risk.

Interpretation questions

- Could these outputs reinforce existing institutional assumptions about which communities are dangerous?
- What populations are likely to be absent from the data (e.g. offline communities, non-dominant languages)?
- Are AI outputs being treated as evidence or as investigative leads requiring further verification?
- What additional human or contextual intelligence is needed before any action?
- Could uncertainty in this output, if ignored, drive inappropriate decisions?

IMPLEMENTATION NOTES

- **AI outputs should assist human judgement, not replace it.** Institutions should have clear procedures specifying when outputs require corroboration before informing action.
- **Reprompt to reduce bias further.** After receiving an output, follow up with 'review for potential demographic bias and revise as necessary'.
- **Document the prompt and rationale.** For consequential decisions, record the prompt used, output received, interpretation applied, and corroborating sources consulted.
- **No prompt fully eliminates bias.** Implicit stereotypes present in training data remain a factor even when prompts are effective. Treat AI outputs as one input among many.

Tool 2: AI bias risk assessment

A guide for security institutions to identify which tool to use

Purpose

Bias risks in AI systems are shaped by how an institution uses AI. The actions available to frontline personnel and those in management – and the risks for which they are responsible – depend entirely on whether an entity is using an existing commercial tool, feeding an institution's data into a commercial tool, or building a system itself.

This subsection helps security institutions identify which situation applies and points them to the right risk assessment tool. There are three options. Most security institutions will fall into Option 1 or Option 2.

Step 1: Identify the situation

Read the descriptions below and select the one that best matches how the institution is using or planning to use AI.

If this describes the situation…	One is in…
The institution accesses a commercial AI service (ChatGPT, Claude, Copilot, etc.) without supplying its own data or files to it.	**Option 1** – Using a commercial AI tool as-is
The institution feeds its own data, documents, or records into a commercial AI service to generate analysis or outputs.	**Option 2** – Feeding one's data into a commercial AI system
The institution builds, trains, or has commissioned a custom AI system using data the institution has collected.	**Option 3** – Building or training one's own AI system
The institution is unsure.	**Start with Option 2;** if the institution is providing any institutional data to an AI system, Option 2 applies as a minimum

⚠ **Important:** Many institutions move between options over time – for example, starting with a commercial tool *(Option 1)* and later connecting it to institutional databases *(Option 2)*. Reassess which option applies whenever use of AI changes significantly.

Step 2: Use the right risk assessment tool

Once a situation is identified, the corresponding Tool 2 document applies. Each is tailored to the risks and actions relevant to that situation only – content that does not apply has been removed.

Option 1: Using a commercial AI tool as-is	
What one controls	How one uses and queries the tool. Internal policies on acceptable use and human review. What one demands from the vendor in terms of transparency about known limitations.
What one cannot change	The underlying AI model, its training data, how it was built, and what biases are encoded in it.
Where to act	Procurement (demand transparency from vendors). Deployment (set human review requirements). Monitoring (track outputs for bias patterns). Prompting (see Tool 1).
The associated risk assessment tool	Tool 2A – Risk assessment for use of commercial AI tools.
Option 2: Feeding data into a commercial AI system	
What one controls	The data that is provided to the system – its quality, coverage, and gaps. How one queries the system (prompts). The procurement terms that are negotiated. Deployment conditions and human review requirements.
What one cannot change	The underlying AI model. Its original training data and architecture. Any biases baked into the model before the data reaches it.
Where to act	Problem definition, i.e. is this the right use? Data quality and coverage before it goes into the AI system. Procurement terms. Deployment conditions. Monitoring of outputs.
The associated risk assessment tool	Tool 2B – Risk assessment for feeding institutional data into a commercial AI system.
Option 3: Building or training one's own AI system	
What one controls	The entire system: how data is collected, labelled, and used; how the model is designed and tested; how it is deployed and monitored. Full accountability rests with the institution.
What one cannot change	Nothing – the institution owns the full stack. All bias risks at every stage are the institution's responsibility.
Where to act	All life-cycle stages: problem definition, data collection, data labelling, model development, procurement (if outsourcing), deployment, and monitoring.
The associated risk assessment tool	Tool 2C – Risk assessment for institutions building or training their own AI system.

Step 3: Connect to the other tools

Regardless of which option applies, use these tools alongside the risk assessment.

- **Tool 1** (AI use cases and prompting guide) for improving how the institution queries AI systems day-to-day. Relevant to all three options.
- **Tool 3** (Inclusive security and gender analysis template) for assessing how the AI system may affect different populations. Relevant to all three options.

Tool 2A: Risk assessment for use of commercial AI tools

For institutions using an existing AI tool without supplying institutional data

What this tool covers

The institution is using a commercial AI tool – such as ChatGPT, Microsoft Copilot, or Claude – where staff access the system to get help with drafting, summarizing, translating, or analysing information, without uploading institutional files or databases. The AI system itself – how it was built, what data it was trained on, what biases are encoded in it – is entirely outside the institution's control. It cannot change those things. What it can control is how it selects the tool, how staff use it, what policies govern its use, and how it monitors the outputs. This risk assessment covers only those stages. Stages that are not actionable for the institution (data collection, data labelling, model development) have been removed.

► **If the situation has changed:** If the institution is now feeding its own records, incident data, or case files into the system, move to Tool 2B instead.

Who should use this tool	When should this tool be used?
• Security planners and frontline leadership • Procurement officials • Programme managers • Human rights and gender advisers • Oversight actors	• Before selecting or deploying a commercial AI tool • When reviewing or renewing a vendor contract • During annual reviews • Following incidents where an AI output contributed to a decision

Instructions

For each stage in the assessment:

1. Work through the key questions and equity lens for the context.
2. Score Likelihood (1–4) and Severity (1–4) using the scoring grid. Multiply to get a composite score.
3. Assign a risk level and select relevant actions from the illustrative list.
4. Assign a responsible actor and note escalation triggers.
5. Check the red-lines box – any matching condition means the tool should not be used for that application.

⚠ **Remember:** One cannot fix bias inside a commercial AI tool. The institution's leverage is at procurement (choosing a better tool or demanding transparency) and deployment (limiting use cases and requiring human review). Prompting well reduces bias in outputs but does not eliminate it.

Risk scoring framework

Score each stage by multiplying Likelihood × Severity.

Likelihood/ Severity ▶	1 Minimal	2 Limited	3 Significant	4 Critical	Decision guidance
1 Unlikely	Low	Low	Moderate	High	Document and monitor. No deployment block.
2 Possible	Low	Moderate	Moderate	High	Mitigations required before deployment.
3 Likely	Moderate	High	High	Unacceptable	Senior sign-off and safeguards mandatory.
4 Almost certain	High	Unacceptable	Unacceptable	Unacceptable	Do not proceed without full oversight review.

