Digital Convergence USP2030 | AI HUB

AI HUB RISK ASSESSMENT FRAMEWORK FOR SOCIAL PROTECTION:

A Practical Guide to Responsible AI Deployment

AI HUB FOR SOCIAL PROTECTION

The AI Hub for Social Protection (AI Hub) is an initiative under the Digital Convergence Initiative (DCI), which helps social protection institutions to harness the benefits of responsible AI for the public good. It supports the development of strategies, frameworks, digital systems and institutional capacities to design, govern and implement AI solutions that are ethical, effective and uphold data sovereignty.

Visit our website at **spdci.org/ai-hub**.

ACKNOWLEDGMENTS

This report was authored by Thomas Byrnes (Consultant, MarketImpact Digital Solutions Ltd), Dr Simona Dobre (Data Scientist, Swiss Tropical and Public Health Institute [TPH]/CEO, DevAIs) and Dr Laura van der Erve (Senior Consultant, Oxford Policy Management). We would like to thank the many colleagues who reviewed the various drafts: Juul Pinxten and Phillippe Leite (World Bank), Rob Worthington (Kwantu), Gabriele Erba (UNICEF), Valentina Barca (independent consultant), Franziska Clausius and Nour Barakat (GIZ). We would also like to thank Susan Sellars (copyeditor) for helping to bring the report into its final form.

REQUIRED CITATION

DCI AI Hub for Social Protection, *AI Hub Risk Assessment Framework for Social Protection: A Practical Guide to Responsible AI Deployment*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Germany, 2026.

AI HUB KNOWLEDGE ARCHITECTURE

The AI Hub is committed to advancing the global dialogue on the use of AI in social protection by creating and sharing knowledge on the responsible adoption of AI solutions. This AI Hub Risk Assessment Framework for Social Protection is part of a broader portfolio of AI Hub knowledge products that support policymakers, practitioners, and partners in navigating the opportunities and challenges of AI in social protection.

Artificial intelligence is a fast-moving field, with new developments, evidence, and applications emerging continuously. While this report reflects the best available information at the time of publication, it may not include subsequent advances. Readers are encouraged to consult our AI Hub website for the most recent developments, publications, and resources: **spdci.org/ai-hub**.

CONTENTS

ACRONYMS

AI	artificial intelligence
AIA	AI Impact Assessment
API	application programming interface
DCI	Digital Convergence Initiative
DPIA	Data Protection Impact Assessment
EU	European Union
FbF	Forecast-based Financing
HiC	human-in-command
HITL	human-in-the-loop
HOOTL	human-out-of-the-loop
HOTL	human-on-the-loop
ICRC	International Committee of the Red Cross
LMIC	low- and middle-income country
MLOp	machine learning operation
OCHA	United Nations Office for the Coordination of Humanitarian Affairs
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
OHCHR	Office of the United Nations High Commissioner for Human Rights
UCT	unconditional cash transfer
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations Children's Fund

ABOUT THIS FRAMEWORK

The AI Hub Risk Assessment Framework for Social Protection is divided into two parts. The first part of this report titled *Understanding and Mitigating the Risks*, outlines the key theoretical and definitional foundations for the Risk Assessment. The second part, titled *Putting the Framework into Practice*, is a hands-on guide to applying the Risk Assessment Framework, referencing multiple use cases making it particularly useful for practitioners. Next to this report, and building on Part 2 of it, the AI Hub Risk Assessment Framework is to be applied using the companion Risks and Safeguards Toolkit which is publicly available for download and use on the AI Hub website https://spdci.org/ai-hub/.

The AI Hub Risk Assessment Framework for Social Protection

Part 1 – This section lays out the **theoretical framework for the Risk Assessment**. It focuses on understanding AI risks for social protection across the AI lifecycle, exploring the five structural sources of risk. It also explains the rights-based principles approach used for the Framework, while discussing human and societal impacts through use case examples to understand the risks in context.

Part 2 – This section **puts the Risk Assessment Framework into practice,** explaining the foundations for the practical application of the accompanying Toolkit. It also discusses the process for monitoring and reassessment after deployment, and how the Toolkit can also support this stage.

Risks and Safeguards Toolkit
(to be read alongside Part 2)

Tool 1. Intrinsic Risk Assessment Worksheet

Tool 2. Context Risk Multiplier Diagnostic

Tool 3. Risk Registration Template

Tool 4. Go/Pilot/No-Go Checklist

EXECUTIVE SUMMARY

While artificial intelligence has many benefits for social protection systems, it also carries with it certain risks. In social protection, decisions determine access to basic income, food security, and essential services, hence, the consequences of errors are uniquely severe. However, when responsibly deployed, AI can enhance responsiveness, extend coverage, and improve service efficiency.

This Risk Assessment Framework and its companion Risks and Safeguards Toolkit are designed to assess the risk posed by the use of AI in social protection to help decision makers decide whether a system is ready to *Go*, *Pilot*, or *No-Go*? It is specifically tailored to non-contributory social assistance. It translates abstract principles into a structured method for answering a fundamental operational question:
Are we ready for this system?

The Framework is based on a review of the literature and conceptual analysis carried out in 2025, drawing on illustrative use cases spanning both programme delivery and policy-level implementation. It is grounded in internationally recognised frameworks governing AI, such as the Organisation for Economic Co-operation and Development's (OECD's) Recommendation on AI (OECD, 2019), the United Nations Educational, Scientific and Cultural Organization's (UNESCO's) Ethics on AI Recommendation (UNESCO, 2021), and the European Union's (EU's) Artificial Intelligence Act (European Union, 2024). Risk analysis was conducted through mapping across the AI lifecycle and by applying a rights-based approach to AI governance.

The Framework identifies five structural sources of risk that accumulate across the AI lifecycle: data-related risks, model-related risks, operational and system integration risks, governance and institutional oversight risks, and market, sovereignty and industry structure risks. Together, these structural risks shape how AI systems behave in real institutions and define how harms materialise in people's lives.

The Framework recognises that risk is co-produced by technology and context. A model that is safe in one setting may be unsafe in another. To operationalise this insight, the report introduces a two-part assessment. Intrinsic Risk examines the system itself, its behaviour, data, and potential impact, across three dimensions: likelihood of harmful or biased outputs, severity of harm if errors occur, and scale of potential impact. The Context Risk Multiplier (CRM) adjusts the Intrinsic Risk by assessing institutional readiness across four dimensions: legal framework, institutional capacity, data ecosystem, and public trust and transparency. This combined logic ensures that deployment decisions reflect not only what the system *could* do, but also what the surrounding environment can realistically manage.

To turn this assessment into practice, the report provides a four-part operational Risks and Safeguards Toolkit:

- **Tool 1. Intrinsic Risk** – profiles the inherent risk of the AI system using evidence-based scoring.

Tool 2. Context Risk Multiplier – assesses whether safeguards and institutional conditions are sufficient to support deployment.

Tool 3. Risk Register – records risks, mitigations, and responsible owners, building institutional memory and preventing uncontrolled scaling ('pilot creep').

Tool 4. Overall Risk (Go/Pilot/No-Go Checklist) – provides a decision gate that confirms whether minimum safeguards, such as lawful basis, human oversight, and grievance routes, are in place before live deployment.

Together, these tools create a transparent, repeatable, and auditable process for evaluating the safety and readiness of any proposed AI system.

The Framework asserts that responsible AI governance includes responsible non-adoption. When Intrinsic Risks are high, or contextual safeguards are weak, a decision to pause ('No-Go') becomes an opportunity. It is not a setback, but rather an opportunity to provide the necessary protection for beneficiaries and for the integrity of social protection systems. Pausing deployment until necessary, and tested, safeguards are in place can often be the most responsible course of action.

By applying this framework, governments, development partners, and implementing agencies can make proportionate, defensible decisions. Innovation can proceed where conditions allow, pilots can generate evidence where uncertainty remains, and deployments can be paused where risks to rights, equity, or trust are too great. This ensures that AI serves social protection, rather than contributing to existing risks.

CHAPTER 1. INTRODUCTION

Artificial intelligence is reshaping social protection across diverse country contexts. Governments are deploying geospatial models to support poverty assessments, predictive analytics for shock anticipation, biometric systems to verify identity, and anomaly-detection tools to enhance payment integrity (Ohlenburg, 2020). This technology can contribute to expanding coverage, improving accuracy, and increasing responsiveness, particularly in low-capacity environments.

However, the adoption of AI in social protection has intensified risks, especially where there are rights implications. Social protection and, more specifically, social assistance, programmes determine access to essential support to the most vulnerable tiers of society. Errors generated by AI systems can deny people food and income stability, and even the ability to meet basic needs. While these risks are not new, what changes with AI and modern machine learning is the speed and scale of the impact. At the same time, AI can add value beyond traditional rule-based automation by enabling more flexible approaches under uncertainty, earlier detection of anomalies, and faster response to shocks, provided its uncertainty and error modes are explicitly managed (Wirtz et al, 2022).

To capitalise on the benefits of using AI in social protection, while at the same time minimising the risks, decision makers need a structured method for assessing such risk. They need to be able to answer the critical question: **Are we ready for this AI system?** To make this decision, governments and agencies require tools that are structured, context-appropriate and concrete.

Accordingly, this report provides a Framework for assessing the risk posed by the use of AI in social protection to help decision makers decide whether a system is ready to **Go**, **Pilot**, or **No-Go**? It is complemented by a Risks and Safeguards Toolkit that operationalises this decision-making process. The Framework is based on a review of the literature and conceptual analysis carried out in 2025, drawing on illustrative use cases spanning both programme delivery and policy-level implementation. It is grounded in internationally recognised frameworks governing AI, such as the Organisation for Economic Co-operation and Development's (OECD's) Recommendation on AI (OECD, 2019), the United Nations Educational, Scientific and Cultural Organization's (UNESCO's) Ethics on AI Recommendation (UNESCO, 2021), and the

THE USE OF AI IN SOCIAL PROTECTION CAN LEAD TO HARM

Examples include the SyRI welfare surveillance model in the Netherlands, which was struck down for violating privacy and discrimination protections (District Court of The Hague, 2020) and Denmark's predictive fraud-detection system, which has faced extensive criticism for algorithmic profiling and opaque decision-making (Amnesty International, 2024a). These failures emerged in high-capacity settings; for low- and middle-income countries (LMICs), weak data ecosystems, unclear mandates, and limited oversight magnify exposure to harm (DCI AI Hub, 2026c).

European Union's (EU's) Artificial Intelligence Act (European Union, 2024). Risk analysis was conducted through mapping across the AI lifecycle and by applying a rights-based approach to AI governance. Lifecyle mapping traces how risks emerge, evolve and are compounded at each stage of the AI lifecycle. The rights-based approach treats AI governance not merely as a technical consideration, but as a legal, ethical, and institutional duty (see Annex 2, Rights-based foundation for AI governance in social protection). Although the Framework can be applied to social protection generally, it is primarily designed for non-contributory social assistance, where the risks are severe (see Annex 1, Methodology).

The Framework is structured into two main parts.

PART 1 focuses on understanding AI risks for social protection, providing an overarching Framework for this. It includes three sections: Section 2 explores how risks are generated at different stages of the AI lifecycle and the five structural sources of risk. For each risk, specific mitigation measures are outlined. The rights-based principles are then set out to operationalise the structural sources of risk. This is followed by Section 3 which discusses the human and societal impacts and Section 4 which provides some typical AI use cases for the sector as examples to understand the risks in context.

PART 2 puts the Framework into practice and sets the foundations for the practical Risks and Safeguards Toolkit that accompanies this document. Section 5 explains the risk-assessment and scoring method adopted by the Toolkit, its component parts (Tools 1–4) and the process to fill these in. Section 6 suggests a workflow and Section 7 discusses monitoring and reassessment. Finally, Section 8 contains a brief conclusion.

Statement on Scope: *What the Risks and Safeguards Toolkit is and is not*. The Toolkit is designed to assess the pre-deployment risk of an AI system in non-contributory social assistance and to inform the final decision on whether to proceed. It can be adapted to other social protection domains with expert judgement. It is not a substitute for legal compliance, a data protection impact assessment, or procurement due diligence, and it is not a technical build specification.

PART 1
UNDERSTANDING AND MITIGATING THE RISKS

So how are AI risks generated in social protection systems?
When algorithms assist in social protection prioritisation choices, targeting, verification, or monitoring, even small errors can translate directly into exclusion, the misallocation of resources, and loss of trust. This part looks at the risks that emerge at different stages of the AI lifecycle, the various structural sources of such risk in social protection systems and how rights-based principles relate to such sources, as well as how these sources interact to produce human and societal impacts. Finally, to elaborate, some examples of situations in which AI has been deployed in social protection are provided.

CHAPTER 2. HOW ARE AI RISKS GENERATED?

Risks along the AI lifecycle

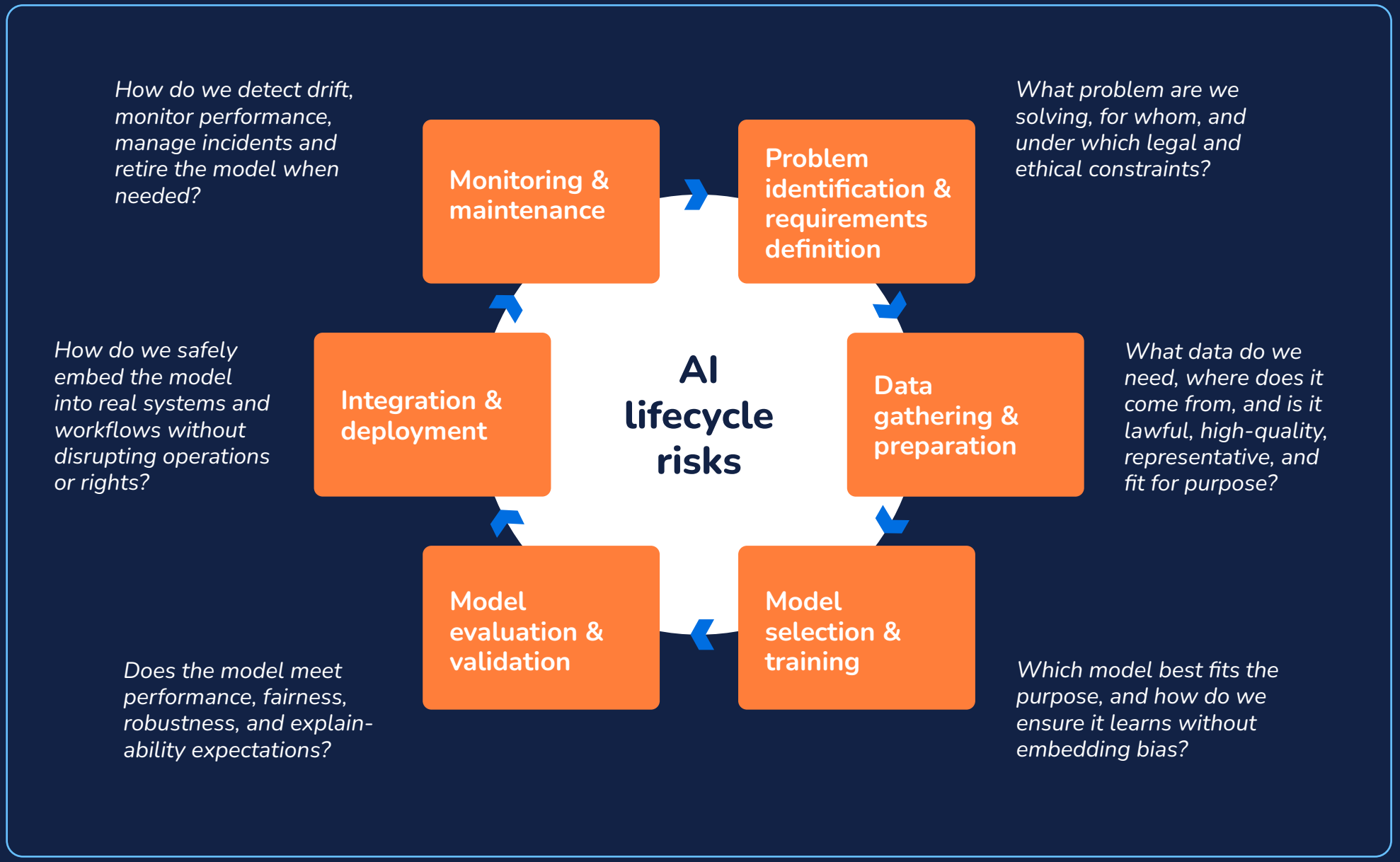
AI-related risks for social protection emerge across the entire AI lifecycle – from data collection and model design to deployment, monitoring, and decommissioning (see Figure 1) – and are shaped by the institutions, infrastructure, and social contexts in which AI operates (OECD, 2019; NIST, 2023; DCI AI Hub, 2026a; DCI AI Hub, 2026c; Zhang et al., 2022). A technically accurate model can still produce harmful or discriminatory outcomes if it is trained on poor data, integrated into fragile systems, deployed without adequate oversight, or used in environments where beneficiaries cannot understand or challenge automated decisions. Conversely, strong governance cannot compensate for flawed datasets, opaque logic, or insufficiently tested integration. To show how these risks accumulate and interact, Figure 1 maps the six stages of the AI lifecycle and the critical questions that determine whether risks are identified, mitigated or allowed to propagate (adapted from OECD, 2019; NIST, 2023).

