WOMEN AT THE TABLE

POLICY PAPER

From Principles to Procurement

AI public purchasing that serves people



Caitlin Kraft-Buchman

21 October 2026

AI & EQUALITY

FRIEDRICH NAUMANN FOUNDATION For Freedom.
Human Rights Hub

Contents

Executive summary 01

01
The power of procurement
p. 02

02
Understanding the AI landscape
p. 05

03
AI readiness and the lifecycle
p. 07

04
Defining the public need
p. 09

05
Gender-responsive AI procurement
p. 11

06
Approaching the market and choosing a supplier
p. 13

07
Contracting, deployment and monitoring
p. 15

08
Making responsible procurement workable
p. 18

09
Policy recommendations
p. 20

Questions for decision-makers
p. 22

Sources and notes
p. 24

Executive summary

Public procurement gives governments the power to make better public services the outcome the AI market is paid to deliver. Exercising that power requires a clear public purpose, evidence that a system can fulfil it, and the capacity to act when it falls short.

This publication sets out a human rights-based approach to AI procurement, grounded in equality and non-discrimination, participation and accountability. These principles help authorities judge whether a system improves access, responds to people's circumstances and allows decisions to be understood and challenged.

The opening chapters explain procurement's influence over what suppliers build and why AI requires particular attention. Systems can reproduce historical inequalities, shape decisions through rankings or summaries, and change after purchase. Readiness therefore means having the expertise and institutional capacity to assess their effects throughout use.

The central chapters turn to defining the public need and making equality part of service quality. Relevant government departments, sector experts, frontline staff and affected people must help establish what should improve before a solution is chosen. Gender-responsive procurement provides an established practice on which to build: examining how systems interpret unpaid care, career breaks or irregular income can expose assumptions that apparently neutral performance measures miss. This gender lens makes the wider human rights approach concrete and helps reveal overlapping barriers affecting different groups.

The later chapters translate these objectives into supplier evaluation, enforceable contracts and continuing oversight. They explain why authorities need relevant testing, meaningful human review, routes to correction and the ability to suspend or replace a system. The policy recommendations identify the shared expertise, funding and institutional responsibility needed to make this workable.

The central test is whether the service becomes more effective for the people who depend on it. Faster processing matters when it contributes to that result. Procurement must define the improvement, require evidence of delivery and preserve government's ability to secure it.

01 The power of procurement

Use purchasing power to improve public services

Public procurement is an opportunity to ask what government could do better with new capabilities. AI may help services identify unmet needs, make information more accessible and support staff in responding to complex circumstances. The ambition should be to improve both people's access to useful services and government's ability to provide them. Purchasing decisions give that ambition practical force by defining what suppliers must deliver.

Achieving this takes deliberate work before the tender. The departments whose rules and records shape a service need to agree its purpose with procurement colleagues, sector experts, frontline staff and affected people. Time spent on that discussion can reveal possibilities that a request for faster processing would miss. It can also expose obstacles that technology alone cannot remove.

Public authorities buying AI are purchasing systems that can shape who receives support, which opportunities people see and what information officials use to make decisions. These systems can help government recognise unmet needs and provide better services. They can also repeat a mistaken assumption across thousands of cases before its consequences become visible. A supplier update can change their behaviour after the purchase. Procurement is therefore a point of public responsibility: it determines what government asks these systems to do, what evidence it demands and whether it retains the ability to correct their effects.

What a purchasing decision can change

Consider a public employment service buying an AI system to help people find work. The system treats a long, uninterrupted record of paid employment as a sign of suitability. A woman returning after years of unpaid care receives recommendations for work below her qualifications. Her experience has been read as a deficit, and that judgement now shapes the opportunities she sees.

This employment example shows how a purchasing decision can turn an unexamined assumption about people into a repeated feature of public provision. It also shows where buyers can intervene. The authority can require evidence that the system recognises relevant capabilities after a career break. It can test the quality of the opportunities offered. It can reject a product that reproduces the barrier it wants its service to remove.