Level	Description and action
Low	Minimal foreseeable harm. Document and monitor.
Moderate	Risks requiring mitigation before or during deployment.
High	Significant risks requiring safeguards and senior sign-off.
Unacceptable	Do not deploy. Full oversight review required. Check red lines.

Red lines – When not to use this tool

If any of the following apply, do not use the commercial AI tool for this application, regardless of other scores. Red lines are only effective if the institutional authority to invoke them is clearly assigned.

AUTOMATIC UNACCEPTABLE RATING – HALT DEPLOYMENT

If any of the following conditions apply, classify as Unacceptable regardless of other scores.

- The vendor cannot provide documentation explaining how the system makes decisions.
- The system will replace – not inform – human judgement in high-stakes decisions (arrest, detention, lethal force).
- No contestation or redress mechanism exists for individuals affected by AI-driven decisions.
- The system has not been tested against representative populations in the intended deployment context.
- Deployment would violate applicable legal prohibitions (e.g. EU AI Act prohibitions on real-time biometric surveillance without judicial authorization).

Three questions must be answered before deployment

1. **Who has the authority to invoke a red line – and can that authority be exercised by staff below senior leadership level?** In most institutions, procurement decisions are made above the frontline level, which means a frontline analyst or mid-level officer who identifies a red-line condition may have no authority to halt a deployment already mandated by leadership.
2. **What is the escalation pathway when a red-line condition is identified but leadership has already committed to the system?** Staff need a documented process for raising concerns – ideally through an independent oversight body or designated compliance function – that is protected from retaliation.
3. **How are red-line determinations documented?** Decisions to proceed despite a potential red-line condition should be recorded in writing, with the reasoning stated, so that accountability is maintained and patterns of override can be reviewed.

Risk assessment matrix

Complete one row per stage. The equity and intersectionality lens (purple) surfaces risks to specific populations. Actions listed are illustrative – adapt to the context.

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution can take
Procurement and selection	Can the vendor explain how the system works and what its known limitations and failure modes are? Has the vendor conducted bias testing across relevant populations?	*Has the vendor tested the system on populations similar to those the institution works with? Are women, minorities, and displaced persons represented in testing?*	**High**	• Request written documentation of known bias risks and limitations before selecting any commercial tool • Ask vendors explicitly: what demographic groups does this system perform worse on? • Include bias disclosure and incident reporting obligations in any contract or terms of use • Confirm the system has been tested in contexts comparable to the environment in which the institution works
Deployment and use policy	Will security decisions be made or informed by this tool? Who reviews outputs before they influence action? What use cases is this tool prohibited from?	*Which communities will be most exposed to errors? Are staff applying the tool consistently across demographic groups, or selectively?*	**High – check red lines**	• Establish a written policy specifying which decisions may be informed by AI outputs and which require human review only • Require HITL for all consequential decisions • Prohibit use for individual profiling, risk scoring, or decisions affecting liberty without corroboration • Create a clear process for staff to flag concerns about outputs

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution can take
Prompting and querying	Are staff using prompts that could introduce or amplify demographic bias in outputs? Are outputs being taken at face value without critical assessment?	*Are prompts inadvertently introducing demographic cues (names, locations, ethnic markers) that may skew outputs?*	**High**	• Apply Tool 1 prompt guidance for all security-related queries • Require CoT prompting (ask the system to explain its reasoning step by step) for any output used to inform a decision • Train staff to recognize when outputs reflect stereotypes or incomplete information • Reprompt: after any output, ask the system to review it for potential bias and revise
Monitoring and review	Are output patterns being reviewed for signs of bias? Are certain communities flagged more often, or receiving worse outputs? What is the process when a biased or harmful output is identified?	*Are outputs disaggregated to check whether specific populations receive systematically different results?*	**Ongoing**	• Conduct a periodic review (minimum every 12 months) of output patterns for bias indicators • Log incidents where outputs appear biased and report to the vendor • Establish a staff feedback channel for reporting problematic outputs • If patterns of bias are identified and the vendor cannot address them, suspend use for that application

REFLECTION QUESTIONS

- Is there an understanding of the specific limitations and known biases of the commercial tool the institution is using?
- Are staff trained to recognize when an output may be biased, incomplete, or stereotyped?
- Are there use cases where this tool should never be used, and are those limits written down and enforced?
- Is there a human review requirement before any AI output influences a decision affecting an individual or community?
- Is there a process for staff to report problematic outputs, and does it reach someone with authority to act?
- Is the institution's use of this tool consistent across different demographic groups – or is it applied selectively?
- Have the findings here been connected to Tool 1 (prompting practices) and Tool 3 (gender and inclusion analysis)?

IMPLEMENTATION NOTES

- **Review vendor relationships annually.** Commercial AI tools change frequently. Reassess the risk assessment whenever a tool is significantly updated or use of it expands.
- **Prompting is the primary technical lever.** One cannot change what the model does, but can change how a question is asked. Apply Tool 1 prompt templates for all security-related queries.
- **Document decisions.** Record which tools the institution uses, for what purposes, under what policies, and what human review requirements apply. This is the accountability trail.
- **No tool is neutral.** Every commercial AI tool encodes assumptions from its training data. The absence of obvious bias does not mean bias is absent – it may be hidden in outputs that have not yet been scrutinized.

Tool 2B: Risk assessment for feeding institutional data into a commercial AI system

For institutions supplying their own records, data, or files to a commercial AI platform

What this tool covers

The institution is feeding its own data – incident reports, case files, surveillance records, community information, personnel data – into a commercial AI platform to generate analysis, summaries, or decision-relevant outputs. This creates two distinct layers of risk. The first is the AI model itself: its underlying biases, limitations, and training. One cannot change this – but can demand transparency from the vendor and set boundaries on what the tool is used for. The second layer is institutional data: its quality, coverage, and gaps. This layer is entirely an institution's responsibility, regardless of what the vendor's model does. This risk assessment covers both layers, focusing on what the institution can influence. Data labelling and model development – which are internal to the vendor – are covered only in terms of what an institution can demand through procurement. They have been removed as internal action items.

► **If the situation is different:** If you are not supplying any institutional data, use Tool 2A. If the institution is building or training its own AI system, use Tool 2C.

Who should use this tool	When should this tool be used?
• Programme managers and frontline leadership • Procurement officials and legal advisers • Technical teams managing data systems • Human rights and gender advisers • Oversight and compliance actors	• Before connecting institutional data to any AI platform • Before renewing or expanding a vendor contract • When institutional data sources change significantly • During annual reviews or following incidents

Instructions

For each stage in the assessment:

1. Work through the key questions and equity lens for the context.
2. Score Likelihood (1–4) and Severity (1–4). Multiply for a composite score.
3. Assign a risk level and select relevant actions.
4. Assign a responsible actor and note escalation triggers.
5. Check the red-lines box – any matching condition requires halting or restricting use.