Human oversight is important in any AI system. The following terms describe the level of oversight:

- **HiC (human-in-command):** Humans set the strategy and define policies in natural language, then the AI operates autonomously within the guidelines defined by humans.
- **HITL (human-in-the-loop):** Humans are active participants, providing input and reviewing critical decisions at every step.
- **HOTL (human-on-the-loop):** Humans act as supervisors, monitoring the system and intervening only when necessary.
- **HOOTL (human-out-of-the-loop):** The AI operates autonomously, with minimal or no human intervention.

While many technical risks are domain-agnostic, their consequences in social protection (and social assistance specifically) are uniquely severe, because they directly affect income security, dignity, and autonomy.

FIGURE 1. QUESTIONS TO ASK TO IDENTIFY RISK ALONG THE AI LIFECYCLE



The six AI lifecycle stages and their risks:

- **Stage 1.** Problem identification and requirements definition – Risk of misframed problem statements, absence of legal basis, exclusion of affected populations, or inappropriate automation of rights-critical decisions
- **Stage 2.** Data gathering and preparation – Risk of representation bias, limited coverage of registries, weak data provenance, fragmented governance, and insufficient data minimisation
- **Stage 3.** Model selection and training – Risk of lack of interpretability, robustness failures, optimisation trade-offs, overfitting, and reliance on proprietary systems that obscure auditability
- **Stage 4.** Model evaluation and validation – Risk of insufficient testing on local or representative data, failure to assess performance across subgroups, reliance on vendor-reported metrics without independent verification, and inadequate stress-testing for edge cases or adversarial inputs
- **Stage 5.** Integration and deployment – Risk of automation bias, inadequate human oversight, operational disruption, and failures to implement HITL or HOTL controls required by *A Taxonomy for AI in Social Protection* (DCI AI Hub, 2026a)
- **Stage 6.** Monitoring and maintenance (includes decommissioning) – Risk of absence of continuous monitoring, insufficient grievance pathways, inability to retrain or adapt models, and institutional capacity constraints

Structural sources of risk and mitigation measures

This Framework identifies five domain-agnostic structural sources of risk (DCI AI Hub, 2026a; Steimers & Schneider, 2022; Slattery et al., 2024). These five structural sources, define the conditions that determine whether an AI system behaves reliably, transparently, and fairly (see Figure 2 below). These sources impact differently on each stage of the AI lifecycle. The consequences of these risks directly affect beneficiaries' rights, autonomy, and access to essential support (see Section 3 for a discussion of the human and societal impacts).

FIGURE 2. STRUCTURAL SOURCES OF RISK

Data-related risks

These risks arise when the information used to train, validate, or operate an AI system does not accurately or fairly reflect the real world (Ntoutsi et al., 2020; Pagano et al., 2023; Ferrara, 2023). A central source of these risks is representation bias (Chu et al., 2023; Norori et al., 2021; Nazer et al., 2023), which occurs when the dataset does not proportionally reflect the populations or situations that the model will encounter in practice. Representation bias can stem from historical bias (when past social, economic, or institutional inequalities are embedded in the data itself); uneven sampling (when the data collection process systematically favours certain groups), or insufficient coverage (when some groups or contexts are underrepresented or entirely absent in the final dataset) (Hiniduma et al., 2024; Ntoutsi et al., 2020; Lindert et al., 2020; World Bank 2024, 2025).

As well as representation bias, data quality and integrity problems can significantly distort predictions and undermine model reliability. To identify these risks systematically, social assistance systems should assess data against the six core dimensions of quality defined in the *Data Governance Framework for Digital Social Protection Systems* (DCI, 2025): completeness, accuracy, currency, consistency, validity, and uniqueness. Failures in these dimensions, such as incomplete records, outdated information (lack of currency), or duplicated values, compromise the integrity of eligibility determination and operational efficiency (Gong et al., 2023; Mohammed et al., 2022; Hiniduma et al., 2024).

DATA BLIND SPOTS

When individuals are wrongly classified, excluded, or disadvantaged, these errors are rarely recorded in administrative datasets. As described by Virginia Eubanks in *Automating Inequality* (2018), this creates persistent blind spots: the data never captures these mistakes, and the system cannot learn from them. Over time, this absence of corrective information reinforces representation bias, because the dataset fails to include the experiences of those who have been misclassified or harmed, leaving the model increasingly unaware of the full range of real-world outcomes.

When data originate from multiple institutions, inconsistencies across datasets, such as mismatched definitions, identifiers, or formats, introduce additional distortions that propagate throughout the system. Over time, real-world conditions may change in ways that make past data less reliable for current decisions. This phenomenon, often called 'data or concept drift', can gradually reduce system accuracy without being immediately visible[1] (NIST, 2023; Munz et al., 2023). Models may rely on superficial correlations or indirect indicators that do not truly capture the intended outcome. When this happens, predictions can break down in new situations or amplify existing disparities (Banerjee et al., 2023; Zhang et al., 2022).

1 Note: Data and concept drift is described here as data-side phenomena; later sections distinguish the operational monitoring mechanisms (operational risks) and the institutional mandate/ownership responsibilities (governance risks).

MITIGATING DATA-RELATED RISKS

To address data-related risks, mitigation strategies should follow the data lifecycle, from intake and structuring to model training and long-term monitoring. The goal is to prevent distortions from entering the system, ensure that datasets remain representative and traceable, and detect changes that may erode model performance over time. Key strategies include:

- **Enforce data-quality rules** in order to detect incomplete, inconsistent, noisy, or duplicated records.
- **Reinforce lineage and provenance controls** to document data origins, transformations and assumptions.
- **Adopt harmonised schemas and shared definitions** across agencies to reduce cross-dataset inconsistencies.
- **Conduct systematic bias analysis** (e.g. distribution checks, subgroup comparisons, feature audits) to identify representation gaps, label bias or structural imbalances before model training.
- **Apply fairness-oriented preprocessing** (rebalancing, reweighting, or synthetic data generation) based on the results of bias analysis to mitigate identified distortions.
- **Implement continuous monitoring** to detect data drift and concept drift during deployment.
- **Use purpose limitation and data-minimisation practices** to ensure that only relevant and proportionate data influence model outcomes.

Note: When models or scoring outputs developed for one programme are reused for eligibility determination in other programmes, the original purpose limitation, data coverage, and accuracy thresholds must be reassessed for the new context. Raising eligibility thresholds alone does not address the mismatch between the population the model was trained on and the population it is now being applied to.

Model-related risks

Some risks originate in the architecture, objectives, and internal behaviour of the algorithm itself (OECD, 2019; Zhang et al., 2022; He & Wang, 2025). They include opacity and limited explainability, which make it difficult to inspect how the model reaches its conclusions, as well as objective misalignment (Kelly et al., 2019; Zhang et al., 2022), which is where the optimisation target does not match the real policy or social goal. Another key source of harm arises when the model is built on incorrect assumptions about how factors relate to each other. This issue, known as model misspecification, can lead to systematic errors that cannot easily be corrected.

Models may also exhibit algorithmic bias, producing uneven or discriminatory outcomes across groups, even when the underlying data appear balanced (Ferrara, 2023; Gray et al., 2023; Nazer et al., 2023; Ntoutsi et al., 2020; Zhou et al., 2022). They can also engage in shortcut learning (Banerjee et al., 2023), relying on superficial correlations or proxy variables, rather than stable, generalisable signals relevant to the target outcome. This can amplify disparities or lead to distorted predictions under context shift.

A broader set of risks relates to reliability and generalisation (Fu et al., 2024; Munz et al., 2023; Mohammed et al., 2022; Zhang et al., 2022). These arise when a model performs well on training data, but fails to behave consistently under real-world conditions. They

include 'overfitting', when the model memorises noise and fails to generalise, and 'underfitting', when the model is too simple to capture genuine patterns. Reliability issues also emerge when models encounter unusual or unfamiliar inputs that differ from the data used during training. In such cases, performance may become unstable or less fair. Complex or adaptive systems may display emergent behaviours that arise from internal dynamics rather than opacity, while self-updating or autonomous models may exhibit behavioural drift, gradually diverging from intended outcomes.

Finally, security-related risks can also originate at the model level. 'Model poisoning', through tampered training data, corrupted gradients, or manipulated parameters, can degrade performance or introduce hidden vulnerabilities. Generative or stochastic models may produce misinformation or hallucinations, particularly when operating outside their intended context or without adequate safeguards (Mauri & Damiani, 2022; Zhang et al., 2022; Ferrara, 2023).

MITIGATING MODEL-RELATED RISKS

To address model-related risks, mitigation strategies should, once again, follow the model lifecycle, from objective/architecture choices through training and pre-deployment validation. The aim is to ensure that models generalise reliably, behave consistently across subgroups, remain well-calibrated, and are resilient to adversarial manipulation, with additional, model-specific controls for generative AI to reduce hallucinations and misinformation. Key strategies include:

- **Define model objectives and decision rules so that they reflect the real-world consequences of errors,** such as wrongly excluding eligible individuals, rather than focusing only on overall statistical accuracy.
- **Control overfitting and underfitting** through robust validation designs (independent test sets, repeated cross-validation/bootstrapping) and model-complexity controls (regularisation, early stopping, pruning, constraints) to improve generalisation.
- **Institutionalise structured error analysis,** including systematic false-positive/false-negative review and threshold selection aligned with acceptable risk.
- **Integrate fairness safeguards** directly into model development and require performance to be assessed across relevant groups before deployment. More advanced technical methods may be used where appropriate, but the key requirement is systematic subgroup evaluation.
- **Test the system under stress conditions,** including deliberate attempts to manipulate or deceive it, to ensure it behaves safely and predictably.
- **For generative AI** (large language models and other generative models), apply model-specific controls, alignment/post-training, factuality and hallucination-reduction measures, constrained generation, adversarial/jailbreak stress-testing,[2] and bias/harm evaluation/mitigation to reduce unsafe, unsupported, misleading, or discriminatory outputs.

2 Adversarial testing is a method for evaluating a machine learning model to learn how it behaves when provided with malicious or inadvertently harmful input. Jailbreaking is another way to do so and refers to manipulating large language models to bypass their intended restrictions or safety protocols. These methods of testing are useful to strengthening the protection of generative AI models.

Operational and system integration risks[3]

Risks arise when AI models are deployed in real infrastructure, data pipelines, and organisational workflows, where technical performance depends not only on the model itself, but also on the reliability of the surrounding system. They include integration failures with legacy applications and data architecture, which can create brittle interfaces, incompatible formats, and unstable production deployments (Zimelewicz et al., 2024; Singh, 2025). Operational risks also emerge from broken or delayed data pipelines. Changes in data structure, naming conventions, measurement units, or data volumes can silently spread through workflows and lead to undetected errors or cascading failures (Eken et al., 2024; Sribhashyam, 2025). In addition, scalability and latency constraints may prevent timely model serving, retraining, or updates, especially in dynamic or streaming environments, leading to stale outputs, degraded service quality, and increased downtime (Vijayan, 2024; Eken et al., 2024).

A second cluster of risks concerns weak monitoring, validation, and machine learning operation (MLOp) practices in production. Many of the models in use are insufficiently monitored, limiting drift detection and incident response (Paleyes et al., 2020; Zimelewicz et al., 2024). Weak operational management of deployed models, such as limited record-keeping, incomplete version control, or the absence of rollback mechanisms, reduces reproducibility, delays recovery, and increases the risk of prolonged failures or unsafe updates (Paleyes et al., 2020; Vijayan, 2024; Eken et al., 2024).

MITIGATING OPERATIONAL AND SYSTEM INTEGRATION RISKS

Reducing operational and system integration risks should cover the full deployment lifecycle, from system design and interface definition to pipeline automation, staged release, monitoring, and incident response. The goal is to keep AI systems interoperable with existing infrastructure, resilient to failures, observable in real-world use, and recoverable when issues occur. In practice, this means combining phased integration, reliable pipelines, scalable infrastructure, strong monitoring, real-world validation, and disciplined MLOp practices for reproducibility, rollback, and safe updates. Key strategies include:

- **Introduce new systems gradually** through controlled testing, such as running them in parallel with existing processes before full replacement, and define clear system interfaces to reduce disruption and fragile integration with legacy systems.
- **Harden data and deployment pipelines** with orchestration, automated validation, schema checks, and lineage tracking so data-quality or configuration issues are detected early.
- **Design for scalability, latency, and resilience** using appropriate serving architectures (e.g. autoscaling, hybrid cloud/edge setups) and model optimisation techniques (e.g. pruning, quantisation, distillation).
- **Implement end-to-end monitoring and incident response** across data quality, model behaviour, infrastructure health, and operational performance, with clear alerts and response procedures.
- **Validate in real deployment conditions** using staged release/testing approaches (e.g. shadow mode, A/B testing, canary rollout) and domain-relevant metrics, including subgroup-level checks.
- **Institutionalise robust MLOp practices** such as versioning of models/data/code, logging, model registries, and rollback mechanisms.
- **Embed governance and risk management in operations** through access controls, approval gates, audit trails, and formal AI risk management processes to keep deployment decisions accountable and documented.

3 Note: This section covers technical operational controls; the next section on Governance and Institutional Oversight Risks covers decision rights, accountability and resourcing for operational follow-up.

Governance and institutional oversight risks

These risks emerge when institutions lack the structures, decision rights, capacity, and accountability needed to keep an AI system under effective control across its lifecycle, from design and procurement through deployment, change management, monitoring, and retirement (Zhou et al., 2022; Zhang & Zhang, 2023). In practice, these risks are less about the model's technical accuracy and more about whether the organisation can reliably see what the system is doing, justify why it is doing it, intervene when it fails, and ensure that responsibilities and legal duties are met (DCI AI Hub, 2026a; OHCHR, 2021; Jobin et al., 2019).

These risks typically manifest in various ways, including: (i) unclear governance structures (no clear ownership, fragmented roles, weak escalation pathways, and blurred decision authority); (ii) lack of accountability (responsibility gaps between teams, vendors, and public authorities, making harms hard to attribute and remediate); (iii) insufficient risk management and institutional capacity (limited AI literacy, inadequate resourcing, and weak internal controls); and (iv) insufficient monitoring (issues such as drift, degraded performance, or unexpected failure modes persist because there is no defined institutional mandate for continuous supervision and incident follow-up). These risks are reinforced by organisational culture and strategic misalignment, especially when incentives prioritise rapid deployment over safe operation and staff lack the authority, or feel pressure not, to challenge automated outputs, leading to automation complacency.

Compliance-related risks fit naturally within this category, because many requirements are institutional rather than purely technical. They include failure to establish or evidence a lawful basis for processing, enforce purpose limitation and data minimisation, meet transparency and documentation obligations, ensure auditability and traceability, or align deployments with sectoral rules and procurement constraints. Finally, governance failures often become visible through weak or missing redress and contestability: affected individuals cannot understand, challenge, or appeal AI-influenced outcomes; internal teams lack clear processes to review contested decisions; and errors do not feed back into corrective action, allowing harms (including bias or inequity originating in data/model mechanisms) to persist unaddressed in production.