Public procurement helps determine what suppliers develop and whose needs they address. The United Nations Economic Commission for Europe (UNECE) Transformative Innovation Charter makes this role explicit. It identifies public purchasing as a major source of demand for innovation and calls for that demand to support wider societal change.[1] For AI, the implication is direct. Governments can create a market for systems that improve access to public services and advance equality. They do this by making those outcomes part of what they buy.

Why AI requires closer attention

AI gives these choices unusual reach. A shared model can apply the same assumption across thousands of cases and several public bodies. In a benefits service, a summary that leaves out evidence of irregular earnings could lead a caseworker to misunderstand an applicant's circumstances. If the same summarisation practice is repeated across the service, a small omission becomes a recurring barrier. The consequences may accumulate much faster than complaints or periodic reviews reveal them.

The speed of change also matters. An authority may evaluate one version of a product and find that a supplier update has changed how it performs. A function purchased to assist staff may become the default basis for their decisions. Neither change necessarily brings the service back before a procurement team. Yet both can alter how public power is exercised. Oversight must follow changes in the system and in the way staff use it.

Experienced procurement officials already manage complex software, uncertain performance and supplier dependence. AI intensifies these difficulties and adds a particular question: are the patterns the system has learned a sound basis for the public task? Software can follow a fixed rule that the authority has specified. A machine-learning model may instead infer relationships from data, including relationships the authority never chose. Even when it performs its technical function correctly, the resulting ranking, prediction or summary may be unsuitable for the service. Buyers must assess what those outputs mean for the people who depend on them.

The concern extends across public services. An employment system may undervalue unpaid care; a benefits tool may misread irregular income; an information service may work poorly for people who use a minority language. Each can make an existing barrier harder to see and easier to reproduce. Equality and non-discrimination therefore belong within the assessment of service quality. Gender provides a particularly revealing lens because apparently neutral records of work, income and care can conceal unequal circumstances.

Build on procurement's existing public objectives

Procurement is already a means of putting public policy into practice. The innovation approach described by UNECE connects purchasing to societal objectives.[1] The World Bank's rated-criteria guidance provides for assessing quality, innovation and relevant social or environmental factors alongside cost.[12] Its GovTech approach links digital reform to accessible services, citizen engagement and coordination across government.[11] Applied to AI, these foundations support a clear expectation: the purchase should deliver an identifiable public improvement.

A human rights-based approach gives these public objectives a common foundation: equality and non-discrimination, participation and accountability should shape what authorities buy and how they assess its effects.[5,6] Gender-responsive procurement offers an established example of putting equality into purchasing practice. The European Institute for Gender Equality provides guidance across the procurement cycle, and the European Investment Bank has published a practical guidance note.[2,3] AI procurement can build on this practice by examining how a system's assumptions affect the people it is meant to serve.

Retain the ability to act

Purchasing power has limits. A local authority may have little influence over a widely used model or an incumbent software provider. Cooperation between buyers can strengthen their position. Individually, buyers should establish before award what evidence the supplier will provide: test results for the intended use, known limitations, records needed to investigate errors and information about significant changes. They also need enforceable arrangements for correction, suspension and transition to another service. These are the means by which an authority can act when experience challenges the supplier's claims.

This publication argues that public buyers can direct AI towards better services when they can establish its value, examine its effects and require correction. The argument builds on established procurement guidance and the lifecycle approach to AI risk and impact assessment discussed below.[2,3,6,7] Its practical test is straightforward: has the service improved for the people who need it, and can the authority act when it has not? Processing people faster is only one part of that judgement.

02 Understanding the AI landscape

Identify what the system influences

An AI system can influence a public decision without making the final decision itself. It may rank applications, select information for a caseworker or recommend which cases deserve attention. Each function affects what people see and what happens next. Buyers need to establish that influence before they can decide what evidence and oversight the purchase requires.

A predictive system estimates an outcome from patterns in data. In an employment service, it might estimate who is likely to find work quickly. That estimate does not establish who most needs support, or who would benefit most from training. If the authority allocates help according to the prediction, it has made a policy choice. The choice remains the authority's responsibility even when the score comes from a supplier.