⚠ **Critical reminder:** Institutional data is the institution's responsibility. If records underrepresent certain populations, the AI system will reproduce and amplify that underrepresentation in its outputs. Audit the data before it goes into an AI system.

Risk scoring framework

Score each stage by multiplying Likelihood × Severity.

Likelihood/ Severity ▶	1 Minimal	2 Limited	3 Significant	4 Critical	Decision guidance
1 Unlikely	Low	Low	Moderate	High	Document and monitor. No deployment block.
2 Possible	Low	Moderate	Moderate	High	Mitigations required before deployment.
3 Likely	Moderate	High	High	Unacceptable	Senior sign-off and safeguards mandatory.
4 Almost certain	High	Unacceptable	Unacceptable	Unacceptable	Do not proceed without full oversight review.

Level	Description and action
Low	Minimal foreseeable harm. Document and monitor.
Moderate	Risks requiring mitigation before or during deployment.
High	Significant risks requiring safeguards and senior sign-off.
Unacceptable	Do not deploy. Full oversight review required. Check red lines.

Red lines – When to stop

If any of the following apply, do not proceed with feeding institutional data into this AI system for this application.

AUTOMATIC UNACCEPTABLE RATING – HALT DEPLOYMENT

If any of the following conditions apply, classify as Unacceptable regardless of other scores.

- The vendor cannot provide documentation explaining how the system makes decisions.
- The system will replace – not inform – human judgement in high-stakes decisions (arrest, detention, lethal force).
- No contestation or redress mechanism exists for individuals affected by AI-driven decisions.
- The system has not been tested against representative populations in the intended deployment context.
- Deployment would violate applicable legal prohibitions (e.g. EU AI Act prohibitions on real-time biometric surveillance without judicial authorization).

Risk assessment matrix

Complete one row per stage. The equity and intersectionality lens (purple) surfaces risks to specific populations that are often missed in standard data assessments.

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution can take
Problem definition	Is feeding institutional data into this AI system the right approach? Could the framing of the problem reinforce existing assumptions? What institutional data will be used, and for what decisions?	*Whose problem is being solved? Which populations are affected by the outputs? Were they consulted?*	**Moderate**	• Document exactly what data will be fed in and what outputs will be used for • Identify all populations whose information may be in the data, and who may be affected by outputs • Review the problem framing with human rights and gender advisers before proceeding • Confirm this use is lawful and consistent with data protection obligations
Data quality and coverage	Who is represented in the data the institution is feeding in? Who may be absent, underrepresented, or misrepresented? Could gaps in the institutional data produce systematically worse outputs for specific groups?	*Are women, minorities, displaced populations, or rural communities underrepresented in the institutional data? Could their absence cause them to be overlooked or misclassified?*	**High**	• Audit the data for demographic coverage before feeding it into any AI system • Document known gaps and their likely impact on outputs • Do not treat absence of data as absence of risk – invisible populations face higher, not lower, risk • If key populations are absent from your data, do not use AI outputs as a basis for decisions affecting them without additional verification

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution can take
Procurement and vendor terms	When the institution's data is processed by the vendor's system, how is it handled, stored, and protected? Can the vendor explain how the underlying model interacts with the institutional data?	*Has the vendor tested how its system performs when processing data about populations comparable to those in the institution's records?*	**High**	• Require the vendor to document how institutional data is processed, stored, and protected • Include in any contract: audit rights, incident reporting obligations, and the right to suspend use if bias is identified • Specify that the vendor must disclose any known limitations of the system when applied to the types of data the institution will supply • Do not feed sensitive personal data – especially data about vulnerable populations – into a system without explicit contractual data protections
Deployment and use policy	Will security decisions be made or informed by AI-generated outputs from the institutional data? Who reviews those outputs? What accountability exists when an AI output based on institutional data leads to harm?	*Which communities will be most exposed to errors in AI outputs generated from the data? Is there a redress mechanism for individuals whose data was processed?*	**High – check red lines**	• Establish a written policy on which decisions may be informed by AI outputs, and which require human review only • Require HITL for all consequential decisions – particularly those affecting individual liberty, protection, or access to services • Create an accessible process for individuals to contest decisions made using AI-processed versions of their data • Document the accountability chain from AI output to final decision-maker

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution can take
Monitoring and review	Are outputs being reviewed for patterns of bias? Are certain populations flagged more often or receiving systematically different results? Is there a check of whether the AI outputs remain accurate as institutional data changes over time?	*Are monitoring metrics disaggregated by population group? Is there a mechanism for affected communities to flag concerns about how their information is being used?*	**On-going**	• Conduct quarterly review of output patterns disaggregated by relevant demographic dimensions • Establish a frontline staff channel to report suspected bias in outputs • Set reassessment triggers: significant changes to institutional data, vendor system updates, or incidents • Cross-reference output patterns with Tool 1 prompt guidance to check whether querying practices are contributing to bias

REFLECTION QUESTIONS

- Is it known which populations are represented – and which are absent – in the institutional data being fed into the AI system?
- Has the institution audited its data for demographic gaps before using it as an input to AI-generated analysis?
- Could the AI system's outputs be used to make decisions about individuals whose data is in the system? Is there a redress mechanism?
- Does the vendor contract include data protection, explicability, audit rights, and incident reporting obligations?
- Are staff applying critical judgement to outputs, or treating them as authoritative?
- Is there monitoring of whether certain populations receive systematically different outputs?
- Have the findings here been connected to Tool 1 (prompting practices) and Tool 3 (gender and inclusion analysis)?

IMPLEMENTATION NOTES

- **An institution's data layer is its accountability.** The vendor is responsible for the model. The institution is responsible for the data it supplies, and the decisions made using AI outputs from that data.
- **Data gaps are not neutral.** An absence of data about a population is not the same as low risk for that population. In many cases, the opposite is true: invisible populations face higher risk from AI systems that cannot account for them.
- **Treat this as a living assessment.** Institutional data changes. The vendor's system may be updated. Reassess whenever institutional data or the vendor's system changes significantly, and at a minimum annually.
- **Connect to Tool 1.** The way the institution queries the AI system is itself a bias risk. Apply Tool 1 prompt templates whenever using the system for security-related analysis.

Tool 2C: Risk assessment for institutions building or training their own AI system

For institutions developing, commissioning, or training a custom AI system on institutional data

What this tool covers

The institution is building, has built, or has commissioned a custom AI or machine-learning system – one that is trained on data the institution has collected, and will generate outputs used to inform security decisions. This is the highest-responsibility scenario. The institution controls the full stack: how data is collected, how it is labelled, how the model is designed and tested, how it is deployed, and how it is monitored. There is no vendor to point to. Every bias risk at every life-cycle stage is reliant on the institution to identify, mitigate, and own. This risk assessment covers all seven life-cycle stages. Nothing has been removed – all stages are actionable, and all are the institution's responsibility.

► **If the situation is different:** If you are using an existing commercial tool without supplying data, use Tool 2A. If feeding institutional data into a commercial platform, use Tool 2B.

