Tool 1 – Intrinsic Risk Worksheet

This Toolkit accompanies the AI Hub Risk Framework for Social Protection: Potentials, Risks, and Safeguards (Byrnes, Dobre and van der Erve, 2026). It consists of four tools designed to be used in sequence: Tool 1 (Intrinsic Risk Assessment), Tool 2 (Context Risk Multiplier), Tool 3 (Risk Register), and Tool 4 (Go/Pilot/No-Go Checklist). Each tool can be consulted independently, but the full assessment requires completing all four. The companion report provides the conceptual foundations and detailed guidance for applying this Toolkit.

This tool helps governments assess the inherent risks of an AI component before any safeguards are applied. It provides a structured, qualitative approach to identify what could go wrong, the severity and spread of potential harm, and which rights may be affected. The outcome is an **Intrinsic Risk Tier** (Low / Medium / High).

While developed for social assistance use cases, this tool is adaptable for broader use across social protection pillars and other public-sector domains.

Sources and Methodology

This tool draws on established risk assessment frameworks from multiple domains, adapted for AI applications in social assistance:

- **Humanitarian data governance:** The three-dimensional risk structure (likelihood, severity, scale) adapts the Human Rights Impact Assessment (HRIA) methodology for AI set out in the *Handbook on Data Protection in Humanitarian Action* (ICRC, 2024, Chapter 17.6), which assesses risk through likelihood (probability × exposure) and severity (gravity × effort to overcome). The Handbook's guidance on assessing data quality, bias risks, and the importance of meaningful human oversight directly informs this tool's diagnostic questions.
- **Humanitarian data responsibility:** The OCHA Data Responsibility Guidelines (OCHA, 2025) inform the tool's approach to evidence-based assessment, documentation requirements, and the precautionary principle when evidence is incomplete.
- Social protection standards: The tool reflects principles from the DCI-endorsed Implementation Guide on Data Protection and Privacy in Social Protection Systems (2022) and Data Governance Framework for Digital Social Protection Systems (2025).

A note on rating scales: This tool rates intrinsic risk as Low, Medium, or High, where High means greatest concern. Tool 2 uses a different scale (Strong / Moderate / Weak) because it measures institutional readiness rather than risk.

The two scales are combined in the decision matrix described in the companion report (Byrnes, Dobre and van der Erve, 2026, Section 5.4).

Relationship to the risk framework: This tool examines the first three of the five structural sources of risk identified in the companion report (Byrnes, Dobre and van der Erve, 2026), namely Data Risks, Model Risks, and Operational and Integration Risks, through the lens of likelihood, severity, and scale. The remaining two sources, Governance and Institutional Oversight Risks and Market and Sovereignty Risks, are assessed in Tool 2.

When to apply this Tool

Applying this tool early—during scoping, pilots, scale-up, or major changes—helps ensure AI supports rather than replaces human judgment in critical decisions like eligibility or appeals. Reviews should be conducted by a cross-functional team, typically including a **data-science lead**, a **programme or MIS manager**, a **legal or data-protection officer**, and an **accountability or grievance specialist**.

Tool 1 must be reapplied after major data changes, distribution shifts, or model retraining to account for model drift and evolving risk.

System Information

Field	Details
AI Component / Module	
Brief description	
Version number	
Date assessed	
Date for next review	

Intrinsic Risk Tier

Fill in the intrinsic risk level, as determined by the detailed assessment performed below: □ Low □ Medium □ High

Sign-off

Role	Name	Signature	Date
AI Component / Module		_____	_____
Brief description		_____	_____
Version number		_____	_____
Date assessed		_____	_____

How to use this tool

Step 1 – Complete Pre-Assessment Screening

Complete the screening questions on proportionality, do-no-harm, and accountability. If significant concerns arise, document them and ensure they are addressed before proceeding. If fundamental red lines are crossed (e.g., system enables social scoring or mass surveillance with no safeguards), do not proceed with detailed assessment until resolved.

Step 2 – Review the evidence and determine risk level for each dimension

Start by examining all available documentation for the three dimensions. Look for completeness, clarity, and reliability—flag any gaps, outdated sources, unverifiable claims, or vendor-only evidence. If evidence is weak or uncertain, apply the precautionary principle and lean toward a higher rating.

To reach a decision:

- Compare evidence against the rating criteria (Low / Medium / High)
- Discuss uncertainties openly (consensus is good practice because it surfaces disagreements and helps uncover overlooked issues)
- Confirm that another team could replicate the assessment using the same evidence.

The goal is a shared, defensible rating for each dimension based strictly on what is documented.

Step 3 - Document the justification

Record a concise but complete narrative explaining the justification for any decision, including:

- What evidence was reviewed?
- What uncertainties remain?
- What real-world impacts were considered?
- Why this tier appropriately reflects risk to beneficiaries?