MITIGATING GOVERNANCE AND INSTITUTIONAL OVERSIGHT RISKS

The goals here are multiple and linked to the actions that help ensure adequate governance and accountability, e.g. to ensure that oversight is risk-tiered and meaningful (not symbolic), that accountability is explicit and enforceable, that assurance activities (audits, assessments, monitoring) are continuous and action-oriented, and that institutions have the capacity, culture, and safeguards to challenge AI outputs, meet compliance obligations, and provide redress when harms occur. Key strategies include:

- **Adopt risk-based human and institutional oversight** by mapping *model influence and decision consequences* to appropriate oversight modes (e.g. HiC, HITL, HOTL), with stricter control for high-impact use cases or failures that are difficult to detect, including by:
  - **Adjusting oversight intensity over time** as context, data conditions, model behaviour, or risk profile changes, not only at initial deployment
  - **Ensuring oversight is operationally feasible** (workload, authority to override, and usable explanations), otherwise HITL becomes a compliance label rather than a safeguard

  ...

...

- **Formalise AI governance structures, roles, and decision rights** to close responsibility gaps and prevent 'responsibility vacuums' in monitoring, incident response, and retirement. This includes clear allocation of roles for design approval, deployment authorisation, change control, monitoring ownership, incident handling, and decommissioning.
- **Embed AI governance into existing institutional governance and compliance systems** (e.g. information governance, programme governance, clinical governance where relevant), rather than creating a parallel silo. This reduces fragmentation and strengthens enforceability and continuity.
- **Institutionalise continuous auditing, assessment, and operational assurance,** so that oversight is evidence-based and corrective, not once-off, including by:
  - Conducting **risk assessments** and maintaining operational audit trails that support accountability and compliance verification
  - Treating algorithm auditing (fairness metrics, adversarial tests, participatory audits) as a **socio-technical process** linked to corrective actions, not a stand-alone technical check
  - For higher-risk deployments, incorporating **red teaming[4] and stress testing** against domain-specific harms and failure modes; where relevant, use structured oversight frameworks (e.g. healthcare-oriented stewardship/oversight frameworks) to systematise review
- **Strengthen transparency, documentation, and traceability** that make oversight possible in practice: maintain documentation such as model cards, use manuals, decision logs, and clear communication to affected stakeholders to support scrutiny, contestation, and accountability.
- **Strengthen institutional capacities to foster a culture of informed AI oversight[5]** through sustained AI literacy programmes and trainings, ensuring staff at different levels can understand system limitations, critically interpret outputs, and safely question or override systems. Governance structures should embed this capacity through clear escalation pathways, defined override protocols, and regular audit cycles. Institutional leadership should actively signal the value of questioning AI outputs, create incentives and an environment of psychological safety to encourage staff to question system outputs and avoid automation complacency.
- **Operationalise compliance and redress as institutional functions** (not just legal statements): ensure decision pathways are reviewable, responsibilities for handling complaints/appeals are assigned, and stakeholders have channels to contest outcomes and trigger review, so governance failures surface early and lead to remediation rather than persisting invisibly.
- **For extreme or systemic risk profiles, adopt adaptive governance and coordination mechanisms** that go beyond local, reactive controls, recognising that some risks require proactive safeguards and cross-organisational alignment.

4 Red teaming is an evaluation methodology to discover flaws and vulnerabilities in AI systems, typically through deliberate adversarial testing designed to simulate real-world misuse or failure scenarios.

5 A challenge culture, applied to AI oversight, is an environment in which questioning or overriding AI-generated outputs is encouraged, so that human judgement remains an active check on automated decision-making rather than a passive rubber stamp.

Market, sovereignty and industry structure risks

These risks arise when AI systems in production depend on external vendors, cloud infrastructure, proprietary application programming interfaces (APIs), and global technology ecosystems.
In these settings, risks not only include technical performance, but also concentrated control over compute, models, toolchains, pricing, and standards, which can reduce institutional autonomy, create asymmetric bargaining power (often most acute for LMIC institutions with limited negotiation leverage and fewer local alternatives), and expose public-sector services to shocks originating outside the implementing organisation (DCI AI Hub, 2026b; Schmitz et al., 2024; Zeng et al., 2024; Uuk et al., 2025; Van Der Vlist et al., 2024; Leslie & Perini, 2024; Pirrone et al., 2025).

These risks typically appear through three intertwined channels:

- **Market and concentration risks.** A small number of major cloud and platform providers can become 'default' infrastructure for model training, deployment, monitoring, and marketplaces. This creates lock-in via technical coupling (APIs, managed services, proprietary tooling) and contractual constraints, raising switching costs and exposing implementers to unilateral pricing, policy, or product changes (e.g. API deprecations, usage limits, feature gating). Such dependence is often reinforced by limited audit access to vendor-controlled systems, absent or weak exit and migration pathways, and rapid technology churn (frequent model/API changes), which increase long-term costs and make continuity planning harder. Dependency also extends upstream: foundation-model ecosystems rely on concentrated inputs, compute, chips, data, capital, and specialised talent, creating chokepoints where disruptions or strategic behaviour can cascade to downstream deployments. This dependence is amplified when institutions rely on off-the-shelf models accessed via vendor-hosted APIs, with which implementers have limited visibility into training data, design choices, and evaluation methods, and limited control over vendor updates. This can constrain independent assurance of stable accuracy and subgroup performance over time and create exposure to 'silent' shifts in behaviour when models, safety layers, or routing policies are modified upstream.
- **Sovereignty and geopolitical risks.** Reliance on foreign clouds, foreign-hosted models, and cross-border data flows can undermine operational sovereignty over critical digital services: control over data location, access conditions, audit rights, and model roadmaps may sit outside national or institutional governance. Geopolitical and regulatory shifts, export controls on chips, sanctions, data localisation rules, or changes in foreign jurisdictional requirements can suddenly restrict access to computing resources or models, with direct effects on service continuity. Sovereign AI initiatives may still mask material dependencies if they remain reliant on imported graphics processing units (GPUs), foreign cloud platforms, or vendor-controlled foundation models.

For more information on AI data sovereignty risks and how they can be assessed and mitigated for social protection, read the DCI AI Hub's report entitled *AI in Social Protection: Data Sovereignty Considerations* (2026c).

- **Industry-structure and systemic risks.** The rapid expansion of large, pre-trained AI models developed by a small number of major providers concentrates technical and governance power in a limited set of actors. Failures, misuse, security incidents, or governance errors at this upstream layer can propagate widely through dependent applications, creating systemic fragility. In addition, AI deployments increasingly resemble multi-tier supply chains (pre-trained models, agents, data sources, managed services, evaluation tooling), yet these dependency graphs are often under-documented and under-governed, especially in critical public services. Dominant vendors can also shape benchmarks, toolchains, and best practices, embedding path dependence and limiting the diversity of architectures and governance approaches available to downstream users.

MITIGATING MARKET, SOVEREIGNTY AND INDUSTRY STRUCTURE RISKS

To address these risks, mitigation should treat external AI dependencies (vendors, clouds, APIs, and foundation-model ecosystems) as core risks, not implementation details. The aim is to reduce single points of failure and bargaining asymmetries, strengthen public-sector negotiation and vendor-management capacity, strengthen portability and auditability, engineer sovereignty and jurisdictional control into architectures, and add governance layers that can withstand upstream shocks (pricing, access restrictions, model/API deprecations, and systemic incidents) (Van Der Vlist et al., 2024; Wirtz et al., 2022; Uuk et al., 2025). Key strategies include (see also DCI AI Hub's *AI in Social Protection: Data Sovereignty Considerations,* 2026c):

Market and concentration risks:

- **Diversify vendors and critical inputs** by avoiding single-provider chokepoints (e.g. multi-cloud/multi-model strategies where feasible) and adopting supply-chain esilience practices such as alternative suppliers, contingency planning, and risk forecasting to reduce disruption exposure and bargaining power asymmetries.
- **Reduce lock-in through procurement and portability requirements** (open standards where possible, modular architectures, explicit exit/migration pathways), so switching is a governed option rather than a crisis response.
- **Embed contractual and audit controls** with vendors, including service-level commitments (availability, latency, incident communication), change-control and notification clauses for API/model updates, termination rights, and enforceable data/model portability provisions, including audit-access rights to vendor-controlled systems where feasible. Complement internal assurance with independent third-party audits and safety incident reporting for high-impact systems.
- **Use competition-oriented policy levers** (procurement rules, licensing conditions, open standards, and support for open-source alternatives) to counter structural concentration and increase credible options for downstream users.

Sovereignty and cross-border risks:

- **Adopt sovereignty-compatible hosting patterns** (local/regional hosting, sovereign cloud zones, locality-aware routing) to retain control over where data and computing reside and which jurisdictions apply.
- **Engineer data sovereignty by design** through data classification, locality constraints, and, where appropriate, federated or distributed learning approaches that keep sensitive data in-jurisdiction while enabling learning and service improvement.
- **Align with strong regional frameworks while retaining autonomy** by integrating national strategies, data-protection regimes, and local oversight institutions into AI governance, so that compliance obligations and enforcement do not become fully outsourced to external vendors' regulatory environments.

...

...

Industry-structure and systemic risks:

- **Map and manage AI supply chains** (models, APIs, agents, data services, hosting, evaluation tooling) to identify multi-tier dependencies, chokepoints, and upstream failure modes, and to make liability, recourse, and continuity planning explicit.
- **Build resilience through modularity and redundancy** (technical and contractual), so upstream outages, governance failures, or market shocks do not cascade into service collapse.
- **Overlay governance, transparency, and incident-sharing obligations** that improve accountability across dependency chains (including expectations for disclosure, audit trails, and learning from incidents), especially for high-impact deployments.
- **For frontier/systemic-risk profiles, use explicit systemic-risk management** (risk tolerance statements, pre-deployment risk assessment, deployment/containment controls) to reduce the probability and severity of system-wide shocks originating upstream.

Mitigating external dependency risks requires treating AI-enabled services as critical infrastructure: diversify vendors and compute, build sovereignty and locality into architectures, and apply rigorous governance and audit mechanisms that rebalance power between dominant providers, states, and downstream implementers (Wirtz et al., 2022; Uuk et al., 2025; Bengio et al., 2023).

Linking rights-based principles to structural risk sources

The sub-section discusses a set of rights-based principles that provide the normative foundation for managing the five structural sources of AI risk discussed in the previous section. Each principle directly corresponds to a set of vulnerabilities in the AI lifecycle and defines how institutions must act to prevent those vulnerabilities from turning into systemic harm. The key principles (or rather, organic groupings of key principles) are: data privacy and data security; equality, non-discrimination, fairness, and inclusion; autonomy, human dignity, and due process; and accountability, transparency, and redress (see Annex 2 for a discussion of these principles).

PRINCIPLES FOR RIGHTS-BASED USE OF AI IN SOCIAL PROTECTION

A set of guiding principles provide the normative foundation for managing the five structural sources of AI risk. Each principle directly corresponds to a set of vulnerabilities in the AI lifecycle and defines how institutions must act to prevent those vulnerabilities from turning into systemic harm (see Annex 2 for more details on these principles). Key principles (or rather, organic groupings of key principles) include:

Principle 1. Privacy and data security – including responsibility to:

- Safeguard personal data through lawful purpose limitation, data minimisation, and strong technical and organisational security measures.
- Ensure that the collection, use, linking and sharing of data respect confidentiality, integrity, and individual control, including access management, secure processing, and audit logs.

Principle 2. Equality, non-discrimination, fairness, and inclusion – including responsibility to:

- Prevent and correct bias, disparate impact, or structural discrimination across gender, ethnicity, region, disability, socioeconomic status, or other vulnerability factors.
- Ensure equitable access to services, proportionality, and fair treatment in all automated assessments, including safeguards for vulnerable groups.

Principle 3. Autonomy, human dignity, and due process – including responsibility to:

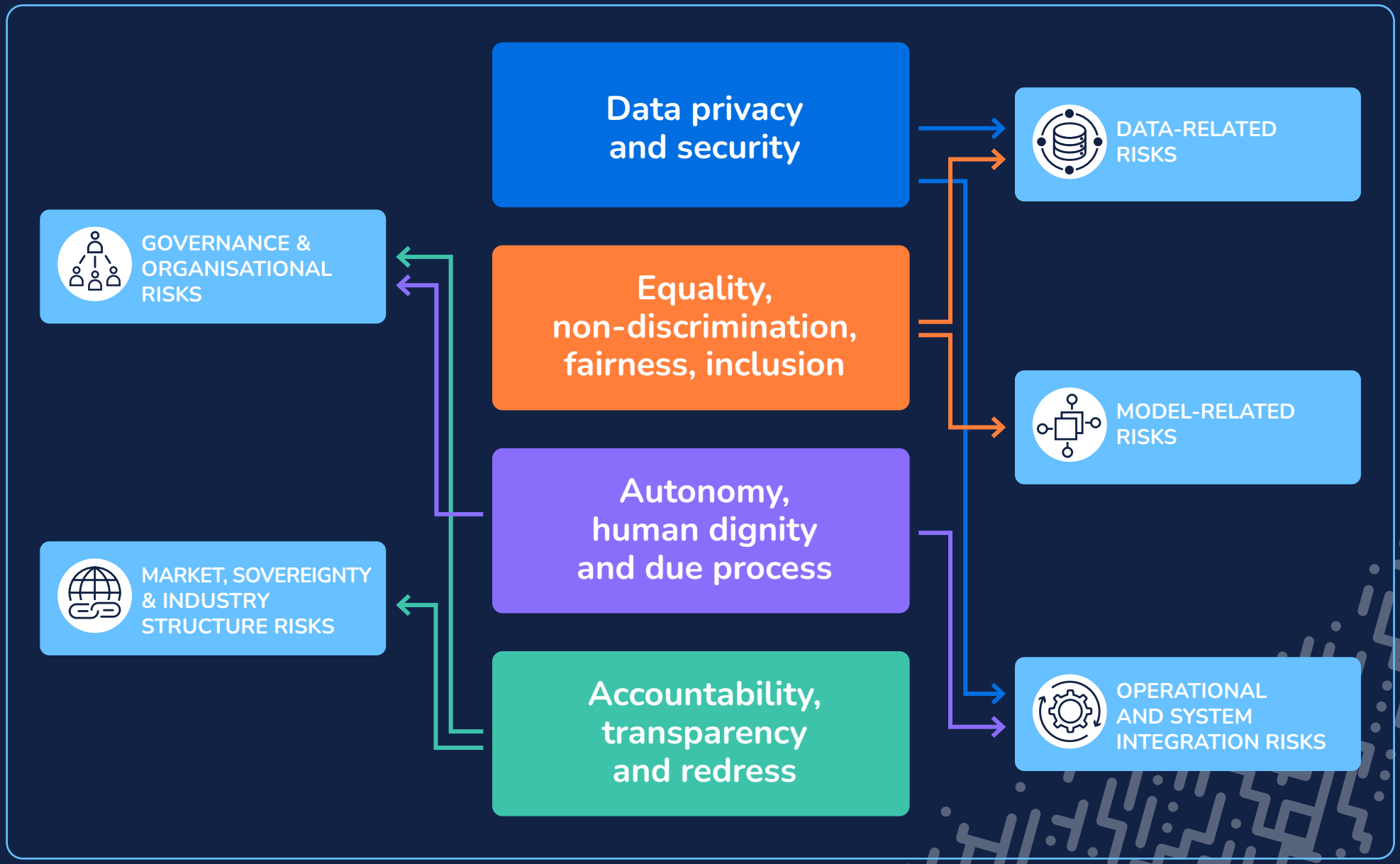
- Preserve people's ability to make meaningful choices and remain recognised as rights-bearing subjects, not passive objects of automated classification.
- Guarantee notice, meaningful explanation, contestability, and the right to be heard.
- Ensure that AI-assisted decisions do not undermine dignity, agency, or procedural justice through opaque and rigid automation.

Principle 4. Accountability, transparency, and redress – including responsibility to:

- Assign clear responsibility to a named public authority for every AI-assisted decision
- Ensure transparency appropriate to the system's impact, through documentation, explainability, traceability and auditability, and provide accessible, timely, and effective grievance and redress mechanisms.