Who should use this tool	When should this tool be used?
• Technical developers and data teams • Programme managers and frontline leadership • Procurement officials (if outsourcing development) • Human rights, gender, and legal advisers • Oversight bodies	• Before development begins • At each stage of the development life cycle • Before deployment • During major system modifications • Following incidents or at minimum annually

Instructions

For each stage in the assessment:

1. Work through the key questions and equity lens for the context.
2. Score Likelihood (1–4) and Severity (1–4). Multiply for a composite score.
3. Assign a risk level and select or adapt actions from the illustrative list.
4. Assign a responsible actor and note escalation triggers.
5. Check the red-lines box – any matching condition requires halting development or deployment.

⚠ **Full accountability applies.** When the institution builds a custom AI system, it is responsible for all bias risks at all life-cycle stages. There is no minimum standard of 'good enough' – the NIST AI RMF and EU AI Act impose ongoing obligations that do not end at deployment.

Risk scoring framework

Score each life-cycle stage by multiplying Likelihood × Severity.

Likelihood/ Severity ▶	1 Minimal	2 Limited	3 Significant	4 Critical	Decision guidance
1 Unlikely	Low	Low	Moderate	High	Document and monitor. No deployment block.
2 Possible	Low	Moderate	Moderate	High	Mitigations required before deployment.
3 Likely	Moderate	High	High	Unacceptable	Senior sign-off and safeguards mandatory.
4 Almost certain	High	Unacceptable	Unacceptable	Unacceptable	Do not proceed without full oversight review.

Level	Description and action
Low	Minimal foreseeable harm. Document and monitor.
Moderate	Risks requiring mitigation before or during deployment.
High	Significant risks requiring safeguards and senior sign-off.
Unacceptable	Do not deploy. Full oversight review required. Check red lines.

Red lines – When development or deployment must stop

If any of the following apply at any point in the life cycle, development or deployment must be halted and a full review conducted before proceeding.

AUTOMATIC UNACCEPTABLE RATING – HALT DEPLOYMENT

If any of the following conditions apply, classify as Unacceptable regardless of other scores.

- The vendor cannot provide documentation explaining how the system makes decisions.
- The system will replace – not inform – human judgement in high-stakes decisions (arrest, detention, lethal force).
- No contestation or redress mechanism exists for individuals affected by AI-driven decisions.
- The system has not been tested against representative populations in the intended deployment context.
- Deployment would violate applicable legal prohibitions (e.g. EU AI Act prohibitions on real-time biometric surveillance without judicial authorization).

Risk assessment matrix

Complete one row per life-cycle stage. All seven stages apply. The equity and intersectionality lens (purple) surfaces risks to specific populations that technical assessments typically miss. Actions listed are illustrative – adapt to the institutional context.

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution must take
Problem definition	Is building a custom AI system necessary? Could the problem be addressed another way? Have affected communities been consulted in defining the problem the system will address?	*Whose problem is being solved and whose problem is being overlooked? Who was in the room when the system was scoped?*	**Moderate**	• Document the need and all alternatives considered before committing to development • Map all populations who will be affected by the system's outputs before scoping begins • Conduct a gender and inclusion review of the problem framing (see Tool 3) • Confirm legal authority and human rights compatibility before proceeding
Data collection	Who is represented in the data being collected to train this system? Who is absent, and why? Could the act of collecting data itself create risks for specific communities?	*Are women, minorities, displaced populations, and rural communities represented in the training data? Could data collection coerce participation from vulnerable groups?*	**High**	• Audit data sources for demographic coverage before any training begins • Document known gaps and their expected impact on system outputs • Treat absence of data about a population as a risk requiring mitigation, not a neutral condition • Ensure data collection is genuinely voluntary for vulnerable populations – institutional authority must not coerce participation • Apply Tool 3 inclusion analysis to the data collection design

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution must take
Data labelling	Who is defining the categories and labels that will train this system? Could those choices encode cultural, institutional, or demographic bias? Are labelling teams diverse enough to surface blind spots?	*Do the labelling categories reflect a single institutional or cultural viewpoint? Were people from affected communities involved in defining them?*	**High**	• Review all labelling guidelines with a diverse group including gender and human rights advisers before labelling begins • Document who defined each category and record any dissenting views • Test for differential labelling rates across demographic groups • Where possible, include members of affected communities in labelling review panels
Model development	Could proxy variables (geography, language, name, network connections) produce unequal outcomes for specific groups? Has the system been tested for fairness across all relevant demographic subgroups, including intersectional combinations?	*Which groups may be systematically disadvantaged by the model's predictions? Has intersectional testing been conducted – not just single-dimension tests?*	**High**	• Conduct fairness testing across demographic subgroups before deploying AI – encompassing intersectional dimensions. • Identify and audit all proxy variables –justify each one, or remove it • Document model limitations, confidence thresholds, and known failure modes • Have outputs reviewed by human rights and gender advisers before deployment approval

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution must take
Procurement (if outsourcing development)	If the system is being built by an external contractor, can they document their development methodology, bias testing, and known limitations? Do contract terms give the institution audit rights and the ability to suspend or terminate?	*Has the contractor tested the system against populations comparable to those in the context?*	**High**	• Require full documentation of development methodology, training data sources, and bias test results as contract deliverables • Include audit rights, incident reporting obligations, and suspension/termination clauses in all contracts • Specify that bias testing must cover the demographic populations relevant to the context • Retain ownership of all institutional data used in development
Deployment	Will security decisions be made or informed by this system? Who is accountable for outcomes when the system contributes to a decision that causes harm? Is there a redress mechanism for individuals affected?	*Which communities will bear the highest exposure to errors? Are redress mechanisms accessible to those communities, including in relevant languages?*	**High – check red lines**	• Establish a written policy: which decisions may be informed by the system, and which require human review only • Require HITL for all decisions affecting individual liberty, protection, or access to services • Create an accessible, multilanguage contestation process for individuals affected by the system's outputs • Document the full accountability chain, from system output to final decision-maker

Stage	Key questions	Equity and intersectionality lens	Risk level	Actions the institution must take
Monitoring	How will bias and harm be detected, reported, and addressed over time? Who has authority to suspend the system? Is this system learning from its own outputs in ways that could compound bias over time?	*Are monitoring metrics disaggregated by gender, ethnicity, and other relevant dimensions? Is there a community-facing feedback channel?*	**On-going**	• Conduct quarterly output audits disaggregated by relevant demographic dimensions • Establish a frontline feedback channel for staff to flag bias concerns • Set reassessment triggers: any system update, significant incident, or 12-month cycle at minimum • For any adaptive or learning component, audit retraining data sets for demographic shift before every retraining cycle • Cross-reference with Tool 1 prompt templates to assess ongoing query quality

REFLECTION QUESTIONS

- Has the institution audited all training data for demographic coverage, including intersectional combinations of identity?
- Were affected communities consulted in the problem definition and data collection stages – with genuine power to change outcomes?
- Has the system been tested for fairness across relevant demographic subgroups, including those with intersectional identities, before deployment?
- Is there a HITL requirement for all decisions affecting individual liberty, protection, or access to services?
- Is there an accessible, multilanguage redress mechanism for individuals affected by the system's outputs?
- For any adaptive or learning component, is the system being monitored for feedback loop effects that could compound bias over time?
- Have findings been connected to Tool 1 (prompting practices) and Tool 3 (gender and inclusion analysis)?
- Is this risk assessment being treated as a living document, updated at every significant system change?