Clear justification ensures traceability, supports audits, and enables learning.

Step 4 - Determine overall risk level

Use the table in the last section to determine the overall risk level based on the individual risk levels for each of the three dimensions of intrinsic risk. Record the resulting risk level on the first page and ensure you record sign off from all relevant actors on the same sheet.

Step 5 – Record Residual Risk

This becomes the baseline for Tool 3 (AI Risk Register), where mitigation and monitoring actions are planned and each risk is tagged with its decision implication (Blocking, Pilot Condition, or Monitor).

Step 6 - Move to next tool

Continue the risk assessment pathway, by moving to Tool 2, which will assess the institutional readiness for AI, and the accompanying contextual risk to come to an overall risk rating.

Step 7 – Reapply When Conditions Change

Repeat the assessment whenever uncertainty increases e.g., new data, model retraining, legislative changes, programme rule shifts, or moving from test to live pilot. Reapplication ensures safeguards stay aligned with evolving risks and that decisions remain transparent and contestable.

Intrinsic Risk Worksheet

Pre-Assessment Screening

Complete this section before proceeding to the detailed risk assessment. If any answer raises significant concerns, document them and ensure they are addressed before continuing.

Proportionality

- Was a non-AI approach considered to achieve the same goal? If AI was chosen, what is the rationale?
- Were simpler or more transparent methods considered (e.g., rule-based systems, simpler models)? Why was this specific approach selected?

Do No Harm

- Could this system be repurposed or adapted for social scoring or mass surveillance? If so, what safeguards prevent this?
- Could the system's impacts be irreversible or affect fundamental rights (e.g., dignity, due process, non-discrimination)?

Accountability

- Is there a named owner accountable for this AI system's performance and impacts?
- Have affected communities or their representatives been consulted in the design or review process?

Note: Governance arrangements are assessed in detail in Tool 2 (Context Risk Multiplier). These questions confirm minimum accountability foundations before proceeding.

Dimension 1 - Likelihood

The probability that the model will generate incorrect, misleading, or unjust outputs. This depends on training quality, validation on local data, bias audits, and stability under changing conditions. Systems with weak testing or opaque error modes have higher likelihood of harm.

Key Question: How often could the system produce harmful or biased outcomes?

Diagnostic Questions:

- Has the model been tested on data that reflects the local population it will be used on (e.g., not just trained on data from other countries or contexts)?
- Do test results show the model works equally well across different groups (e.g., by gender, region, age, disability status, income level)?
- Has someone independent checked whether the model produces unfair or unequal results for certain groups (bias audit)?
- Can technical staff explain why the model makes the decisions it does, and identify when it might go wrong (interpretability)?
- For generative systems (e.g., chatbots): Does the system sometimes produce false or invented information (hallucination)? How often?
- For predictive models (e.g., eligibility scoring, fraud detection): How often does the model wrongly flag someone as ineligible or fraudulent (false positive), or miss someone who should have been flagged (false negative)?
- Are feedback loops in place to capture and learn from errors in real-world operation (e.g., grievance data, manual overrides, outcome tracking)?

Risk Rating Table

Risk level	What it means	Indicators	Tick
Low	Errors are rare and well-characterised	Model thoroughly tested on local data; performs consistently across different groups; independent bias audit completed; error patterns documented and understood. *Example: A flood-forecasting model validated over multiple seasons with local data.*	☐

Risk level	What it means	Indicators	Tick
Medium	Errors occur but are manageable with oversight	Some gaps in testing; moderate concerns about bias or unequal performance across groups; limited independent verification; error patterns partially understood.	☐
High	Frequent, unpredictable, or poorly understood errors	Little or no testing on local data; unclear whether model works fairly across groups; relying mainly on vendor claims that cannot be verified; model behaviour difficult to explain. *Example: A new eligibility classifier trained on thin or outdated data from a different context.*	☐

Evidence & Assumptions

Evidence reviewed	
Evidence gaps/ unverified claims	
Assumptions	

Dimension 2 - Severity

The consequences of an error for beneficiaries or programme operations. Severity rises when harm affects eligibility, sanctions, or appeal rights, or when errors are irreversible and hard to contest. The sensitivity of the data processed 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.

High-Risk Use Case Indicators

The following use cases should be presumed HIGH severity unless strong mitigations are documented:

Phase in Delivery Chain	High-Risk Use Cases
Assess	AI determines who is identified as potentially eligible or vulnerable (not just supports outreach)
Enrol	AI directly influences eligibility decisions, identity verification rejection, or programme entry/exit
Provide	AI triggers payment suspension, fraud flags, sanctions, debt recovery, or benefit reductions
Manage	AI influences case closure, recertification failure, or sanctions without human review

Cross-cutting red flags:

- Decisions are automated without meaningful human review
- Beneficiaries are not informed that AI contributed to the decision
- No accessible mechanism to contest or appeal AI-influenced outcomes
- Errors are difficult to reverse or take significant time to correct

If your use case falls into any of these categories, start with a HIGH severity rating and document specific mitigations that would justify a lower rating.