These principles form a rights-based shield around the structural sources of risk. They ensure that weaknesses in AI data, models, operations, governance, or market dependencies (i.e. the risks in Section 2) cannot silently cascade into human harm, and that public authorities retain both the duty and the capability to act when something goes wrong. Figure 3 illustrates which risks are most tied to each principle, highlighting where governance gaps are most likely to produce adverse outcomes for individuals and communities.

FIGURE 3. RELATIONSHIP BETWEEN STRUCTURAL RISKS AND GUIDING PRINCIPLES



For responsible AI deployment, these guiding principles need to be translated into concrete safeguards that operationalise rights in practice. To avoid remaining merely symbolic, they must be translated into specific, enforceable institutional duties, supported by documentation, monitoring, and auditability. This ensures that safeguards operate not just as principles, but as practical controls that prevent data-, model-, operational-, governance-, and market-level risks from cascading into harm for beneficiaries. More specifically:

- Lawful basis and purpose limitation should anchor every AI use case, supported by Data Protection Impact Assessments (DPIAs) and AI Impact Assessments (AIAs) (Ada Lovelace Institute & DataKind UK, 2020; ICRC, 2024).

  See also the DCI-endorsed ***Implementation Guide – Good Practices For Ensuring Data Protection And Privacy In Social Protection Systems*** **(Wagner et al., 2021)**

- Transparency measures such as plain-language explanations, public model cards, and immutable audit logs make systems understandable and accountable.

- Human agency needs to be preserved through structured oversight models (HITL, HOTL, HOOTL), and contestability guaranteed by accessible grievance and appeal routes.

- Bias and error require managing through ex-ante fairness tests, periodic audits, and published residual-risk statements, ensuring that risks are not only identified, but actively mitigated.

The following case from Bangladesh provides a useful example of how different safeguard mechanisms can combine to mitigate and minimise existing risks.

Case study: Bangladesh Forecast-Based Financing – Safeguards for predictive AI in action

Bangladesh's Forecast-Based Financing (FbF) system illustrates how predictive AI can operate safely under HOTL oversight. Implemented by the Bangladesh Red Crescent Society in partnership with the International Federation of Red Cross and Red Cresent Societies (IFRC) and the Meteorological Department, the FbF model uses ensemble machine-learning forecasts to predict floods and heatwaves before impact. When forecast thresholds are reached, authorised officials validate alerts and approve the release of anticipatory cash assistance to at-risk households. This approach integrates adaptive analytics into national disaster-response protocols while maintaining accountability through multi-agency review (IFRC, 2020; Gros et al., 2022; Bangladesh Humanitarian Country Team, 2025; Anticipation Hub, 2026).

Safeguards in practice

The FbF mechanism applies four proportional controls that balance predictive autonomy with institutional accountability. HOTL validation: Activation decisions are confirmed by authorised disaster-management officials before funds are released. Ensemble modelling and cross-agency data sharing: Forecasts blend inputs from national meteorological services and international partners (IFRC and the Climate Centre) to reduce false positives. Documented activation protocols: Standing Orders on Disaster (2019) provide a clear legal basis for decision-making and transparency. Independent post-event audits: IFRC and the Climate Centre verify that activations were justified and effective, feeding lessons into continuous model calibration.

Lessons learnt

This case demonstrates how proportional governance frameworks can transform high-risk predictive analytics into a safe public good. By combining probabilistic AI with documented oversight and multi-agency review, Bangladesh's FbF model strengthens both efficiency and equity. It shortens delays between forecast and delivery without removing human judgment, reflecting a balanced Assist-Advise delegation mode consistent with HOTL guidance. Its transparency and open documentation offer a transferable template for adaptive social-assistance systems in other climate-vulnerable contexts.

These guiding principles align with a wide range of global normative frameworks on the topic, including the EU Charter of Fundamental Rights, operationalised through General Data Protection Regulation (GDPR) (European Union, 2012, 2016). They are also aligned with OECD AI Principles and UNESCO Ethics Recommendation (OECD, 2019; UNESCO, 2021), while remaining adaptable to local contexts (also see Table A1 in Annex 2).

At the same time, these principles remain fully adaptable to realities in LMICs contexts, where institutional capacity, data infrastructure, and legal protections vary widely. Implementation must, therefore, be proportionate and phased, prioritising non-negotiable safeguards that directly mitigate the most critical sources of risk: meaningful human review for high-impact decisions, transparency appropriate to system impact, accessible grievance and appeal mechanisms, manual fallback channels, and sovereign control over core data and systems (Taeihagh, 2025; Papagiannidis et al., 2022; Wirtz et al., 2022).

CHAPTER 3. HUMAN AND SOCIETAL IMPACTS

The five structural sources of AI risk – data, models, operations, governance, and market dependencies – are deeply interdependent, and failures in one domain often amplify weaknesses in another (Steimers & Schneider, 2022; Slattery et al., 2024). A model trained on biased data will produce unfair outcomes even if technically robust, strong governance cannot compensate for poor integration, and secure data practices can be undermined if market structures restrict auditability or migration.

These vulnerabilities accumulate across the AI lifecycle (OECD, 2019; NIST, 2023; Zhang et al., 2022). When multiple upstream weaknesses align – such as biased datasets, opaque logic, fragile infrastructure, weak oversight, or vendor-controlled environments – their combined effect magnifies downstream harm. For beneficiaries, these compounding risks manifest not as abstract technical issues, but as concrete impacts on rights, welfare, and trust (OHCHR, 2021; UNESCO, 2021). These harms extend beyond individuals, shaping broader social attitudes and institutional legitimacy (Zuiderwijk et al., 2021; Taeihagh, 2021; Wirtz et al., 2022).

The human and societal impacts, therefore, represent the downstream expression of the upstream technical and institutional risks: they capture how systemic vulnerabilities translate into real-world consequences, how upstream failures in data, models, operations, governance, or market structures materialise in people's lived experience. These failures are not abstract: they manifest as exclusion from essential benefits, discriminatory treatment, opaque or unexplained decisions, or a lack of meaningful avenues to contest errors. For social assistance, for example, where decisions determine access to basic support, such failures directly threaten fundamental rights, including to privacy, equality, dignity, due process, and the right to an effective remedy (OHCHR, 2021; Jobin et al., 2019; European Union, 2012, 2016).

People can experience AI failures in concrete, often severe ways:

- **Discriminatory outcomes** that disproportionately affect marginalised groups and undermine the right to equality and non-discrimination
- **Exclusion from benefits** due to misclassification, drift, or rigid automation
- **Loss of autonomy and dignity** when decisions lack transparency, explanation, or a human point of contact
- **Inability to challenge decisions,** especially when notice is missing, reasoning is unclear, or grievance channels are inaccessible, directly impacting on due-process rights and the right to an effective remedy
- **Privacy harms** arising from excessive data collection, unlawful linkage, or leaks of sensitive information
- **Psychological stress and stigma** including stress, fear, or stigma, triggered by intrusive profiling, automated investigations, or classification as 'high risk'

The effects of AI risk are not limited to individuals; they reshape the broader social and institutional environment and, as well as reinforcing structural inequities and eroding trust in public institutions, can lead to:

- **Administrative inefficiencies,** where poorly integrated systems increase the workload, complexity, or error rates for frontline staff
- **Reduced legitimacy** of automated decision-making across the welfare state, jeopardising the credibility of digital reforms
- **Deepened digital divides** further marginalising people with lower literacy, limited connectivity, or reduced ability to navigate digital interfaces
- **Political backlash, litigation and public controversy**

Each risk source affects individuals and institutions in distinct, but interconnected, ways. Data bias translates into discriminatory exclusion for beneficiaries and legal or reputational exposure for agencies. Opacity undermines people's autonomy and contestability, while creating accountability gaps for public authorities. Weak governance leads to rights violations for citizens and compliance failures for institutions. Poor robustness or security produce unsafe outcomes for households and operational disruptions for agencies. Vendor lock-in and system misuse generate surveillance and dependence risks for individuals, and a loss of control and political vulnerability for governments.

CHAPTER 4. SHIFTING RISKS ACROSS AI USE AND DEPLOYMENT

AI uses across the delivery chain

To understand the risks of AI in social protection, one must first understand where and how it is currently deployed. AI is not a speculative or future technology, but is already embedded in decision-making and frontline delivery systems (DCI AI Hub, 2026b). A distinctive pattern of adoption emerges across income settings. In low-income contexts, AI is primarily used to fill data gaps, drawing on satellite imagery, mobile network data, and climate-risk models to compensate for incomplete administrative registries. In high-income contexts, AI is concentrated in the 'Provide' and 'Manage' phases, where advanced administrative datasets support fraud detection, automated case management, and real-time monitoring (ISSA, 2025; DCI AI Hub, 2026b).

Table 1 outlines how AI is used across the four-phases of the delivery chain: assess, enrol, provide, and manage. While we focus here on implementation/delivery, it is important to remember many AI use cases also apply at the policy level and programme design level, as the AI Hub Taxonomy and Global Evidence Review (DCI AI Hub, 2026a and 2026b) thoroughly showcase.

TABLE 1. EXAMPLES OF THE USE OF AI IN THE FOUR PHASES OF THE SOCIAL PROTECTION DELIVERY CHAIN: ASSESS, ENROL, PROVIDE AND MANAGE

DELIVERY CHAIN PHASE	EXAMPLE	FUNCTIONAL SHIFT DUE TO AI	EMERGING RISKS
Assess	AI-enabled: User communication and interaction (e.g. chatbots); vulnerability, needs and risk assessment (e.g. via geospatial mapping), including predictive analytics (e.g. anticipatory action models); matching and recommendation; identification, verification and record linking	These tools shift the assessment of needs and conditions from static, survey-based methods to dynamic, continuously updated predictions, allowing faster and potentially more granular identification and the addressing of emerging needs, including via enhanced user interaction.	Where underlying data is biased or partial, reliance on proxies can systematically overlook marginalised groups, embedding exclusion beneath a veneer of quantitative objectivity. The opacity of proprietary or black-box models further limits the ability of governments or affected communities to understand, contest, or correct these patterns.
Enrol	AI-enabled: Identification, verification and record linkage (e.g. biometric deduplication); data quality and anomaly detection; decision support for determination of eligibility, and benefits and services packages; operational and process automation (e.g. Optical Character Recognition [OCR] methods for document processing)	Manual verification steps are replaced with AI-driven checks and decision-support designed to reduce leakage and speed up onboarding.	Biometric and document-recognition failures, due to poor connectivity, worn fingerprints, low image quality, or demographic bias, can block legitimate beneficiaries at the point of entry. In systems without robust manual override or clear appeal routes, a technical error effectively functions as a denial of rights (DCI AI Hub, 2026c).

DELIVERY CHAIN PHASE	EXAMPLE	FUNCTIONAL SHIFT DUE TO AI	EMERGING RISKS
Provide	AI enabled: Data quality and anomaly detection (e.g. payment-integrity algorithms that flag anomalous patterns in transaction data); identification, verification and record linkage (confirms identity and relevant documentation to comply with know-your-client [KYC] requirements); matching and recommendations (to flag the most appropriate payment channel etc.)	Systems move from passive, periodic disbursement to active, continuous monitoring of payment flows and, in some cases, of beneficiary behaviour.	Fraud-detection systems can reproduce harms, such as in SyRI (the Netherlands) and Robodebt (Australia): opaque suspicion scores, automated sanctions, and reversal of the burden of proof, leaving households to contest algorithmically inferred misconduct after income has been reduced or withdrawn (District Court of The Hague, 2020; Royal Commission into the Robodebt Scheme, 2023). Where such systems operate without meaningful human review, they undermine due process and can generate unlawful debts or wrongful suspensions in social assistance programmes.
Manage	AI enabled: Vulnerability, needs and risk assessment, including predictive analytics (predictive risk profiling to identify complex or multi-dimensional need cases that require specialist referral); operational and process automation (e.g. AI-assisted case file management)	Case management becomes partly algorithmic, with model-generated scores increasingly influencing which families receive attention and how that attention is framed.	These systems introduce and amplify automation bias, the tendency for caseworkers to defer to algorithmic recommendations even when these are poorly understood. Risk flags can predispose staff to view certain households with suspicion, hardening attitudes, increasing the likelihood of sanctions, and reducing the quality of care, often without beneficiaries having any visibility into or recourse against these classifications.

Use cases

The following sub-sections provide some examples of the risk-implications of AI use cases[6] and draw transferable lessons, acknowledging that these same use cases apply to other social protection pillars and functions too.

Institutional lessons from large-scale data integration – Safeguards in practice vs safeguards on paper

Large-scale data integration platforms that use AI-enabled entity resolution to link tax, property, and welfare databases highlight a recurring governance challenge: the gap between planned safeguards and their implementation in practice. In one well-documented example from South Asia, a state-level platform was designed with four robust safeguards: HOTL verification of all algorithmic flags before beneficiary action, transparency documentation through a data governance portal, independent random-sample audits, and role-based privacy controls (ITE&C Department, Government of Telangana, 2019) [Design claims cited to the government's own World Bank presentation (ITE&C, 2019), per DCI AI Tracker case IND-005 – TB].

In practice, institutional pressures and capacity constraints meant these safeguards were not consistently applied. Officials often accepted algorithmic determinations without meaningful review. The proprietary entity-resolution model operated with limited explainability, leaving administrators and beneficiaries unable to understand its logic. Appeals processes were complex and slow. The result was that erroneous linkages led to the exclusion of some vulnerable households from food-security and pension benefits (Al Jazeera, 2024; Amnesty International, 2024b).

Lessons: This experience offers the following lessons:

- **Requiring human review is not sufficient** – officials need the authority, training, time, and resources to challenge algorithmic outputs effectively.
- **Transparency must extend beyond publishing linkage rules to providing meaningful explainability** for administrators and beneficiaries.
- **Grievance mechanisms must operate at the same speed as automated determinations** to prevent harm from accumulating.
- **Vendor contracts should include accountability provisions** such as independent audits, explainability requirements, and access to performance metrics.

These lessons are not unique to any single country. They reflect systemic challenges that arise wherever automation scales faster than institutional capacity, and they directly inform the governance readiness assessment in Tool 2 of the Risks and Safeguards Toolkit that accompanies this document.

6 For AI use cases in the social protection sector and how these map to key social protection functions, at policy, programme design and delivery chain levels, we refer to DCI AI Hub, 2026b (see also Ohlenburg, 2020).

UNICEF Yemen UCT pilot – Human oversight in AI-enabled integrity management

The UNICEF Yemen unconditional cash transfer (UCT) pilot shows how anomaly detection can be deployed safely when human oversight remains decisive.

Beginning in 2023, in the midst of Yemen's humanitarian crisis, UNICEF and its partners introduced an unsupervised machine-learning model to analyse mobile-money transaction data. The goal was to flag potential irregularities such as duplicate SIM use, repeated withdrawals, or unusual transaction timing. Crucially, the system never issued automatic suspensions or exclusions. Every flagged case was reviewed manually before any action was taken (while noting risks may still emerge, as HITL can go into 'cognitive surrender'[7]). This showcases the importance of strong safeguards at the assess and provide stages of social assistance delivery (UNICEF Office of Innovation, 2025; UNICEF Venture Fund, 2026) [Case verified against UNICEF sources and the KP3 use-case inventory (LIC-YEM-001); start date and payment-data wording aligned to those sources – TB].

Lessons: This experience offers the following lessons:

- **Human oversight is a core safeguard:** The model served only to prioritise field audits, not to determine fraud outcomes. Each anomaly was validated by UNICEF payment agents and implementing-partner monitors, ensuring that decisions remained human-led. False-positive rates were tracked, and periodic model validation by UNICEF data teams reinforced accountability and proportionality. This approach demonstrated that AI can assist integrity management without undermining rights when governance is strong.
- **Operationalising rights in a high-risk context:** By combining anomaly analytics with full human adjudication and detailed documentation, the Yemen UCT pilot improved payment integrity and reduced administrative delays without infringing on beneficiaries' rights. It stands as a practical example of HITL governance in fragile environments, where automation risk is high, but operational safeguards can mitigate harm.
- **Closing the loop between entitlement and trust:** Reliable delivery depends on more than technology – it requires governance that links payments, communication, and grievance mechanisms under clear accountability. This also aligns with principles for digital public goods (DPGs): open standards, local control, interoperability, and transparent governance. These features turn responsible design into a reusable public asset, reducing vendor dependency and strengthening long-term accountability.