This distinction matters because the data may record past inequality. If people with career breaks previously received less support, their recorded employment outcomes may be poorer. A model trained on those records could learn to expect poorer outcomes from similar applicants. Using that expectation to limit support would reinforce the pattern. The buyer needs to understand what the system predicts and whether that prediction serves the public purpose.

Distinguish predictions, content and actions

Generative AI raises a different problem. It produces content, including correspondence, summaries and answers to questions. A large language model can write a fluent explanation that contains an invented fact or omits a decisive qualification. Linking it to approved documents can improve the basis for its answers, but does not guarantee that it selects the right passage or represents it accurately.[4] For a public service, confident wording must never substitute for reliable information.

Consider a document-management platform that gains a case-summary feature through an upgrade. The original system stored and retrieved records. The new feature decides which information to include in an account of a person's circumstances. If a caseworker relies on that account, an omission can affect the decision even though the software is described as an assistant. The authority needs to evaluate the new function before it becomes routine practice.

Some products can also act through connected software. They may send a message, update a record or initiate another process. Here the question extends beyond whether the output is accurate. Buyers must establish which actions the system can take, where approval is required and how an incorrect action can be reversed. Permission to draft a letter should not silently become permission to send it.

Follow the function and the supplier chain

An AI service may depend on several companies. The application supplier can use a model provided by another business, hosted through a separate service. The company signing the contract may control the interface while having limited control over changes to the model. Buyers need to identify these dependencies because they determine whether promises about investigation, access and correction can be fulfilled.

The acquisition route can obscure the change. AI may arrive through a pilot, a consultancy assignment or an existing software contract. It may be enabled by an administrator rather than purchased by a procurement team. An inventory limited to contracts labelled AI will miss some consequential uses. Oversight needs to follow the function being introduced and the public service it affects.

Not every automated process is AI, and not every AI function carries the same consequences. A tool that helps locate an internal document warrants a different assessment from one that shapes access to housing. The useful distinction is what the system can influence, how serious an error would be and whether people can obtain a correction. These are the grounds for deciding how much scrutiny is needed.

03 AI readiness and the lifecycle

Establish the capacity to take responsibility

An authority is ready to buy AI when it can judge whether the proposed use is worthwhile and take responsibility for what follows. Procurement officials need colleagues who understand the service, the technology and the public responsibilities involved. The service owner defines the intended improvement; technical specialists examine performance; legal, equality and data-protection expertise informs the safeguards. Finance and operational teams establish whether those safeguards can be sustained. A named official must have authority to act after deployment.

Examine the model and the information it uses

Readiness requires distinguishing a problem with the model from a problem with the information it receives. A model trained on employment patterns in another country may give unsuitable recommendations even when the local applicant's record is complete. It may have learned relationships between qualifications and jobs that do not hold in the local labour market. The authority needs evidence that the model is suitable for its setting.

A model that performs well in that setting may still produce a poor recommendation if the applicant's record omits a qualification or misrepresents a career break. Here the immediate problem is the information supplied during use. Training data shapes the patterns a model learns; information supplied during use shapes its response to a particular case. Buyers need to understand both, because improving one does not necessarily repair the other.

Administrative records contain what an institution has recorded, not everything it needs to know. A service may have detailed information about successful applicants and almost none about people who abandoned the form. Employment records may document paid jobs while leaving skills acquired through unpaid care unrecorded. These gaps matter when a system uses the records to judge need or capability. More data from the same process will not necessarily resolve them.

Readiness also requires the ability to interpret performance claims. A high average score can conceal poor results for a smaller group. An error rate that sounds modest can affect many people in a large service. A benchmark from another country may say little about local language use or administrative rules. The evidence must address the intended task and the consequences of getting it wrong.

Make assessment affect the purchase

These questions arise at different points in the AI lifecycle: defining the purpose, choosing data, developing and testing the system, putting it into service and assessing its effects.[5] Buyers of existing products enter after some of these choices have been made. They must establish whether those earlier choices are suitable for their service and which can still be changed. A supplier may be able to correct a data connection but unable to alter the underlying model. That distinction affects whether the proposed remedy can work.

The Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law provides an international foundation for considering risks and impacts throughout the lifecycle. It addresses equality, accountability, remedies and risk management.[6] The procurement implication is that assessment must continue into use: the effects may change with a new version, a new population or a different role for the system.

HUDERIA, the Council of Europe's non-binding methodology for AI risk and impact assessment, offers a way to examine the context, involve affected people and identify measures to address potential harm.[7] For procurement, those findings should change the purchase. If assessment finds that a ranking could disadvantage applicants with career breaks, the tender should require testing of those cases. If a serious problem cannot be resolved, the authority should change the proposed use or decide against it. An assessment filed away without affecting the decision provides little protection.

A human rights-based approach makes the questions concrete. Equality asks whether the service reaches people fairly and avoids reproducing unjustified disadvantage. Participation brings the knowledge of affected people into its design. Accountability requires an institution that can explain the decision and secure a correction. These questions connect technical performance to the purpose of public provision; a supplier's general statement of ethical intent cannot answer them.

Ensure oversight works in practice

Human review needs particular attention because the system may influence the reviewer before a decision is made. A caseworker cannot check a summary properly without access to the source record, time to read it and authority to disagree with the recommendation. If ordinary workloads allow only a cursory glance, describing the process as human review overstates the protection.

Before purchase, the authority should test how long meaningful checking takes and fund that work. If it cannot, it should reduce the tool's influence or narrow its use.

Not all AI systems learn continuously after deployment. Change can instead come from a new model version, a different data source or a decision to use an existing feature for another purpose. Readiness includes the capacity to recognise when the original evidence no longer covers the service. If that capacity is absent, the authority needs to build it, narrow the proposed use or defer the purchase.

04 Defining the public need

Define the improvement that matters

The public need determines what counts as success. In the employment example, recommending more vacancies is easy to count. Helping people reach work that reflects their capabilities is harder to establish, but much closer to the purpose of the service. A procurement framed around the first measure may reward a product that fails the second.

AI makes this distinction especially important because suppliers can optimise a system against a chosen measure. If speed is the main target, the system may become very good at producing quick answers. That does not establish whether those answers help people obtain the right support. Buyers need to examine what the proposed measure leaves out before allowing it to define performance.

A backlog illustrates the issue. Some delays arise from repetitive work that software could reduce. Others arise because eligibility rules are unclear or staff lack the authority to resolve exceptions. Automating an early step may simply move the queue. The authority needs to locate the cause of the delay before deciding what technology should do.

Suppose a benefits service assumes that applicants struggle to find information and proposes a chatbot. Conversations with applicants reveal a different problem: different offices ask for different documents, and the form does not accommodate irregular income. A chatbot could repeat those instructions more quickly while leaving the obstacle intact. The first improvement may be to agree to consistent requirements across the offices and change the form. AI can then be assessed for a defined supporting role.

Bring the right people into the decision

Defining the need therefore requires people beyond the purchasing team. Other parts of government may own the records, rules or services on which the proposed system depends. Sector experts can test whether a suggested measure reflects a worthwhile outcome. Frontline staff know where procedures fail in practice. Citizens and other service users know what it takes to navigate them. Bringing these perspectives together can reveal an opportunity that no single department or supplier would identify alone.

In the employment service, this could involve colleagues responsible for training, recognition of qualifications and access to childcare, alongside labour-market specialists and people seeking work. A recommendation is of limited use if a recognised qualification is missing from the record, the training is unavailable or the working hours are incompatible with available care. These are connected service problems. The authority should decide which it can address through the purchase and which require action elsewhere in government. Buying a more sophisticated matching tool cannot settle them by itself.

People who do not reach the end of a process are particularly important to this enquiry. Successful applicants cannot explain every barrier faced by those who gave up. Nor can service records reveal all the circumstances they never captured. Frontline staff and organisations trusted by affected communities can help identify where the official description of the service differs from people's experience.