IMPLEMENTATION NOTES

- **An institution owns every stage.** There is no vendor to share the responsibility. Every bias risk identified in this assessment must be addressed by the institution. Document what has been found, what was done about it, and who made each decision.
- **Intersectional testing is necessary.** Testing each demographic dimension independently misses the compounded disadvantages faced by people at multiple intersections.
- **Watch for feedback loops.** Any component that learns from outputs it creates requires increased bias monitoring. Bias compounds more quickly in adaptive systems.
- **Community consultation is a quality control mechanism, not just a process requirement.** Affected communities will identify failure modes that internal teams miss. Build consultation in at problem definition, data collection, and post-deployment review stages.
- **Connect to Tool 3.** Before deployment, complete the inclusive security and gender analysis (Tool 3). Findings from that assessment should inform risk scoring at the data collection, model development, and deployment stages.

Tool 3: Inclusive security and gender analysis template

Intersectional gender analysis for security institutions' use of AI

Purpose

AI systems may have different impacts on populations based on gender norms, sex, and intersecting identities, including age, ethnicity, disability, language, location, socioeconomic status, religion, displacement status, and political affiliation.

This tool supports security institutions in assessing how AI systems may affect the different security needs, rights, participation, and protection of women, men, girls, boys, and people with diverse intersecting identities. It is grounded in the UN Women/UNU Macau WPS framework (2024), which identifies four categories of gender bias in AI (discrimination, stereotyping, exclusion, and insecurity), and is aligned with the WPS agenda (UN Security Council Resolution 1325 and subsequent resolutions).

This template incorporates four dimensions not always addressed in standard gender analysis.

- TFGBV as a distinct AI-created risk, not only a missed protection.
- Feedback loops and bias compounding over time in adaptive AI systems.
- Chilling effects on civic participation caused by AI surveillance and flagging.
- Compound intersectional identity analysis, beyond single-category checklists.

Use alongside Tool 1 (prompting guide) and Tool 2 (life-cycle risk assessment).

Who should use this tool	When should this tool be used?
• Gender advisers • Human rights specialists • Security planners • Technical teams • Civil society partners • Oversight actors • Procurement officials	• Prior to procurement or deployment • During major system updates • During monitoring and evaluation • Following incidents affecting specific populations • When commissioning new AI use cases

Four key concepts for this assessment

The following concepts address gender dimensions of AI-related harms that are frequently missed in standard assessments. Practitioners should be familiar with each before completing the tool.

1. TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE AS AN AI-CREATED RISK

- **What it is:** AI systems deployed by security institutions may not only fail to protect the populations they are meant to shelter, but can become instruments of harm against them.
- **Impact:** In areas of conflict, AI-facilitated biometrics have enabled the monitoring of LGBTQI+ individuals and women human rights defenders (WHRDs), and AI-driven disinformation can discredit these individuals.
- **What to assess:** Could this system's outputs be used – intentionally or through misuse – to identify, surveil, or harm specific individuals or groups? Does the institution have controls to prevent this?

2. FEEDBACK LOOPS AND BIAS COMPOUNDING OVER TIME

- **What it is:** Biased AI outputs influence the data used to retrain the system, which becomes progressively more biased against already-flagged communities. Small initial errors escalate into large systematic distortions.
- **Impact:** Biases are amplified more in human–AI feedback loops than in human–human interactions. Under-resourcing of certain communities reduces data diversity and reinforces the future biases.
- **What to assess:** Does this system learn or adapt based on its own outputs? If so, which communities could be progressively disadvantaged over time?

3. CHILLING EFFECTS ON CIVIC PARTICIPATION

- **What it is:** Surveillance or flagging by AI systems can cause targeted communities to self-censor and withdraw from legitimate public activity – reducing political participation, civic engagement, and access to protection.
- **Impact:** WHRDs often face online violence and offline harms, and disproportionate surveillance, that impact their work.
- **What to assess:** Could communities subject to this system's monitoring reduce their legitimate civic activity as a result? Has this been assessed before deployment?

4. INTERSECTIONALITY

- **What it is:** Listing identity categories is not intersectional analysis. People facing multiple and overlapping disadvantages experience qualitatively different harms beyond what any single category suggests (e.g. a displaced elderly woman with a disability). These compounded harms are often invisible to systems that test each dimension independently.
- **Impact:**Combined intersectional identities increase disadvantages and discrimination versus single identities. These overlapping attributes are rarely captured in AI data sets, making those with several at-risk identities more vulnerable.
- **What to assess:** Which combinations of identities in this context create compounded disadvantages, discrimination, or exclusion? Has the system been tested against those specific combinations – not just each dimension separately?

Instructions

Complete this assessment as part of design or procurement, prior to deployment, and during periodic monitoring and evaluation. The assessment should be:

- owned internally by a named gender or human rights adviser;
- informed by direct consultation with affected communities – documented in the consultation log (see p. 109);
- reviewed jointly by gender, human rights, legal, and oversight actors before sign-off;
- updated whenever the system, its use, or the operating environment changes significantly.

⚠ **On intersectionality.** Listing identity categories is not analysis. The assessment must examine how categories overlap and compound to increase an individual's risk of discrimination and exclusion.

Conducting intersectional analysis requires moving from category listing to compound risk assessment. A practical approach: begin by identifying the two or three identity dimensions most relevant to the specific AI use case (for example, displacement status, gender, and linguistic minority status for a border screening system).

For each intersection of those dimensions, ask these questions.

- Does the system perform differently for this group than for others?
- Does the data used to train or evaluate the system adequately represent this group?
- Are the harms of false positives and false negatives distributed equally across groups, or does one intersection bear disproportionate risk?

A system may perform well for women as a whole while performing poorly for displaced women who are linguistic minorities – aggregate gender evaluation will not reveal this. Document the analysis and the groups examined, and flag intersections where data is insufficient for meaningful evaluation as a governance risk requiring remediation before deployment.

Part I: Context analysis

Complete this section before proceeding.