7 i.e. The frequent phenomenon whereby a user uncritically accepts AI-generated answers or decisions, replacing their own critical thinking and reasoning with the machine's output.

Bangladesh Dynamic Social Registry pilot – Balancing automation and human validation

Bangladesh's Dynamic Social Registry pilot offers a practical example of how automation can be integrated responsibly into continuous eligibility management for national cash-transfer programmes. The system links registry data, periodic survey updates, and community feedback to identify household changes that may affect eligibility for social assistance. Machine-learning models flag records for review, but cannot alter beneficiary status automatically, ensuring that decision-making authority remains firmly human-centred.

Lessons: This experience offers the following lessons:

- **Human oversight remains the core safeguard:** Programme field officers from the Department of Social Services validate all algorithmic flags through household visits or document checks. Each decision is logged and communicated to beneficiaries, who retain the right to contest outcomes through accessible grievance mechanisms. Independent audits by the General Economics Division and development partners assess model precision and bias annually, reinforcing accountability.

- **Turning automation into a rights-compliant tool:** By embedding explainability, documentation, and redress, the pilot transformed automation into an adaptive mechanism, improving inclusion while safeguarding dignity. The system identifies households requiring review without pre-determining outcomes, ensuring that algorithmic support accelerates, rather than replaces, human judgment. This approach operationalises key principles of privacy, fairness, dignity, accountability, and equality in a live social-assistance environment.

- **Adaptive management for resilient systems:** The pilot demonstrates how AI can support forecasting, monitoring, and evaluation within transparent oversight frameworks. When combined with strong governance, these tools enable social-assistance systems to evolve responsibly, balancing efficiency with rights protection. The experience underscores a broader lesson: automation should assist, not replace, human decision-making – especially in processes that determine access to essential entitlements.

PART 2
PUTTING THE FRAMEWORK INTO PRACTICE

This part provides a Companion Guide to the Risks and Safeguards Toolkit. It is designed to assess whether or not to introduce an AI system in social protection to determine whether to Go, Pilot or No-Go. The first section looks at the tools used to assess risk and the scoring method. This is followed by a suggested workflow and a word on monitoring and reassessment.

CHAPTER 5. RISK ASSESSMENT AND SCORING METHOD

Before an AI system interacts with beneficiaries in a social protection programme, especially social assistance, a risk assessment should be carried out (Ada Lovelace Institute & DataKind UK, 2020; NIST, 2023). This section outlines the rationale and methodology for the Risks and Safeguards Toolkit that accompanies this framework. The Toolkit requires implementers to assess risk systematically rather than relying on intuition, vendor claims, or political pressure, complementing this more theoretical Framework discussed in Part 1 of this document. It was designed to help teams take a deliberate pause to verify safety before going live.

The structure of the Toolkit is given in Figure 4 and consists of a combination of the following. Overall Risk is determined by technology characteristics and by the governance environment that surrounds deployment. These are separated through an assessment of Intrinsic Risk followed by adjustment using a Context Risk Multiplier, which can be recorded in a comprehensive Risk Register. Together, these form a structured basis for deciding whether an AI application is ready for deployment (Go), suitable only for a controlled pilot (Pilot), or must be halted (No-Go) until essential protections are in place. In summary:

- Tool 1: Intrinsic Risk assesses how hazardous the AI system is before considering any contextual safeguards. It focuses on what can go wrong, how serious the consequences might be, and how widely harm could spread. This ensures that potential impacts on rights, entitlements, and due process are fully understood at the outset.

- Tool 2: Context Risk Multiplier captures how far local institutions can manage or amplify these technical risks. It reflects the strength of the legal framework, institutional capacity, data ecosystem, and public trust. A system with Moderate Intrinsic Risk may be manageable in a high-capacity setting, but is unsafe when appeal rights are weak or staff cannot meaningfully oversee AI outputs.

- Tool 3: Risk Register serves as the central supervisory log for the entire process.

In other words, an assessment of Intrinsic Risk and the Context Risk Multiplier jointly determine Overall Risk and, thus, Go, Pilot, or No-Go decisions for AI deployment in social protection (Tool 4). This combined approach enables proportionate decisions, rather than one-size-fits-all rules. It also operationalises responsible non-adoption when minimum safeguards cannot be met, even when technology appears attractive or politically favoured.

WHO SHOULD USE THE TOOLKIT AND WHEN?

Who: AI governance requires shared responsibility across technical, legal, and operational teams. No single unit has all the expertise needed to evaluate risks that span data, models, oversight, and rights. The Toolkit, therefore, requires a **cross-functional review group:**

- **Programme leads** define the policy problem, intended users, and decision points affected by the proposed system.
- **IT and data teams** assess system behaviour, model stability, and data quality using Tool 1.
- **Legal, safeguards and oversight officers** examine mandates, rights protections, and grievance channels using Tool 2.
- **Senior leadership** reviews consolidated evidence in Tool 3 and authorises the final *Go*, *Pilot*, or *No-Go* decision using Tool 4.
- **Project owner** is the focal point who is leading on the project, using the toolkit as a reference at every stage and as a resource to make informed decisions in consultation with the entire team.

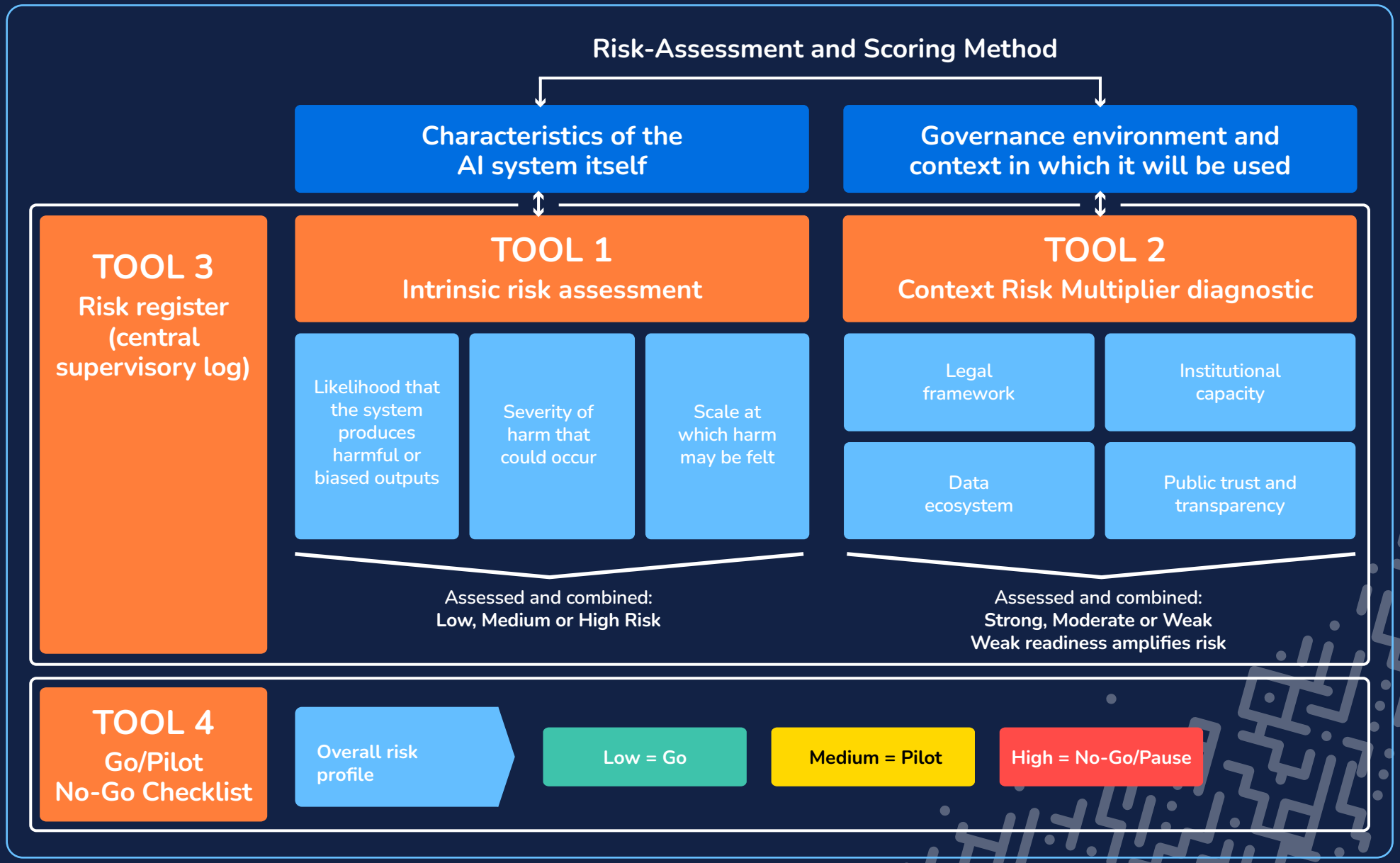
This ensures that any AI system influencing eligibility, exclusion, sanctions, or payment integrity receives meaningful human oversight and institutional scrutiny prior to deployment.

When: The Toolkit must be used at all points where system risk or uncertainty increases. Specifically:

- **At inception** – when any AI system is first proposed
- **Before piloting** – prior to use on real beneficiary or operational data
- **Before scale-up** – when transitioning from pilot environments to large-scale or national deployments
- **After major updates** – including changes to models, data sources, interoperability, vendor arrangements, or relevant legislation

Applying the Toolkit at these stages ensures that past assumptions are revalidated and that safeguards evolve alongside the technology.

FIGURE 4. THE OVERARCHING RISK FRAMEWORK



The five structural sources of risk identified in Section 2 are addressed across the Toolkit as follows (Table 2).

TABLE 2. FIVE STRUCTURAL SOURCES OF RISK AND TOOLS TO ADDRESS THEM

STRUCTURAL RISK SOURCE	PRIMARY TOOL	SECONDARY TOOL
Data-related risks	Tool 1 (Intrinsic Risk)	Tool 3 (Risk Register)
Model-related risks	Tool 1 (Intrinsic Risk)	Tool 3 (Risk Register)
Operational and system integration risks	Tool 1 (Intrinsic Risk)	Tool 2 (Context Risk Multiplier)
Governance and institutional oversight risks	Tool 2 (Context Risk Multiplier)	Tool 4 (Go/No-Go Checklist)
Market, sovereignty and industry structure risks	Tool 2 (Context Risk Multiplier)	Tool 4 (Go/No-Go Checklist)

The following sub-sections explore the process of applying the Toolkit to AI-related decision-making in social protection. Each focuses on a different Tool within the Toolkit, answering the questions: what is it, why is it important and what are its key dimensions?

Tool 1. Intrinsic Risk

Intrinsic Risk is the starting point for all subsequent risk and safeguards decisions. It evaluates the AI system on its own terms: its data, model behaviour, and operational use. The assessment covers three dimensions: (i) the likelihood of harmful or biased outputs, (ii) the severity of harm if errors occur, and (ii) the scale of impact across people and programmes (also see Figure 4).

- Likelihood captures how often the system may generate harmful or biased outputs in real operation. It is driven primarily by data quality, model design, and operational robustness. Models trained on biased, incomplete, or outdated data are more likely to misclassify households or generate unstable predictions. Generative AI models introduce a specific likelihood of hallucinations, inventing facts or regulations that never existed. In practice, this might include a chatbot providing incorrect eligibility advice or a predictive model, systematically excluding certain communities that are underrepresented in training data.
- Severity describes how serious the consequences become once errors occur in the system. Severity is shaped mainly by the type of decision and the degree of automation. Advisory systems, such as decision-support dashboards, generally have lower severity because staff can identify and correct errors before they reach beneficiaries. Systems that directly influence eligibility, sanctions, or payments carry high severity because mistakes can cause wrongful exclusion, unlawful debt, or destitution. Severity increases when harm is irreversible, affects fundamental rights, or people cannot understand or challenge outcomes.

Note: The sensitivity of the data processed also directly affects severity. Systems that rely on sensitive personal data, including biometric identifiers, health or disability status, ethnicity, or detailed household financial records, carry inherently higher severity because errors, breaches, or misuse of such data can cause disproportionate harm, including discrimination, stigma, or irreversible privacy violations. Where a system processes no personally identifiable or sensitive data, severity on this factor is correspondingly lower.

- Scale reflects how many people or programmes could be affected by failures in the system. An error confined to a small, time-bound pilot affects relatively few people and can be contained. Mistakes in a national eligibility model or automated payment-verification system can affect millions. Scale also increases when the system sits at a critical junction in the delivery chain, for example, when eligibility scoring feeds directly into payment systems or triggers sanctions across multiple programmes.

Intrinsic Risk assessment must be documented using structured evidence, rather than assumptions. Teams should use test results, validation logs, coverage statistics, and rights-impact analysis to justify each rating.

Tool 1 provides a standard worksheet for capturing this evidence and assigning qualitative scores. It supports teams to examine the AI system itself: how it behaves, the quality of the data it relies on, and the kinds of harm that could occur if it fails. It uses the three intrinsic dimensions, likelihood, severity, and scale, to arrive at a qualitative risk tier of low, medium, or high. The tool provides prompts and guiding questions to help teams reflect on how often the system might produce errors, how serious those errors could be for beneficiaries, and how widely harm could spread. A good practice is to base ratings on available evidence such as validation results, performance tests, or rights-impact analysis; to note where uncertainties remain; and to apply a precautionary approach where evidence is weak or relies solely on vendor claims. The output of Tool 1 is a clearly documented Intrinsic Risk profile that forms the foundation for the rest of the assessment.

TABLE 3. KEY DIMENSIONS OF INTRINSIC RISK (TOOL 1)

DIMENSION	DESCRIPTION	EVIDENCE/INPUTS REQUIRED	QUALITATIVE RATING
Likelihood	How likely is the system to produce harmful or biased outcomes?	Test results, bias audits, validation logs, red-team exercises	Low / medium / high
Severity	If it fails, how serious is the harm to rights or entitlements?	Rights-impact assessment, programme rules, oversight model	Low / medium / high
Scale	How many people or programmes could be affected?	Coverage data, integration map, population reach	Low / medium / high

It is also important that:

- **Intrinsic Risk ratings must be aggregated conservatively using a simple maximum rule.** Each dimension is scored qualitatively, and the overall Intrinsic Risk level is anchored in the highest of the three scores. This prevents serious risks, such as the possibility of wrongful exclusion at a national scale, from being diluted by lower scores elsewhere. Intrinsic Risk aggregation rules must prevent high-severity scenarios from being understated. Low Risk applies only when likelihood, severity, and scale are all low. Medium Risk applies when at least one dimension is medium and none are high. High Risk applies when any dimension is high, or when all three dimensions are medium.

- **Uncertainty must be treated as a risk, not ignored due to missing evidence.** Where information is weak or incomplete, teams should apply the precautionary principle and raise the rating for that dimension. This ensures safeguards are designed for real-world uncertainty, rather than idealised assumptions.

In practice, most national-scale AI systems will initially score High on the scale dimension, producing an overall High Intrinsic Risk rating. This is by design. Systems that affect hundreds of thousands or millions of people, or that connect to critical infrastructure such as social registries or payment systems, warrant the highest level of scrutiny. The appropriate response is not to abandon the system, but to begin with a controlled pilot at a reduced scale, through which the scale dimension can be assessed as Low or Medium. If the pilot demonstrates acceptable performance across all dimensions, the Toolkit should be reapplied before each subsequent expansion. This staged pathway ensures that large-scale systems are never deployed without proportionate evidence of safety.

Tool 2. Context Risk Multiplier

The Context Risk Multiplier adjusts Intrinsic Risk based on the strength of country and institutional safeguards.

It recognises that the same AI system behaves very differently in a well-regulated, well-resourced environment than in a fragile system with weak rights protections. The multiplier is assessed across four factors: (i) legal framework, (ii) institutional capacity, (ii) data ecosystem, and (iii) public trust and transparency (also see Figure 4).