People returning after unpaid care can explain which capabilities their records fail to show. Their participation should change what the authority asks suppliers to demonstrate: for example, a way for applicants to add relevant experience and challenge unsuitable recommendations. Sector specialists can help assess how that experience relates to job requirements. Government teams can establish how corrected information will reach the staff who need it. Together, they can define a service that gives caseworkers better information and applicants more useful support.

This involvement should begin while the purpose and specification are still open. Participants need a clear account of what they can influence and feedback on what changed. Accessible formats, interpretation or payment for time may be necessary to hear from people who would otherwise be absent. Consultation after the product has been chosen cannot perform the same role. The public authority remains responsible for the decision, including explaining why it did not adopt a proposal.

Compare alternatives and test the proposition

The authority should compare the proposed AI use with credible alternatives. Could better search meet the need? Would changes to the form remove the difficulty? Could staff use a limited tool while retaining the judgement that matters? This comparison keeps the public objective in control of the purchase. It also avoids paying for additional complexity where a simpler intervention would work.

A useful specification connects the intended improvement to observable evidence. In the employment example, the authority could require evaluation of whether applicants with career breaks receive recommendations appropriate to their skills. It would also need to establish how applicants can correct incomplete information. These requirements follow from the identified barrier; a generic promise to improve employability would not resolve it.

The UNECE charter supports small-scale experimentation and learning in innovation procurement.[1] Applied here, that means testing a defined proposition before extending the service. A pilot should help answer whether the proposed system improves the outcome and whether the authority can manage its effects. It should have an explicit point at which the authority can stop. Spending time on a pilot is not evidence that wider deployment is justified. World Bank analysis similarly recommends starting with the service problem, comparing simpler alternatives and using pilots to establish whether wider use is warranted.[13]

05 Gender-responsive AI procurement

Examine assumptions about people's lives

An AI system can treat women and men according to the same rule and still reinforce inequality. If the rule rewards uninterrupted employment, it disadvantages people whose working lives include unpaid care. Removing a field labelled sex does not remove that assumption. Gender-responsive procurement examines how the system reaches its conclusions and what those conclusions mean for people's opportunities.

In the employment example, the first question is whether the system recognises relevant capability after a career break. The next is what work it recommends. Equal numbers of recommendations for women and men would reveal little if women were consistently directed towards lower-paid jobs or work below their qualifications. The assessment needs to reflect the quality of the opportunity offered.

Women do not encounter the service under identical conditions. A migrant's qualifications may be poorly recorded. A disabled applicant may need an accessible route to correct her information. A woman with limited control over a household device may be unable to use a service privately. The authority should identify the circumstances relevant to its population and test whether the proposed system can serve people in those circumstances.

Women make up roughly half the population, yet records and assumptions can still fail to represent their lives adequately. Gender provides a powerful test of whether a supposedly general-purpose system serves the public it claims to serve. The approach also applies to groups excluded or misrepresented because of disability, ethnicity, age, language or other circumstances. The barriers differ and may overlap: each requires evidence relevant to the people affected.

Gender-responsive procurement is already an established area of institutional guidance. EIGE addresses it across the purchasing cycle, while the EIB guidance provides voluntary measures within applicable legal frameworks.[2,3] AI adds a need to examine how historical patterns become predictions and how those predictions influence treatment. Gender expertise must therefore inform the choice of evidence, not arrive only after a product has been selected.

Test who benefits and who is missed

Testing needs to expose the assumption at issue. A demonstration using only applicants with continuous, well-documented careers will not show how the system treats a return after care. Suppliers should demonstrate performance on relevant cases, explain where evidence is weak and show what can be corrected. Average accuracy alone cannot answer an equality concern that arises for a particular group.

Where lawful and appropriate, results broken down by relevant characteristics can reveal differences concealed by an overall figure. Small samples require caution: an apparently good result may rest on too few cases to establish much. Interviews and case review can provide information that numerical measures miss. An applicant can explain why a recommendation is unusable even when the system classifies it as relevant.