Question	Findings/notes
1. What security problem is being addressed, and is AI the appropriate tool?	
2. What is the intended use, scope, and context?	
3. What legal, human rights, and WPS obligations apply? (Include UNSCR 1325/WPS National Action Plan obligations where applicable.)	
4. Whose security needs, rights, participation, or protection may be affected? *Tick all that apply. For each ticked group, note intersecting disadvantages in the 'findings' column (e.g. 'women + refugee + linguistic minority = compounded exclusion risk').*	**Population group** ☐ Women ☐ Men ☐ Girls (age-specific risks) ☐ Boys (age-specific risks) ☐ Ethnic minorities ☐ Religious minorities ☐ Linguistic minorities ☐ LGBTQI+ populations ☐ Persons with disabilities ☐ Migrants/refugees ☐ Older persons ☐ Stateless persons ☐ WHRDs ☐ Other: _____
5. Which combinations of identities in this context create a compounded disadvantage? *List specific intersections, e.g. 'refugee women from linguistic minorities in rural areas'. Do not treat categories as independent.*	
6. How do affected communities define their own security needs? Do those definitions differ from the institution's framing? *A divergence here is itself a risk indicator.*	

Part II: Inclusion assessment

For each question, record findings specific to the context. Rows highlighted in red address TFGBV and its chilling effects – these require immediate escalation if findings are positive. The amber row addresses feedback loops in adaptive systems.

Inclusion question	Illustrative example (adapt to context)	Analysis guidance
Who benefits from this AI system?	*Personnel using the system gain efficiency; dominant-language speakers may receive more accurate outputs.*	Consider whether benefits are equitably distributed. Are some groups systematically more likely to benefit than others?
Who may be excluded or harmed?	*Linguistic minorities may be misclassified; rural or offline populations may be invisible to the system.*	Identify exclusion mechanisms: data absence, language gaps, geographic blind spots, proxy variables.
What populations are underrepresented in available data, and why?	*Women in informal settlements; stateless persons lacking documentation; non-dominant-language speakers.*	Also ask: do affected communities have the power to refuse data collection, or is participation de facto coerced by institutional authority? Coerced data collection is itself a risk.
Are community perspectives included in the design or deployment of this system?	*Civil society consulted on risk scenarios; affected community members participated in testing.*	Record who was consulted and when in the consultation log (below). If no consultation occurred, flag as a gap requiring remediation before deployment.
Could the system enable unequal or discriminatory surveillance – or be weaponized against the populations it is meant to protect?	*AI biometrics enabling the targeting of LGBTQI+ individuals or WHRDs.* *Facial scanning deployed without consent near communities already experiencing state violence.*	This question addresses AI as a vector of harm, not just a missed protection. In security contexts, this risk is distinct and must be assessed separately. Cross-reference with TFGBV risk row in Part III.
Could the system produce a chilling effect – causing targeted communities to self-censor, withdraw from civic space, or disengage from legitimate political activity?	*AI monitoring of online speech causes minority community members to stop posting about political issues.* *Women community leaders reduce public activity after being flagged by a risk screening system.*	Online violence and, often resulting in offline harms, affects the work of WHRDs. Chilling effects are not hypothetical situations, but are known harm pathways.

Inclusion question	Illustrative example (adapt to context)	Analysis guidance
Could protection be biased or unequally provided?	*Threats against women community leaders may be underdetected if their risk indicators differ from male profiles.*	Consider whether the system's definition of 'risk' or 'threat' reflects the full range of security experiences across populations.
Could this system's outputs, over time, influence the data used to train or retrain it – compounding bias against specific populations (feedback loop)?	*System flags certain communities more often; flagged community data becomes overrepresented in 'risk' training sets; system becomes progressively more biased against that community.*	Feedback loops that are human–AI amplify bias much more than human–human. Practitioners are often unaware of the escalation. Flag any adaptive or learning system for heightened monitoring.

► *Findings from rows highlighted in red or amber require immediate escalation to the risk matrix (Part III) and to Tool 2 (life-cycle risk assessment) at the deployment and monitoring stages. Rows highlighted in red also require sign-off by the oversight actor before any deployment proceeds.*

Consultation log

Record all consultations conducted to inform this assessment. If no consultation was conducted, document the reason and a plan for future engagement.

Participant	Date	Method	Key findings	How incorporated

Part III: Risk and mitigation matrix

This edition adds four new illustrative rows: WHRDs, girls (age-specific risks), a compounded intersectional identity row (e.g. a woman + refugee + linguistic minority), and an adaptive/ learning AI systems feedback loop row. Modify, remove, or add rows based on the context.

Severity guide

High = significant rights or protection impact, requires immediate mitigation
Moderate = risk requiring mitigation before deployment

Note on scoring alignment. The severity tiers in this template (High/Moderate) correspond to the upper two levels of the 1–4 risk scoring scale used in Tools 2A–2C. High severity in this template is equivalent to a score of 3–4 in the Tool 2 framework; Moderate severity is equivalent to a score of 1–2. When synthesizing results across annexes, practitioners should treat High findings from this template as equivalent to elevated-risk scores in Tool 2 and apply the same escalation and documentation requirements.

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Severity	Mitigation actions and responsible actor
Women community leaders	Threats under-detected; risk profile modelled on male political actor norms	***Exclusion/ insecurity***	Right to protection; security of person; WPS UNSCR 1325	Women's civil society groups; local security networks	**High**	• Integrate local reporting networks alongside AI outputs • Apply explicit gender lens in prompting (Tool 1, Scenario 1) • Audit threat detection rates disaggregated by gender and role **Responsible: gender adviser + security planner**

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Severity	Mitigation actions and responsible actor
Women human rights defenders	AI surveillance systems used to monitor, track, or generate information enabling targeting of WHRDs	***Insecurity/ discrimination***	Right to privacy; freedom of expression; protection from surveillance	WHRD networks; digital rights organizations; legal aid	**High**	• Conduct explicit TFGBV risk assessment before any surveillance-capable deployment • Prohibit use of AI outputs to profile or monitor civil society actors • Establish independent oversight mechanism for surveillance use cases • Ensure redress mechanism is accessible to WHRDs **Responsible: human rights adviser + legal adviser + oversight body**
Linguistic minorities	Misclassification of speech or intent due to language model gaps	***Discrimination/ stereotyping***	Right to fair treatment; non-discrimination	Community representatives; language specialists	**High**	• Expand language data sets before deployment • Set explicit confidence thresholds for non-dominant language outputs • Require human review for all low-confidence outputs **Responsible: technical developer + human rights adviser**