Key Question: If the system fails, how serious is the harm to rights or entitlements?

Diagnostic Questions:

- If the model makes a mistake, what actually happens to people? (e.g., delayed payment, denied benefit, wrongly flagged for investigation)
- Could a household lose access to cash transfers, food assistance, or essential services because of an error?
- Could an error result in someone being wrongly sanctioned, investigated, or required to repay benefits?
- Could an error undermine someone's right to appeal or to have their case reviewed by a human?

- Is the harm reversible? How quickly can errors be identified and corrected? (e.g., days, weeks, months)
- Does the system directly influence who is included or excluded from a programme, or who faces sanctions?
- Can beneficiaries understand why a decision was made and realistically contest it if they believe it is wrong?

Risk Rating Table

Risk level	What it means	Indicators	Tick
Low	Failure mostly causes inconvenience, not harm	Mistakes lead to minor delays or incorrect information but do not affect access to benefits or rights. *Example: Chatbot gives wrong answer about office hours.*	☐
Medium	Failure affects service quality but not access	Mistakes cause delays, extra steps, or frustration but can be corrected before benefits are affected. *Example: Case routed to wrong caseworker, causing delay*	☐
High	Failure risks exclusion, sanctions, or rights violations	Mistakes could lead to wrongful denial of benefits, payment blocks, fraud flags, sanctions, or debt notices; harm is hard to reverse or contest. *Example: Automated eligibility decision wrongly excludes a household.*	☐

Evidence & Assumptions

Evidence reviewed	
Evidence gaps/ unverified claims	
Assumptions	

Dimension 3 - Scale

The breadth of impact—how widely harm could spread if the system fails. Scale increases when systems cover large populations or connect to other critical components like ID, payments, or social registries. Errors in a small pilot affect few people, while mistakes in a national eligibility model can cascade across programmes.

Key Question: How many people or programmes could be affected?

Diagnostic Questions:

- How many beneficiaries or households fall under the model's influence? (e.g., hundreds, thousands, millions)
- Is this a national system affecting the whole country, a regional programme, or a small-scale pilot?
- Does the system connect to other critical systems (e.g., national ID, payment infrastructure, social registry) where errors could cascade?
- Could errors in this system trigger consequences in other programmes (e.g., exclusion from one programme affecting eligibility for others)?
- Does the system operate in real time (decisions happen immediately) or batch mode (decisions processed periodically, allowing time for review)?

Risk level	What it means	Indicators	Tick
Low	Small-scale, contained	Affects fewer than 5,000 people; standalone system not connected to other programmes; errors unlikely to cascade. *Example: Small pilot in one district.*	☐
Medium	Moderate coverage	Affects tens of thousands of people; regional scope; some connections to other systems but with safeguards. *Example: Regional programme covering several districts.*	☐
High	National or cross-programme reach	Affects hundreds of thousands or millions; connected to national ID, payments, or social registry; errors could cascade across multiple programmes. *Example: National eligibility model linked to social registry.*	☐

Guidance on scale and the pathway to deployment:

A high Scale rating does not mean the system can never be deployed. It means the system should not be deployed at full scale without first demonstrating safety at a smaller scale. The recommended pathway is to pilot the system in a contained environment (where Scale would be rated Low or Medium), gather evidence on performance, and reapply the full toolkit before each expansion. This staged approach ensures that high-scale systems receive proportionate scrutiny at every step, rather than proceeding directly fromdesign to national deployment.

Evidence & Assumptions

Evidence reviewed	
Evidence gaps/ unverified claims	
Assumptions	

Overall risk rating

The overall intrinsic risk rating is anchored in the highest of the three dimensions, as summarised in the table below. If any one dimension is scored as high, the intrinsic risk profile should be treated as high. This prevents serious risks—such as the possibility of wrongful exclusion, harmful sanctions, or widespread errors—from being diluted by lower scores in other areas. A system that fails infrequently but affects a large population, or one that operates reliably but has the potential to cause significant harm if errors do occur, should be considered high risk. **In cases where all three dimensions fall into the medium range, the rating should also be elevated to high.**

Intrinsic risk rating	Aggregation rule	
Low	• Scale, severity and likelihood **all** low	
Medium	• Any of scale, severity and likelihood Medium **and** • No dimension rated High **and** • **At least one dimension rated Low**	
High	• **Any** of scale, severity and likelihood High, **or** • **All** of the dimensions rated Medium	