- **Legal framework** determines whether or not beneficiaries have enforceable rights against harmful automated decisions. Strong contexts have clear legal bases, explicit mandates, robust data-protection rules, and practical rights to appeal algorithm-assisted decisions. Weak contexts feature ambiguous authorisation, limited regulatory oversight, or grievance systems that exist on paper, but not in practice.

- **Institutional capacity** determines whether staff can supervise AI outputs, rather than deferring to them. Sustained oversight requires clear accountability, sufficient staffing, technical competence, and routine audits. Thin capacity increases the risk of automation bias, where busy or underconfident staff simply rubber-stamp AI recommendations without meaningful review (ICRC, 2024). This converts advisory tools into de facto automated decision makers, even when policy requires HITL control.

- **The data ecosystem** shapes how predictable and controllable the system remains over time. Reliable, interoperable, and locally controlled data enables stable model behaviour and effective correction of errors. Fragmented registries, weak security, or heavy reliance on vendor-hosted data increase uncertainty, reduce sovereignty, and make it harder to diagnose and fix failures.

- **Public trust and transparency** affect whether people use systems, challenge errors, or disengage in frustration. High-trust environments maintain transparent communication, accessible grievance channels, and meaningful engagement

with stakeholders. Low-trust settings feature black-box deployments, fear among staff or beneficiaries, and underused appeal routes. These conditions increase the risk that harms go undetected and unresolved.

Context readiness must be rated for each factor before adjusting the Overall Risk tier.

Tool 2 provides a diagnostic checklist for scoring each factor as Strong, Moderate, or Weak based on observable indicators. This supports consistent scoring across programmes and countries.

TABLE 4. KEY DIMENSIONS OF CONTEXT RISK MULTIPLIER (TOOL 2)

DIMENSION	INDICATORS OF READINESS	STRONG READINESS (MITIGATES RISK)	WEAK READINESS (AMPLIFIES RISK)
Legal framework	Data protection; appeal rights; lawful mandate	Enforced legal basis; clear authorisation; active supervision	Unclear mandates; weak appeal rights; inactive regulator
Institutional capacity	Oversight bodies; staffing; audits	Named accountability; trained staff; routine reviews	No clear owner; thin staffing; automation bias risk
Data ecosystem	Quality; localisation; interoperability	Reliable, locally controlled data; strong security	Fragmented registries; vendor control; weak safeguards
Public trust	Transparency; communication; grievance uptake	Open communication; active grievance handling	Opaque systems; low confidence; inaccessible channels

Context Risk Multiplier aggregation must be rule based rather than ad hoc.

Strong context applies where at least three dimensions are rated Strong and none are Weak and may justify lowering the Overall Risk tier by one level.

Moderate context applies where strengths and gaps balance each other and leaves the intrinsic tier unchanged. Weak context applies where two or more factors are Weak and requires raising the Overall Risk tier by one level.

Tool 3. Risk Register

Tool 3 The Risk Register serves as the central supervisory log for the entire process. While Tools 1 and 2 identify risks, the register documents them, assigns responsibility, tracks mitigation, and records evidence that risks have been addressed. It is, therefore, both a governance tool and an accountability mechanism.

- **It is good practice is to record all risks that emerge from the intrinsic and context assessments,** as well as those identified through DPIAs or AIAs, audits, testing, stakeholder feedback, and ongoing monitoring. Each entry in the register should include a short description of the risk, its likelihood and severity, the rights it touches, and any link to Context Risk Multiplier governance factors. It should also specify the mitigation required, the team responsible, a timeline, and the evidence that will confirm when the risk is resolved.

- **Each risk should also be tagged with its decision implication.** A blocking risk (must be resolved before deployment), pilot condition (acceptable for controlled deployment with monitoring), or monitor (acceptable for Go with ongoing tracking). This tagging ensures that the Risk Register serves as the single source of truth for the Go/Pilot/No-Go decision in Tool 4.

- **The register is not a once-off compliance document.** It should remain active throughout the system's lifecycle and be updated whenever new risks appear, when conditions change, or when mitigation measures are complete. It is reviewed at governance meetings and again before any new pilot or scale-up.

Many of the categories used in traditional project Risk Registers are relevant here: technical reliability, legal compliance, operational continuity, data quality and security, human oversight capacity, and grievance pathways. What distinguishes AI Risk Registers is the emphasis on rights impacts, model behaviour, and transparency. Maintaining a live, documented record of these risks strengthens accountability and ensures that teams can demonstrate not only that risks were identified, but also how they were addressed.

Tool 4. Overall Risk

Overall Risk combines Intrinsic Risk and the Context Risk Multiplier to guide the decision to Go, Pilot or No-Go: Tool 4 is thus the final decision gate. The resulting overall profile guides whether a system should be deployed, piloted under tight control, or paused. This ensures that Go, Pilot, and No-Go decisions are evidence based, documented, and repeatable across different AI proposals.

Tools 1 and 2 use different rating scales, because they measure different things. Tool 1 rates Intrinsic Risk as Low, Medium, or High, where High means greatest concern. Tool 2 rates institutional readiness as Strong, Moderate, or Weak, where Strong means greatest capacity to manage risk safely. A High Intrinsic Risk combined with Weak institutional readiness produces the most serious concern; Low risk in a Strong context is the safest profile. The matrix below shows how the two scales interact to produce the combined risk tier.

- **Low Overall Risk justifies deployment with proportionate safeguards and routine monitoring.** Systems in this category should still have documented oversight arrangements, but do not require intensive auditing or exceptional controls. HOTL supervision is usually sufficient, with clear procedures for escalation when anomalies appear.

- **Medium Overall Risk requires controlled piloting and structured learning before any scale-up.** These systems should be deployed only in limited pilots with strict eligibility boundaries, enhanced monitoring, and HITL verification for rights-critical decisions. Independent reviews should be scheduled, and continuation should depend on documented evidence of safety and effectiveness.
- **High Overall Risk requires a deliberate pause and stronger safeguards before any deployment.** For these systems, responsible non-adoption is the default until legal, institutional, and technical safeguards are demonstrably improved. This includes clear lawful mandates, robust human oversight, independent audit, and the ability to suspend or reverse decisions where harm is detected.

Tool 4 The Go, Pilot, No-Go logic must be captured transparently in the Tool 4 Toolkit templates. Tool 3, the Risk Register, and Tool 4, the Go or No-Go checklist, ensure that the reasoning behind each decision is recorded. This enables later audit, supports public accountability, and helps teams avoid repeating past failures when new AI systems are proposed.

- **The checklist begins with the adjusted risk tier.** A system with a low tier may be considered for deployment under standard monitoring; a medium tier typically supports a pilot; and a high tier generally results in a No-Go decision. However, this preliminary classification must be validated against a set of minimum requirements.
- **The checklist confirms that essential safeguards are complete.** These include lawful basis for data processing; completion of a DPIA or AIA (with at least a public summary available) (Ada Lovelace Institute & DataKind UK, 2020; ICRC, 2024); clear human oversight arrangements; and a clear prohibition on automating eligibility, exclusion, sanctions, or appeal decisions. It also requires that named human decision makers are assigned at critical points; that beneficiaries have access to notice, explanation, and appeal; and that time-bound, trackable grievance channels are operational.
- **The checklist also requires confirmation that all risks flagged as blocking in the Risk Register (Tool 3) have been resolved or mitigated to an acceptable level.** No outstanding blocking conditions may remain unaddressed at the point of decision.
- **Further checks ensure that data governance measures are in place, including documentation of inputs and assumptions, model versioning, data minimisation, security controls, and retention policies (ICRC, 2024).** A monitoring plan must be defined, including thresholds for errors or bias and clear responses when those thresholds are exceeded. Deployment scope must also be explicit, covering geography, population, and timeframe, as well as stop conditions defined.

If any item on the checklist is not met, the system is automatically classified as No-Go, regardless of the adjusted risk tier. This ensures that minimum standards are non-negotiable, especially where rights are at stake.

The final decision – Go, Pilot, or No-Go – should be recorded in writing, including a short narrative justification and references to evidence. A good practice is to require sign-off from multiple units, including legal, programme, technical, and oversight representatives, to ensure that responsibility is shared and transparent.

RESPONSIBLE NON-ADOPTION IS NOT A FAILURE!

It is important to realise that a No-Go verdict is not a failure of innovation, but an example of responsible non-adoption, the appropriate outcome when safeguards or governance readiness are insufficient. Some AI systems should not be deployed when risks exceed the institution's ability to manage them safely, whether because data foundations are too weak, the lawful basis for processing is unclear, grievance and appeal systems cannot provide timely remedies or because governance safeguards are not yet strong enough to prevent or correct harmful outcomes.

In these situations, choosing not to proceed protects beneficiaries, maintains trust in the social protection system, and avoids the cost and harm of intervening after problems occur. Responsible non-adoption allows institutions to strengthen data quality, legal frameworks, and oversight structures before revisiting the use case. It ensures that AI is introduced only when benefits are clear, risks are manageable, and rights can be safeguarded effectively.

CHAPTER 6. SUGGESTED WORKFLOW

To help institutions understand how the Toolkit operates in practice, a simple workflow is proposed that shows how the four tools fit together. The process begins with defining the use case and ends with a documented, auditable deployment decision. Each step is led by a specific team, but requires shared review and accountability. Table 5 summarises the recommended sequence:

TABLE 5. TOOLKIT WORKFLOW, FROM IDENTIFICATION TO DECISION

PHASE	TOOL/ACTION	RESPONSIBLE	PURPOSE
1 **Identification & scoping** *Is AI being proposed?*	**Tool 1** Pre-Assessment Screening	Programme lead; IT/ management information system (MIS) director	Confirm proportionality, do-no-harm, and accountability before detailed assessment. Flag fundamental red lines.
2 **Technical assessment** *How risky is the system?*	**Tool 1** Intrinsic Risk Worksheet	Data science/IT team	Assess likelihood, severity, and scale of potential harm: produces Intrinsic Risk Profile – Low/ Medium /High.
3 **Governance readiness** *Are we ready for this?*	**Tool 2** Context Risk Multiplier	Policy/legal officers; data protection officer (DPO)/ oversight	Evaluate legal framework, institutional capacity, data ecosystem, and public trust: produces Context Risk Multiplier tier – Strong/Moderate/Weak.
4 **Risk synthesis** *Document and assign*	**Tool 3** Risk Register	Cross-functional working group	Consolidate all risks, assign mitigations, owners, and dead-lines: identify blocking vs monitoring conditions.

...

PHASE	TOOL/ACTION	RESPONSIBLE	PURPOSE
5 **Decision conference** *Should we proceed?*	**Tool 4** Go / Pilot / No-Go Checklist	Inter-agency steering committee	Apply decision matrix, confirm minimum safeguards: lawful basis, human oversight, grievance route, transparency.
6 **Endorsement** *Authorise and document*	Formal sign-off and publication	Senior management; regulator (where applicable)	Finalise deployment decision: archive documentation for transparency, audit, and institutional memory.

Go

Low overall risk. Deploy with standard safeguards, routine monitoring, and annual audit.

Pilot

Medium overall risk. Controlled, time-bound deployment with enhanced oversight and HITL verification.

No-Go

High overall risk. Defer until legal, institutional, or technical safeguards are demonstrably improved.

Note: The Toolkit must be reapplied at each stage at which risk or uncertainty increases: at inception, before piloting, before scale-up, and after major system, data, or legislative changes. The Risk Register (Tool 3) remains active throughout the system lifecycle.

This workflow illustrates how the Toolkit operates as an integrated sequence rather, than a set of standalone tools. The programme unit begins by clarifying the purpose and scope of the AI system, setting the basis for all subsequent assessment. Technical specialists then assess Intrinsic Risk using Tool 1, determining the system's inherent exposure before any safeguards are considered. Legal, policy, and oversight teams follow with Tool 2 to assess governance readiness, ensuring that the institution has the authority, capacity, and data foundations required to manage the identified risks.

Once these assessments are complete, a cross-functional working group consolidates all risks and mitigation actions into the Risk Register, which becomes the central record that senior decision makers will rely on. The inter-agency steering committee then reviews the evidence, applies the decision matrix in Tool 4, and determines whether the system can proceed, should be piloted under enhanced oversight, or must pause until safeguards improve. The final step is formal endorsement and publication, which ensures transparency and provides a documented audit trail for internal and external oversight.

Statement on Scope: ***What the Toolkit is and is not.*** **The Toolkit is designed to assess the pre-deployment risk of an AI system in non-contributory social assistance and to inform the final decision on whether to proceed. It can be adapted to other social protection domains with expert judgement. It is not a substitute for legal compliance, a data protection impact assessment, or procurement due diligence, and it is not a technical build specification.**

CHAPTER 7. MONITORING AND REASSESSMENT

Approval to deploy or pilot an AI system is not the end of the governance process. Once a system is in use, teams must continue to monitor performance, update the Risk Register, and revisit the assessments whenever conditions change. New training data, model updates, integration with additional registries, shifts in programme rules, or changes in legal frameworks all introduce new uncertainties that may alter both intrinsic and contextual risk.

Good practice is to reapply the Toolkit at regular intervals, such as during annual audits, or sooner if evidence suggests emerging risks. This ensures that safeguards remain proportionate as systems evolve and that institutional oversight keeps pace with technical and operational changes. Continuous monitoring helps detect early signs of bias, drift, or unintended effects, allowing timely correction before harm occurs.

By treating deployment as the start of an ongoing cycle, rather than a once-off decision, institutions maintain the transparency, accountability, and rights protections that responsible AI adoption in social protection requires.

CHAPTER 8. CONCLUSION

Artificial intelligence holds significant promise for strengthening social protection systems. AI can improve targeting accuracy, support early crisis response, streamline registration and verification, and enhance programme integrity. However, it is only beneficial when deployed responsibly, with safeguards that are proportionate to the level of risk, and in ways that firmly preserve human judgment in all rights-affecting decisions. AI must never replace trained staff in decisions about eligibility, sanctions, or appeals. These decisions require discretion, contextual understanding, and respect for due process, which are elements that no automated system can fully replicate.

The risks of poorly governed AI are real and well documented. The use of AI in social protection can amplify existing inequalities and unintentionally exclude vulnerable groups, reduce contestability and limit institutional oversight, and errors caused by automation can go uncorrected, causing harm to households that rely on assistance. These risks are heightened in low- and middle-income settings, where institutional capacity, legal protections, and data governance frameworks may still be evolving.

This report offers a practical way to navigate these opportunities and risks. This Risk Assessment Framework and accompanying Risks and Safeguards Toolkit translate global principles into a structured method that institutions can use to evaluate AI use cases, assess governance readiness, and apply safeguards that match the level of exposure. By separating intrinsic technical risk from contextual institutional capacity, the method encourages careful, evidence-based reflection on both the capabilities of the technology and the readiness of the surrounding system. The four-tool workflow supports transparent, repeatable, and auditable decision making, which helps teams demonstrate not just that AI is useful, but that it is safe, lawful, and accountable.

The Framework's central conclusion is that responsible AI governance includes responsible non-adoption. When Intrinsic Risks are high, or when contextual safeguards are weak, a No-Go decision is not a setback, it is a necessary protection for beneficiaries and for the integrity of social protection systems. Pausing deployment until safeguards are in place is often the most responsible course of action. Ultimately, the goal is not faster deployment, but safer and more accountable deployment. AI should not be used when it would automate rights-critical decisions; when lawful processing bases are unclear; when oversight capacity is insufficient for required monitoring; when data quality cannot support reliable model performance; or when vendor arrangements undermine sovereignty, auditability, or exit options. In such cases, pausing or declining adoption protects beneficiaries, avoids irreversible harm, and allows governments to strengthen foundational systems before revisiting the use case.

At its core, responsible AI in social assistance is about protecting people. With transparency, accountability, data sovereignty, and meaningful human oversight at the centre, AI can help governments deliver support more quickly, more fairly, and with greater integrity, thereby reinforcing both social protection systems and public trust in an increasingly digital age. By grounding adoption in these principles, governments can harness innovation while maintaining the fairness, dignity, and accountability that social protection systems are built to uphold.