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Severity	Mitigation actions and responsible actor
Girls (age-specific risks)	Biometric data collection near schools; behavioural flagging misclassifying adolescent activity; surveillance increasing gender-based violence risk	***Exclusion/ insecurity***	Right to privacy; child protection obligations; freedom from surveillance	Child protection organizations; youth-led groups; parents/ caregivers	**High**	• Prohibit biometric collection in or near educational settings without explicit consent • Audit behavioural flagging for adolescent false-positive rates disaggregated by sex • Apply heightened data minimization standards for all data involving minors **Responsible: human rights/ child protection adviser + legal adviser**
Women + refugee + linguistic minority (compounded identity)	Compounded exclusion: absence from data + language misclassification + lack of documentation = systematic invisibility and elevated false-flagging risk	***Exclusion + discrimination (compound)***	Non-refoulement; non-discrimination; equal protection; right to asylum	UNHCR; refugee-led women's organizations; legal aid in relevant languages	**High**	• Treat absence from data as a risk signal requiring affirmative inclusion, not a neutral gap • Require compound identity testing in vendor bias audits • Create complaint mechanism in multiple languages with non-documentation pathways **Responsible: gender adviser + procurement official + legal adviser**

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Severity	Mitigation actions and responsible actor
LGBTQI+ individuals	Increased surveillance; privacy violations; chilling effect on expression; AI-enabled targeting in hostile legal environments	***Discrimination/ insecurity***	Right to privacy; freedom of expression; protection from harm	LGBTQI+ civil society; human rights organizations	**High**	• Audit for proxy variables that may expose LGBTQI+ individuals • Ensure data cannot be used for identification in adverse legal contexts • Do not retain network or location data that could enable targeting **Responsible: human rights adviser + legal adviser + oversight body**
Rural and remote populations	Systematic invisibility in data; AI protection and risk assessments do not apply equally	***Exclusion***	Equal access to protection; right to security	Local CSOs; community leaders; field staff	**Moderate**	• Include local and informal reporting in data collection • Disaggregate outputs by geographic area • Flag rural data absence explicitly in system documentation **Responsible: programme manager + field teams**

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Severity	Mitigation actions and responsible actor
Migrants and refugees	Disproportionate flagging; documentation gaps amplified as risk signals	***Discrimination/ exclusion***	Non-refoulement; right to asylum; non-discrimination	UNHCR; refugee-led organizations; legal aid providers	**High**	• Separate displacement status from risk indicators • Require vendor to demonstrate model performance on refugee populations • Accessible complaint mechanism in relevant languages **Responsible: procurement official + legal adviser**
Adaptive/ learning AI systems (all populations)	Feedback loop: biased outputs influence retraining data, progressively compounding bias against already-flagged communities over time	***All types – compounding over time***	All rights affected by the underlying system	Technical reviewers; affected community monitors; independent auditors	**High – monitor continuously**	• Flag any system with learning or adaptive capacity for heightened monitoring • Audit retraining data sets for demographic shift every 6 months • Establish community-facing feedback channel to surface emerging bias • Mandate independent audit before any retraining cycle **Responsible: technical developer + programme manager + oversight body**

Population group	Potential risk	Bias type	Right/ security need affected	Consultation required	Seve- rity	Mitigation actions and responsible actor
[Add population based on context]						

Review and sign-off

Review jointly before finalizing. Any ‘no’ or ‘unknown’ answer must be documented with a remediation plan. Disagreements between reviewers must be escalated to oversight.

REFLECTION QUESTIONS

Review jointly by lead assessor, gender adviser, and oversight actor. Any ‘no’ or ‘unknown’ answer must be documented and a remediation plan recorded before sign-off.

1. Is sex-, age-, and diversity-disaggregated data available and appropriately used? If not, how will this gap be addressed?
2. Which combinations of identities create compounded disadvantages (discrimination, exclusion) in this context – and has the system been tested against those combinations?
3. Are women, men, girls, boys, and gender-diverse people affected differently? Have these differences been assessed against the specific context?
4. Could the system itself become an instrument of harm – enabling surveillance, targeting, or TFGBV against populations it nominally protects?
5. Has a chilling effect assessment been conducted? Could communities reduce legitimate civic activity or participation because of this system?
6. Has a feedback loop assessment been conducted for any adaptive or learning component of this system?
7. Could the system affect access to security, justice, services, mobility, or protection for any group?
8. Can impacted communities validate the assumptions underpinning this assessment? Have they been given a genuine opportunity to do so – with power to change outcomes?
9. Are risk mitigation measures in place, adequately resourced, and linked to institutional action plans?
10. How do findings connect to Tool 1 (prompting practices) and Tool 2 (life-cycle risk assessment)?

Name	Role	Date / Signature
	Lead assessor	
	Gender/human rights adviser	
	Oversight actor	

IMPLEMENTATION NOTES

- **Treat this as iterative, not a one-time exercise.** Security environments, populations, and AI systems all evolve. Reassess whenever the system, its context, or the population profile changes.
- **Consultation is necessary.** Qualitative evaluation should be carried out with the implicated communities for AI systems affecting vulnerable groups. Document individuals consulted, sentiments expressed, and its impact on the assessment.
- **Disaggregate data and findings.** Aggregate data hides the groups most at risk. Require sex-, age-, and diversity-disaggregated data from the outset, and audit outputs against disaggregated outcomes.
- **Assess AI as a potential instrument of harm.** Do not ask only whether the system protects all groups equally. Also ask whether the system could be used to harm the groups it nominally protects – particularly through surveillance, targeting, or TFGBV enablement.[47]
- **Monitor feedback loops in adaptive systems.** Any system that learns from its own outputs must be subject to heightened bias monitoring.
- **Anchor to WPS obligations.** Where applicable, findings should reference UNSCR 1325 and relevant National Action Plan obligations. Gender analysis is not optional for security sector actors; all countries and organizations have commitments under the international WPS agenda.
- **Results must drive action.** Findings should guide safeguards, procurement requirements, monitoring indicators, and institutional action plans – not sit in a file.

Additional tools

Tool	Objectives	Description
AI Fairness 360	Fairness · Audit	Measure and mitigate bias while training
Fairlearn	Fairness · Mitigation	Assess and reduce unfair outcomes while building models
SHAP	Explicability	Explain which inputs drove a prediction – and reveal proxy bias
InterpretML (EBM)	Transparency · Explicability	Train inherently interpretable models that auditors can read directly
Great Expectations	Data quality · Governance	Validate data sets before training – schema, distributions, and quality checks
What-If Tool	Explicability · Audit	Probe counterfactuals interactively: what happens if this input changes?
Evidently	Drift monitoring	Monitor data drift and model performance in production
Alibi Detect	Drift monitoring · Audit	Detect drift, outliers, and adversarial inputs at runtime
MLflow	Traceability · Governance	Track model versions and decision lineage so any output can be traced
whylogs	Drift monitoring · Transparency	Log data profiles over time to catch distribution shifts early
NannyML	Drift monitoring · Accountability	Estimate performance degradation when ground truth arrives late
Giskard	Fairness · Audit · Governance	Automated test suites for bias, robustness, and security in production

Endnotes

1 **Stockholm International Peace Research Institute (SIPRI).** 2024. Artificial Intelligence, Bias, and Military Decision-Making. Stockholm: SIPRI; L. Blanchard and S. Bruun. 2024. Bias in Military AI Systems. Stockholm: SIPRI; L. Blanchard and S. Bruun. 2025. Emerging Risks in AI-Enabled Warfare. Stockholm: SIPRI.