REFERENCES

Ada Lovelace Institute and DataKind UK, *Examining the black box: Tools for assessing algorithmic systems*, Ada Lovelace Institute, London, 2020, https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Al Jazeera, 'How an algorithm denied food to thousands of poor in India's Telangana', *Al Jazeera*, 24 January 2024, https://www.aljazeera.com/economy/2024/1/24/how-an-algorithm-denied-food-to-thousands-of-poor-in-indias-telangana

Amnesty International, *Coded injustice: Surveillance and discrimination in Denmark's automated welfare state*, Amnesty International, London, 2024a

Amnesty International, *Use of Entity Resolution in India: Shining a light on how new forms of automation can deny people access to welfare*, Amnesty International, London, 2024b, https://www.amnesty.org/en/latest/research/2024/04/entity-resolution-in-indias-welfare-digitalization/

Anticipation Hub, 'Anticipatory action in Bangladesh', *Anticipation Hub*, accessed June 2026, https://www.anticipation-hub.org/global-overview/countries/bangladesh

Banerjee, I., Bhattacharjee, K., Burns, J., et al., '"Shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation', *Journal of the American College of Radiology*, 20(10), 2023, https://doi.org/10.1016/j.jacr.2023.06.025

Bangladesh Humanitarian Country Team, *Anticipatory Action Framework: Monsoon Floods*, April 2025, https://reliefweb.int/attachments/b9e1bf90-a96b-490f-bfc7-a31e98de669d/Bangladesh%20AA%20Framework%20-%20Flood%20-%20April%202025%20-%20FINAL%20.pdf

Bengio, Y., Hinton, G., Yao, A., et al., 'Managing extreme AI risks amid rapid progress', *Science*, 384(6698), 842–845, 2023, https://doi.org/10.1126/science.adn0117

Chu, C., Donato-Woodger, S., Khan, S., et al., 'Age-related bias and artificial intelligence: A scoping review', *Humanities and Social Sciences Communications*, 10, 1–17, 2023, https://doi.org/10.1057/s41599-023-01999-y

DCI, *A Data Governance Framework for Digital Social Protection Systems*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, 2025, https://spdci.org

DCI AI Hub for Social Protection, *A Taxonomy for AI in Social Protection: Defining Capabilities and Application Areas*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Germany, 2026a

DCI AI Hub for Social Protection, *AI Adoption in Social Protection: A Global Evidence Review (2020–2025)*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Germany, 2026b

DCI AI Hub for Social Protection, *AI in Social Protection: Data Sovereignty Considerations*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Germany, 2026c

District Court of The Hague, *Judgment in Case C-09-550982-HA ZA 18-388 (SyRI)*, The Hague, 2020

Eken, B., Pallewatta, S., Tran, N., Tosun, A., and Babar, M., 'A multivocal review of MLOps practices, challenges and open issues, *ACM Computing Surveys*, 58, 1–35, 2024, https://doi.org/10.1145/3747346

Eubanks, V., *Automating inequality: How high-tech tools profile, police, and punish the poor*, St Martin's Press, New York, 2018

European Commission, *Ethics guidelines for trustworthy AI*, High-Level Expert Group on AI, Brussels, 2019

European Union, *Charter of Fundamental Rights of the European Union (2012/C 326/02)*, Official Journal of the European Union, 2012

European Union, *Regulation (EU) 2016/679 (General Data Protection Regulation)*, Official Journal of the European Union, 2016

European Union, *Artificial Intelligence Act*, Council of the European Union, Brussels, 2024

Ferrara, E., 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies', *Sci*, 6(1), 3, 2023, https://doi.org/10.3390/sci6010003

Fu, Y., Huang, Z., Deng, X., et al., 'Artificial intelligence in lymphoma histopathology: Systematic review', *Journal of Medical Internet Research*, 27, e62851, 2024, https://doi.org/10.2196/62851

Gray, M., Samala, R., Liu, Q., et al., 'Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science', *Clinical Pharmacology & Therapeutics*, 115, 2023, https://doi.org/10.1002/cpt.3117

Gros, C., et al., 'Household-level effects of providing forecast-based cash in anticipation of extreme weather events: Quasi-experimental evidence from humanitarian interventions in the 2020 floods in Bangladesh', *International Journal of Disaster Risk Reduction*, 82, 103388, 2022, https://doi.org/10.1016/j.ijdrr.2022.103388

He, P. and Wang, X., Ethical risk research on the interpretability of AI algorithms, *Proceedings of the 4th International Conference on Artificial Intelligence, Internet and Digital Economy*, 100–105, 2025, https://doi.org/10.1109/icaid65275.2025.11034567

Hiniduma, K., Byna, S. and Bez, J., *Data readiness for AI: A 360-degree survey*, ACM Computing Surveys, 2024, https://doi.org/10.1145/3722214

IFRC, 'Protecting families in Bangladesh against floods: Early Action Protocol activated', International Federation of Red Cross and Red Crescent Societies, 2020, https://www.forecast-based-financing.org/2020/07/02/protecting-families-in-bangladesh-against-floods-early-action-protocol-activated/

ILO, Governing *AI in the world of work: A review of global ethics guidelines*, International Labour Organization, Geneva, 2023

International Committee of the Red Cross (ICRC), *Handbook on data protection in humanitarian action (3rd Edition)*, Edited by Massimo Marelli, Cambridge University Press, Cambridge, 2024, https://doi.org/10.1017/9781009414630

ISSA, *AI applications in social security: Evidence and insights from the TechByte*, International Social Security Association, Geneva, 2025

ISSA, *Guidelines on Information and Communication Technology*, revised edition, International Social Security Association, Geneva, 2019, https://www.issa.int/guidelines/ict

ITE&C Department, Government of Telangana, *Samagra Vedika – Telangana's Integrated Platform*, World Bank-hosted presentation, 2019, https://thedocs.worldbank.org/en/doc/945071576869997489-0310022019/original/GTVenkateshwarRaoPresentationonSamagraVedikatoWordlBankseminatDec19.pdf

Jobin, A., Ienca, M. and Vayena, E., 'Artificial intelligence: The global landscape of ethics guidelines', *Nature Machine Intelligence*, 1, 389–399, 2019, https://doi.org/10.1038/s42256-019-0088-2

Leslie, D., and Perini, A., 'Future Shock: Generative AI and the International AI Policy and Governance Crisis', *Harvard Data Science Review*, Special Issue 5: Grappling With the Generative AI Revolution, 2024, https://doi.org/10.1162/99608f92.88b4cc98

Lindert, K., George Karippacheril, T., Rodríguez Caillava, I., and Nishikawa Chávez, K., *Sourcebook on the foundations of social protection delivery systems*, World Bank, Washington, DC, 2020

Mauri, L. and Damiani, E., 'Modeling threats to AI-ML systems using STRIDE', *Sensors*, 22(17), 6610, 2022, https://doi.org/10.3390/s22176662

Mohammed, S., Budach, L., Feuerpfeil, M., et al., 'The effects of data quality on machine learning performance on tabular data', *Information Systems*, 132, 102549, 2022, https://doi.org/10.1016/j.is.2022.102549

Munz, P., Hennick, M. and Stewart, J., 'Maximizing AI reliability through anticipatory thinking and model risk audits', *AI Magazine*, 44(2), 173–184, 2023, https://doi.org/10.1002/aaai.12099

Nazer, L., Zatarah, R., Waldrip, S., et al., 'Bias in artificial intelligence algorithms and recommendations for mitigation', *PLOS Digital Health*, 2(9), e0000278, 2023, https://doi.org/10.1371/journal.pdig.0000278

NIST, *Artificial intelligence risk management framework (AI RMF 1.0)*, National Institute of Standards and Technology, Gaithersburg, MD, 2023

Norori, N., Hu, Q., Aellen, F., Faraci, F., and Tzovara, A., 'Addressing bias in big data and AI for healthcare: A call for open science', *Patterns*, 2(9), 100347, 2021, https://doi.org/10.1016/j.patter.2021.100347

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al., 'Bias in data-driven artificial intelligence systems: An introductory survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1356, 2020, https://doi.org/10.1002/widm.1356

OECD, *OECD recommendation on artificial intelligence*, OECD Publishing, Paris, 2019

OHCHR, *The right to privacy in the digital age (A/HRC/48/31)*, Office of the United Nations High Commissioner for Human Rights, Geneva 2021

OCHA, *OCHA data responsibility guidelines*, United Nations Office for the Coordination of Humanitarian Affairs, Centre for Humanitarian Data, Geneva, 2025, https://data.humdata.org/dataset/2048a947-5714-4220-905b-e662cbcd14c8/resource/8bc5b848-8ece-4f1f-a78b-18dd972bb21a/download/data-responsibility-guidelines-2025.pdf?_gl=1*1vtc8z*_ga*NjgzMTY0NjgxLjE3Njg0OTAwMzU.*_ga_E60ZNX2F68*czE3Njg1NjU5NTEkbzIkZzEkdDE3Njg1NjY2MzYkajYwJGwwJGgw

Ohlenburg, T., *AI in social protection – exploring opportunities and mitigating risks*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) & Asian Development Bank (ADB), 2020

Pagano, T., Loureiro, R., Lisboa, F., et al., 'Bias and unfairness in machine learning models: A systematic review', *Big Data and Cognitive Computing*, 7(1), 15, 2023, https://doi.org/10.3390/bdcc7010015

Paleyes, A., Urma, R., and Lawrence, N., 'Challenges in deploying machine learning: A survey of case studies', *ACM Computing Surveys*, 55, 1–29, 2020, https://doi.org/10.1145/3533378

Papagiannidis, E., Enholm, I., Dremel, C., Mikalef, P., and Krogstie, J., 'Toward AI governance: Identifying best practices and potential barriers and outcomes', *Information Systems Frontiers*, 25, 123–141, 2022, https://doi.org/10.1007/s10796-022-10251-y

Pirrone, C., Fricano, S., and Fazio, G., 'Exploring Vulnerability in AI Industry', *ArXiv*, abs/2510.23421, 2025, https://doi.org/10.48550/arxiv.2510.23421

Royal Commission into the Robodebt Scheme, *Report of the Royal Commission into the Robodebt Scheme*, Commonwealth of Australia, Canberra, 2023.

Schmitz, A., Mock, M., Görge, R., Cremers, A. and Poretschkin, M., A global scale comparison of risk aggregation in AI assessment frameworks, *AI and Ethics*, 5, 1407–1432, 2024, https://doi.org/10.1007/s43681-024-00479-6

Singh, M., 'Integrating artificial intelligence with legacy systems: A systematic analysis of challenges and strategic considerations', *European Journal of Computer Science and Information Technology*, 2025, https://doi.org/10.37745/ejcsit.2013/vol13n323845

Slattery, P., Saeri, A., Grundy, E., et al., 'The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence', *arXiv preprint*, arXiv:2408.12622, 2024, https://doi.org/10.48550/arXiv.2408.12622

Sribhashyam, V., 'Key challenges and limitations of MLOps in context of machine learning', *International Journal of Scientific Research in Engineering and Management*, 2025, https://doi.org/10.55041/ijsrem46346

Steimers, A. and Schneider, M., 'Sources of risk of AI systems', *International Journal of Environmental Research and Public Health*, 19(6), 3611, 2022, https://doi.org/10.3390/ijerph19063641.

Taeihagh, A., 'Governance of artificial intelligence', *Policy and Society*, 40(2), 137–157, 2021, https://doi.org/10.1080/14494035.2021.1928377

Taeihagh, A., *Governance of generative AI*, Policy and Society, 2025, https://doi.org/10.1093/polsoc/puaf001

UNESCO, *Recommendation on the ethics of artificial intelligence*, UNESCO, Paris, 2021

UNICEF Office of Innovation, 'AI for accountability: How artificial intelligence is strengthening humanitarian cash transfers in Yemen', UNICEF, 24 November 2025, https://www.unicef.org/innovation/stories/ai-accountability

UNICEF Venture Fund, 'AI anomaly detection of cash transfers – Yemen', accessed June 2026, https://www.unicefventurefund.org/portfolio/ai-anomaly-detection-cash-transfers-yemen

Uuk, R., Brouwer, A., Schreier, T., et al., 'Effective mitigations for systemic risks from general-purpose AI', *Superintelligence – Robotics – Safety & Alignment*, 2(1), 2025, https://doi.org/10.70777/si.v2i1.13975

Van Der Vlist, F., Helmond, A., & Ferrari, F., 'Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence', *Big Data & Society*, 11, 2024, https://doi.org/10.1177/20539517241232630

Vijayan, N., 'Building scalable MLOps: Optimizing machine learning deployment and operations', *International Journal of Scientific Research in Engineering and Management*, 2024, https://doi.org/10.55041/ijsrem37784

Wagner, B., Ferro, C. and Stein-Kaempfe, J., *Implementation guide: Good practices for ensuring data protection and privacy in social protection systems*, Commissioned by GIZ's Sector Initiative Social Protection for the SPIAC-B Workstream on Data Protection, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn, 2021

Wirtz, B., Weyerer, J., and Kehl, I., 'Governance of artificial intelligence: A risk and guideline-based integrative framework', *Government Information Quarterly*, 39(1), 101685, 2022, https://doi.org/10.1016/j.giq.2022.101685

World Bank, *Playbook on digital social protection delivery systems: Towards dynamic inclusion and interoperability*, World Bank, Washington, DC, 2024

World Bank, *Global insights on social registries: Coverage and beyond*, World Bank Washington, DC, 2025

Zeng, Y., Klyman, K., Zhou, A., et al., AI risk categorization decoded (AIR 2024), *AGI – Artificial General Intelligence*, 1(1), 2024, https://doi.org/10.70777/agi.v1i1.10603

Zhang, X., Chan, F., Yan, C., and Bose, I., 'Towards risk-aware artificial intelligence and machine learning systems: An overview', *Decision Support Systems*, 159, 113800, 2022, https://doi.org/10.1016/j.dss.2022.113800

Zhang, J., and Zhang, Z, 'Ethics and governance of trustworthy medical artificial intelligence', *BMC Medical Informatics and Decision Making*, 23, 144, 2023, https://doi.org/10.1186/s12911-023-02103-9

Zhou, N., Zhang, Z., Nair, V. and Singhal, H., 'Bias, fairness and accountability with artificial intelligence and machine learning algorithms', *International Statistical Review*, 90(3), 468–480, 2022, https://doi.org/10.1111/insr.12492

Zimelewicz, E., Kalinowski, M., Méndez, D., Giray, G., Alves, A., Lavesson, N., Azevedo, K., Villamizar, H., Escovedo, T., Lopes, H., Biffl, S., Musil, J., Felderer, M., Wagner, S., Baldassarre, T. and Gorschek, T., ML-Enabled Systems Model Deployment and Monitoring: Status Quo and Problems, In: Software Quality as a Foundation for Security, SWQD 2024, *Lecture Notes in Business Information Processing*, Vol 505, Springer, Cham, 2024, https://doi.org/10.1007/978-3-031-56281-5_7

Zuiderwijk, A., Chen, Y., and Salem, F., 'Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda', *Government Information Quarterly*, 38(3), 101577, 2021, https://doi.org/10.1016/j.giq.2021.101577

ADDITIONAL READING

Aiken, E., and Ohlenburg, T., *Novel digital data sources for social protection: Opportunities and challenges*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn, 2023

Alavi, M., Leidner, D., and Mousavi, R., 'Knowledge management perspective of generative artificial intelligence', *Journal of the Association for Information Systems*, 25, Article 15, 2024, https://doi.org/10.17705/1jais.00859

Albahri, A., Duhaim, A., Fadhel, M., et al., 'A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion', *Information Fusion*, 91, 364–395, 2023, https://doi.org/10.1016/j.inffus.2023.03.008

Ateeq, K., Oswal, N., Jawabri, A., et al., 'The transformative impact of artificial intelligence (AI) on organisational behaviour (OB): A study of employee engagement, performance, and ethical implications', *Journal of Posthumanism*, 5(4), 2025, https://doi.org/10.63332/joph.v5i4.1221

Camilleri, M., 'Artificial intelligence governance: Ethical considerations and implications for social responsibility', *Expert Systems*, 41(2), e13406, 2023, https://doi.org/10.1111/exsy.13406

GIZ, *AI in Social protection: Exploring opportunities and mitigating risks*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn, 2020

Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L., 'A survey on dataset quality in machine learning', *Information and Software Technology*, 162, 107268, 2023, https://doi.org/10.1016/j.infsof.2023.107268

Humphreys, D., Koay, A., Desmond, D., and Mealy, E., 'AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business', *AI and Ethics*, 4, 791–804, 2024, https://doi.org/10.1007/s43681-024-00443-4

Juhász, P., 'Operational risks caused by AI use at work and their management in professional literature', *Economy & Finance*, 12(1), 2025, https://doi.org/10.33908/ef.2025.1.5

Kelly, C., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D., 'Key challenges for delivering clinical impact with artificial intelligence', *BMC Medicine*, 17, 195, 2019, https://doi.org/10.1186/s12916-019-1426-2

Kjærum, A., and Madsen, B.S., 'Pushing the boundaries of anticipatory action using machine learning', *Data & Policy*, 7, e8, 2025

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., and Gebru, T., Model cards for model reporting, *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT '19), 220–229, 2019, https://doi.org/10.1145/3287560.3287596

Mäntymäki, M., Minkkinen, M., Birkstedt, T., and Viljanen, M., 'Defining organizational AI governance', *AI and Ethics*, 2, 603–609, 2022a, https://doi.org/10.1007/s43681-022-00143-x

Mäntymäki, M., Minkkinen, M., Birkstedt, T. and Viljanen, M., 'Putting AI ethics into practice: The hourglass model of organizational AI governance', *arXiv preprint*, arXiv:2206.00335, 2022b, https://doi.org/10.48550/arXiv.2206.00335

Martin, N., and White, J. 'Water resources: AI–ML data uncertainty risk and mitigation using data assimilation', *Water*, 16(19), 2758, 2024, https://doi.org/10.3390/w16192758

Priyanshu, A., Maurya, Y., and Hong, Z., 'AI governance and accountability: An analysis of Anthropic's Claude', *arXiv preprint*, arXiv:2407.01557, 2024, https://doi.org/10.48550/arXiv.2407.01557

Rana, N., Chatterjee, S., Dwivedi, Y., and Akter, S., 'Understanding dark side of artificial intelligence (AI) integrated business analytics', *European Journal of Information Systems*, 31, 364–387, 2021, https://doi.org/10.1080/0960085X.2021.1955628

Schneider, J., Abraham, R., Meske, C., and vom Brocke, J., 'Artificial intelligence governance for businesses', *Information Systems Management*, 40(3), 229–249, 2020, https://doi.org/10.1080/10580530.2022.2085825

Shevlane, T., Farquhar, S., Garfinkel, B., et al., 'Model evaluation for extreme risks', *arXiv preprint*, arXiv:2305.15324, 2023, https://doi.org/10.48550/arXiv.2305.15324

Tejani, A., Ng, Y., Xi, Y., and Rayan, J., 'Understanding and mitigating bias in imaging artificial intelligence', *Radiographics*, 44(5), e230067, 2024, https://doi.org/10.1148/rg.230067

Varma, Y., 'The evolution of AI model governance: Navigating risks with advanced monitoring', *International Journal of Science and Engineering*, 9(2), 2023, https://doi.org/10.53555/ephijse.v9i2.299

Wagner, S., Baldassarre, M., and Gorschek, T., 'ML-enabled systems model deployment and monitoring: Status quo and problems', *ArXiv*, abs/2402.05333, 2024, https://doi.org/10.48550/arxiv.2402.05333

ANNEX 1. METHODOLOGY

The methodology used to develop this framework was designed to ensure that the framework is grounded in evidence and can be applied practically by its users, particularly in low-capacity environments. All analytical decisions, therefore, aimed to translate high-level principles into tools that institutions can apply when assessing real AI systems.

Research approach and evidence gathering

The research combined a structured evidence review with targeted conceptual analysis. Work was carried out between August and November 2025 with a common guiding question: *What risks arise when AI is applied to social protection (and especially assistance) functions, and what safeguards are required to manage them proportionately?* The use cases examined are those discussed in the publication titled *AI Adoption Social Protection: A Global Evidence Review (2020–2025)* (DCI AI Hub, 2026b) and cover social protection implementation along the delivery chain primarily, but also other use cases at the policy level (e.g. informing financing allocations or policy decisions) and at the programme design level (e.g. informing decisions on transfer values or eligibility criteria to be used).

A structured synthesis process was used to validate risk patterns across multiple evidence sources. These sources included the DCI AI Hub's Global Evidence Review (DCI AI Hub, 2026b), multilateral guidance, technical papers, legal judgments, and civil-society analyses.[8] The methodology did not create new categories of risk; instead, it confirmed whether recurring issues in documented AI deployments corresponded to the rights-based principles outlined in Annex 2 of this report, and the governance requirements described in the Data Sovereignty report (DCI AI Hub, 2026c).

Alignment with foundational frameworks and international standards

The methodology aligned the framework with recognised international standards while ensuring feasibility for institutions in low- and middle-income countries (LMICs). External reference points included the OECD AI Principles (OECD, 2019), UNESCO's ethics guidance on AI (UNESCO, 2021), OHCHR recommendations on rights-based digital governance (OHCHR, 2021), and the EU's Artificial Intelligence Act (European Union, 2024). These frameworks informed the principles of proportionality, transparency, accountability, and human agency that underpin the safeguards in this report, but were adapted to reflect the practical realities of constrained social protection environments.

The analytical tools draw on operational risk instruments familiar to public-sector teams. Checklists, rubrics, and risk registers are structures used in humanitarian data governance frameworks, such as those produced by the ICRC (ICRC, 2024) and UN humanitarian data standards (OCHA, 2025), as well as algorithmic assessment toolkits developed for public-sector contexts (Ada Lovelace Institute & DataKind UK, 2020). These instruments were selected because they are specifically designed for fast-moving contexts characterised by high data risks, ensuring that rigorous safeguards

8 Examples include the District Court of The Hague judgment on SyRI (2020) and Amnesty International's analysis of Denmark's welfare fraud-detection system (2024a), both referenced in *AI Adoption in Social Protection: A Global Evidence Review (2020–2025)* (2026b).

can be applied even under operational pressure. Consequently, the core data protection principles applied in this framework are aligned with the *Handbook on Data Protection in Humanitarian Action* (ICRC, 2024), ensuring usability by programme managers, grievance officers, legal advisers, IT teams, and oversight bodies, not only technical specialists.

Analytical lens: mapping risks across the AI lifecycle

All risks were analysed using a six-stage AI life-cycle model adapted for the social protection sector. This model synthesises structures used in established governance frameworks (OECD, 2019; NIST, 2023) and adapts them to reflect the realities of social protection. It recognises that risks accumulate over time rather than appearing at a single stage (see Section 2).

Mapping risks to lifecycle stages clarifies when safeguards must be applied and who is accountable for them. This structure directly informs the design of the four tools within the AI Hub Risk Toolkit (Section 5), ensuring that mitigation actions correspond to identifiable responsibilities across programme, IT, legal, and oversight functions.

This study took a rights-based approach to looking at the use of AI in social protection. The rights-based approach is particularly suited to social protection, and social assistance specifically, as AI influences decisions about eligibility, benefit amounts, and sanctions, and, therefore, must be evaluated not only for technical accuracy, but also for compatibility with administrative justice and rights protection.

This rights-based approach transforms AI governance from a technical exercise into a framework that protects people and upholds public trust. In this context, *trustworthy AI* means that rights come first and risk controls follow. AI systems must operate within clear legal limits, be transparent and explainable, and remain under accountable human oversight. Institutions, not algorithms, must ultimately answer for decisions and provide justification and redress where needed. Because social protection determines access to essential benefits, public authorities must ensure that every use of AI is lawful, necessary, proportionate, fair, and subject to meaningful human review (OHCHR, 2021; UNESCO, 2021; European Union, 2012). These obligations derive from fundamental rights including privacy, equality and non-discrimination, human dignity, due process, and the right to an effective remedy (see Annex 2 for key principles of rights-based foundation).

Applicability and limitations of framework

Although the framework is universal and can be applied across any social protection use case, across different pillars (social insurance, social assistance and labour market policies), it was primarily designed and tailored for application to non-contributory social assistance, where risks and rights impacts are severe, as errors or bias in eligibility, sanctions, verification, or payments can cause immediate harm and undermine due process. Other pillars of social protection exhibit distinct risk profiles, legal mandates, and data structures. Social insurance and labour market interventions operate under different accountability regimes, meaning additional attention and possible modifications will have to be made if applying the framework to these pillars.

ANNEX 2. RIGHTS-BASED FOUNDATION FOR AI GOVERNANCE IN SOCIAL PROTECTION

The key principles of a rights-based framework for AI governance in social protection – privacy and data security; equality, non-discrimination, fairness, and inclusion; autonomy, human dignity, and due process; and accountability, transparency, and redress – are derived from a range of global legal frameworks, as synthesised in Table A1.

TABLE A1. KEY PRINCIPLES DERIVING FROM RIGHTS-BASED FRAMEWORKS FOR THE USE OF AI

KEY PRINCIPLES	FRAMEWORK
Beneficence; non-maleficence; justice; autonomy; explicability	**ILO, *Governing AI in the World of Work: A Review of Global Ethics Guidelines* (2023)**
Benefiting people and the planet; human-centred values and fairness; transparency and explainability; robustness, security and safety; accountability	**OECD, *AI Principles and OECD – Advancing Accountability in AI* (2019)**
Proportionality and do-no-harm; safety and security; right to privacy and data protection; multistakeholder and adaptive governance; responsibility and accountability; transparency and explainability; human oversight and determination; sustainability; awareness and literacy; fairness and non-discrimination	**UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021); OCHA, *Data Responsibility Guidelines* (2025)**
Human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability	**European Commission, *Ethics Guidelines for Trustworthy AI* (2019)**
Transparency; justice and fairness; non-maleficence; responsibility; privacy; beneficence; freedom and autonomy; trust; sustainability; dignity; solidarity	**Jobin et al., *Artificial Intelligence: The Global Landscape of Ethics Guidelines* (2019)**
Privacy; security; data protection; fairness; efficiency; service quality; administrative integrity; member protection principles	**ISSA, *Guidelines on Information and Communication Technology* (2019)**

Privacy and data security

Privacy and data security risks arise when the collection, storage, or use of personal information exceeds what is lawful, necessary, or proportionate. They apply to non-AI use cases, but can be exacerbated by AI. These risks originate primarily in ***data-related risks and vulnerabilities***, but are reinforced by weaknesses in *governance and institutional oversight*, as well as ***market, sovereignty and industry structure*** (Section 2). If data are gathered without clear purpose limitation, minimisation, or a lawful basis, individuals lose control over information that directly affects their rights and dignity. Poor data quality, undocumented transformations, or uncontrolled linkage of sensitive datasets can create hidden exposures and make it impossible for people to understand or challenge how their data influence decisions.

Operational failures, such as weak access controls, insecure interfaces, or poor incident-response capability, can turn minor errors into large-scale breaches. Governance gaps exacerbate the problem when there is no clear accountability for data stewardship, no oversight of data-sharing agreements, or no mechanisms to detect and correct misuse. Finally, reliance on external vendors or foreign cloud providers can limit national control over sensitive information, undermining both sovereignty and the ability to ensure lawful processing.

Protecting privacy and data security, therefore, requires strong governance of the entire data lifecycle: strict purpose limitation and minimisation, secure storage and access controls, clear documentation of data flows, and the ability to audit and correct how personal data are used in AI systems. Ensuring these safeguards is essential not only to maintain legal compliance, but also to uphold dignity, trust, and the fundamental right to privacy in social protection.

Note: These safeguards are also extensively discussed in two key DCI-endorsed resources, the 2022 *Implementation Guide*, which focuses on privacy and data protection best practices for practitioners (Wagner, Ferro & Stein-Kaempfe, 2021); and the 2025 *Data Governance Framework*, which provides a broader architecture for managing social protection data lifecycles (DCI, 2025).

Equality, non-discrimination, fairness, and inclusion

Equality, non-discrimination, fairness, and inclusion are primarily threatened by biases in AI data and biases in models, which together determine how AI systems classify, prioritise, or exclude individuals. Data-related risks, such as representation gaps, historical inequalities, or weak data quality in remote regions, form the root causes of discriminatory outcomes. Model-related risks then amplify these disparities through misaligned objectives, opaque logic, or unequal error rates across groups.

These risks are further compounded by others (discussed in Sections 2). Operational and system integration risks (such as misconfigured thresholds or weak subgroup validation) and governance and institutional oversight failures (such as missing audits, inadequate oversight, or inaccessible grievance channels) further entrench these patterns, allowing small technical disparities to become systemic exclusion. Market, sovereignty and industry structure risks indirectly affect fairness when external providers impose tools that cannot be audited or adapted to local equity considerations.

Together, these structural vulnerabilities determine whether AI systems uphold or undermine equality and fairness in social protection.

Autonomy, human dignity, and due process

Autonomy, human dignity, and due process are primarily threatened by model risks, operational and system integration failures, as well as weak governance and institutional oversight. Opaque or misaligned models undermine the ability of individuals to understand and challenge decisions, turning them into passive objects of automated classification. Operational weaknesses, such as rigid automation, missing human oversight, or broken appeal channels, further erode procedural justice by removing the human judgment needed to contextualise decisions. Governance failures are particularly damaging: without clear accountability, documentation, or grievance mechanisms, it becomes impossible to guarantee notice, explanation, contestability, or proportionality.

Other AI-related risks also play a role. Data risks also contribute when incorrect or undocumented data form the basis of decisions that beneficiaries cannot meaningfully dispute. Finally, market, sovereignty and industry structure risks, such as vendor lock-in or restricted audit access, can prevent institutions from providing explanations or correcting errors, thereby undermining autonomy and the right to an effective remedy.

Together, these risk sources determine whether AI systems respect or erode individuals' agency, dignity, and procedural rights in social protection.

Accountability, transparency, and redress

Accountability, transparency, and redress ensure that AI systems in social protection remain under the control of public institutions, not automated processes or private vendors. These principles address the core governance and institutional oversight risks that arise when responsibilities are unclear, documentation is weak, or oversight mechanisms are absent. Without clear accountability, no authority can justify or correct an AI-assisted decision; without transparency, beneficiaries cannot understand how decisions affecting their rights were made; without effective redress, errors become systemic injustices rather than isolated failures.

This principle reinforces control across the AI lifecycle. It requires institutions to document decision pathways, maintain audit trails, ensure explainability, and assign a named human decision owner at each critical step. It also demands accessible grievance mechanisms and the legal authority to remedy harm. By strengthening institutional capacity to trace, explain, and correct AI decisions, this principle mitigates governance, but also operational and system integration, and market, sovereignty and industry structure risks.

IMPRINT

Published by
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn

Friedrich-Ebert-Allee 32 + 36
53113 Bonn, Germany
T +49 228 44 60-0
F +49 228 44 60-17 66
info@giz.de

Dag-Hammarskjöld-Weg 1–5
65760 Eschborn, Germany
T +49 6196 79-0
F +49 6196 79-11 15

Required citation
DCI AI Hub for Social Protection, *AI Hub Risk Assessment Framework for Social Protection: A Practical Guide to Responsible AI Deployment*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Germany, 2026.

Authors
Thomas Byrnes
Simone Dobre
Laura van der Erve

Responsible for the content
Bernd Schramm, GIZ

Editing
Susan Sellars

Layout
EYES-OPEN, Berlin

Bonn, August 2026

AI Hub for Social Protection

The AI Hub is a global partnership, with close cooperation among nine core implementing partners: German Social Accident Insurance Institution for the Construction Sector (BG BAU), Government of Canada – Employment and Social Development Canada (ESDC), Dataprev (the Social Security Information and Technology Enterprise of the Brazilian government), Federation of German Pension Insurance Institutions (DRV Bund), Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), International Labour Organization (ILO), International Social Security Association (ISSA), Organisation for Economic Co-operation and Development (OECD) and the World Bank.

The AI Hub is implemented by:

BG BAU Berufsgenossenschaft der Bauwirtschaft

Government of Canada / Gouvernement du Canada

DATAPREV

Deutsche Rentenversicherung

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

issa

OECD

THE WORLD BANK

International Labour Organization