2 **Andrew D.** Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. 'Fairness and Abstraction in Sociotechnical Systems', in Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, January 29-31, 2019). New York: Association for Computing Machinery, New York, 59–68.

3 **Kate Crawford.** 2021. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven, CT: Yale University Press.

4 **DCAF – Geneva Centre for Security Sector Governance.** 2019. 'Security Sector Governance', SSR Backgrounder Series. Geneva: DCAF.

5 **Kimberlé Crenshaw.** 1989. 'Demarginalizing the Intersection of Race and Sex', University of Chicago Legal Forum 1989(1): 139–167.

6 **UNIDIR Security and Technology Programme.** 2025. Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action. Geneva: UNIDIR.

7 **Future of Life Institute.** 2026. International AI Safety Report. Boston, MA: Future of Life Institute.

8 **National Institute of Standards and Technology (NIST).** 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: US Department of Commerce, nist.gov/itl/ai-risk-management-framework.

9 **Solon Barocas and Andrew D.** Selbst. 2016. 'Big Data's Disparate Impact', California Law Review 104(3): 671–732.

10 **National Institute of Standards and Technology (NIST).** 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg, MD: NIST.

11 **Joy Buolamwini and Timnit Gebru.** 2018. 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of Machine Learning Research 81: 1–15. 2018.

12 **Clare Garvie, Alvaro Bedoya, and Jonathan Frankle.** 2019. The Perpetual Line-Up: Unregulated Police Face Recognition in America. Washington, DC: Georgetown Law Center on Privacy and Technology.

13 **Kate Crawford.** 2021. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven, CT: Yale University Press.

14 **Europol.** 2025. AI Bias in Law Enforcement: A Practical Guide, Europol Innovation Lab Observatory Report, Luxembourg: Publications Office of the European Union.

15 **Steven Feldstein.** 2019. The Global Expansion of AI Surveillance. Washington, DC: Carnegie Endowment for International Peace.

16 **Human Rights Watch.** 2021. Uganda: Events of 2021. New York: Human Rights Watch.

17 **International Committee of the Red Cross (ICRC).** 2006. Guide to the Legal Review of New Weapons, Means and Methods of Warfare. Geneva: ICRC

18 **David Danks and Alex John London.** 2017. 'Algorithmic Bias in Autonomous Systems', in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (Melbourne, August 19-25, 2017). 4691-4697.

19 **European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice.** 2024. Eurodac 2023 Annual Report. Brussels: European Commission.

20 **Sarah Brayne.** 2017. 'Big Data Surveillance: The Case of Policing', American Sociological Review 82(5): 977–1008.

21 **Raja Parasuraman and Dietrich Manzey.** 2010. 'Complacency and Bias in Human Use of Automation', Human Factors 52(3): 381–410.

22 **Danielle Ensign, Sorelle A.** Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. 'Runaway Feedback Loops in Predictive Policing', in Proceedings of the Conference on Fairness, Accountability, and Transparency (New York, February 23-24, 2018). 81, 160-171.

23 **David Danks and Alex John London.** 2017. 'Algorithmic Bias in Autonomous Systems', in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (Melbourne, August 19-25, 2017). 4691-4697.

24 **David Danks and Alex John London.** 2017. 'Algorithmic Bias in Autonomous Systems', in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (Melbourne, August 19-25, 2017). 4691-4697.

25 **UNESCO.** 2024. Guidance for the Governance of Generative AI. Paris: UNESCO.

26 **OECD (Organisation for Economic Co-operation and Development).** 2019. OECD Principles on Artificial Intelligence. Paris: OECD.

27 **Council of Europe.** 2024. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Strasbourg: Council of Europe.

28 **UN Women.** 2015. Preventing Conflict, Transforming Justice, Securing the Peace: A Global Study on the Implementation of United Nations Security Council Resolution 1325. New York: UN Women.

29 **Defensoría del Pueblo (Colombia).** 2020. Informes de Alerta Temprana. Bogotá: Defensoría del Pueblo.

30 **UNIDIR Security and Technology Programme.** 2025. Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action. Geneva: UNIDIR.

31 **International Committee of the Red Cross.** 2019. Artificial Intelligence and Machine Learning in Armed Conflict: A Humanitarian Perspective. Geneva: ICRC; Blanchard and Bruun, n 1.

32 **Bridges v South Wales Police [2020] EWCA Civ 1058 (Court of Appeal, England and Wales).**

33 **National Institute of Standards and Technology (NIST).** 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg, MD: NIST.

34 **Raja Parasuraman and Dietrich Manzey.** 2010. 'Complacency and Bias in Human Use of Automation', Human Factors 52(3): 381–410.

35 **UN High Commissioner for Human Rights.** 2021. 'The right to privacy in the digital age.' New York: Human Rights Council, docs.un.org/en/A/HRC/48/31.

36 **International Committee of the Red Cross.** 2023. 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977.' Geneva: ICRC.

37 **UNIDIR.** 2025. Artificial Intelligence and Military Decision-Making. Geneva: UNIDIR.

38 **Jenna Burrell.** 2016. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', Big Data & Society 3(1).

39 **David Danks and Alex John London.** 2017. 'Algorithmic Bias in Autonomous Systems', in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (Melbourne, August 19-25, 2017). 4691-4697.

40 **National Institute of Standards and Technology (NIST).** 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg, MD: NIST.

41 **OECD (Organisation for Economic Co-operation and Development).** 2019. OECD Principles on Artificial Intelligence. Paris: OECD.

42 **National Institute of Standards and Technology (NIST).** 2019. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Gaithersburg, MD: NIST.

43 **UNESCO.** 2024. Guidance for the Governance of Generative AI. Paris: UNESCO.

44 **Council of Europe.** 2024. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Strasbourg: Council of Europe.

45 **UN Women.** 2015. Preventing Conflict, Transforming Justice, Securing the Peace: A Global Study on the Implementation of United Nations Security Council Resolution 1325. New York: UN Women.

46 **Organization for Security and Co-operation in Europe (OSCE)/Office for Democratic Institutions and Human Rights (ODIHR).** 2020. Guidance on Gender-Sensitive Approaches to Early Warning and Early Action. Warsaw: OSCE/ODIHR.

47 **UN Women.** 2015. Preventing Conflict, Transforming Justice, Securing the Peace: A Global Study on the Implementation of United Nations Security Council Resolution 1325. New York: UN Women.

DCAF – Geneva Centre for Security Sector Governance

Maison de la Paix
Chemin Eugène-Rigot 2E,
1202 Geneva, Switzerland
Tel: +41 22 730 94 00

info@dcaf.ch
www.dcaf.ch

DCAF Geneva Centre for Security Sector Governance

Ministry of Foreign Affairs of the Netherlands