Collecting more personal information is not an automatic solution. The authority should establish which information is necessary to investigate the concern and how it will be protected. Controlled test cases may answer some questions; others require carefully governed evaluation in the service. If a serious concern remains untested, that uncertainty must affect the deployment decision.

Widen supplier participation and test the product

Public purchasing also determines who can compete to supply the service. EIGE warns that generic turnover, insurance and previous-contract requirements can exclude smaller or newer organisations, including women-owned businesses.[8] This issue predates AI. Its relevance here is that unnecessary entry barriers can narrow the available expertise while favouring established suppliers. Requirements should match the actual demands of delivery.

Supplier participation and product performance need separate scrutiny. Ownership by women does not establish that a system treats women fairly. A large supplier's diversity policy does not show that its employment recommendations are suitable. The authority needs evidence about the service it is purchasing. EIGE's guidance on EU award criteria makes the connection to the contract explicit: a company's overall gender balance is not itself a basis for additional marks.[9]

The objective is concrete. A woman returning after care should be assessed on relevant capabilities and have a usable way to correct an incomplete account of her experience. Procurement can make that an expectation of delivery. It can also create demand for better methods of recognising skills that existing services have overlooked.

06 Approaching the market and choosing a supplier

Ask the market to address the service problem

The supplier must demonstrate that its system can improve the service the authority intends to provide. General claims about a model's capabilities cannot establish that it can interpret local records correctly, work in the languages people use or recognise experience acquired outside a conventional career. Evaluation needs to bring those conditions into view before the authority commits.

Market engagement should begin with the problem. In the employment example, the authority wants to help applicants reach work appropriate to their capabilities, including after unpaid care. It should ask suppliers how they would address the gaps in existing records and what evidence would establish that their approach works. This leaves room for innovation while keeping the public purpose clear. It also allows buyers to discover that an apparently straightforward request may require changes to data, staffing or connected services.

The colleagues and service users involved in defining the need have a role here too. Labour-market specialists can assess whether recommendations are credible. Staff can identify information they would need to check. Applicants can explain why a vacancy that appears suitable would be inaccessible in practice. Their contribution should help define realistic test cases and assessment criteria. The authority should then apply those criteria consistently across offers and document the basis for selection. This gives service expertise and lived experience a clear role while preserving a transparent purchasing decision.

Require relevant evidence

Evidence must match the claim. If a supplier reports high accuracy, the authority needs to know which task was measured, whose cases were included and which version was tested. A model's ability to answer general questions says little about whether the proposed application can summarise a benefits file without losing a decisive fact. An independent assessment is useful only to the extent that it covers the system and use under consideration. The buyer should make unresolved gaps visible rather than allowing a headline score to stand for the whole service.

Testing should include cases that could defeat the purpose of the purchase. For the employment tool, these include career breaks, unfamiliar qualifications and incomplete records. For a public information assistant, they might include questions expressed in everyday language and situations in which the correct response is to refer someone to a member of staff. The authority needs to see how the system handles uncertainty and how easily people can recover from a wrong answer. A good result on straightforward cases is only part of the evidence.

A pilot should test the proposed improvement against the current service or another credible alternative. Staff can assess the quality of recommendations and record the time spent checking and correcting them. Service users can assess whether the information helps them take the next

step. Where outcomes take time to emerge, the authority should distinguish early evidence, such as more relevant recommendations, from a demonstrated improvement in employment. A decision to extend the AI system to more users, offices or functions should reflect what the trial has actually established.

Assess support and turn requirements into criteria

The supplier's ability to investigate and correct failures is also part of what is being purchased. An application provider may use a model operated by another company. If summaries begin to omit important facts, the application provider may need that company's cooperation to identify what changed. Buyers should establish who can examine the failure, identify the version involved and implement a remedy. A promise of support has little value if the contracting supplier cannot obtain the information or assistance required to deliver it.

Equality requirements should be expressed through the task. A matching service that systematically overlooks qualified applicants after career breaks is failing to recognise relevant capability. The requirement is therefore to demonstrate suitable recommendations for those applicants and a way to correct incomplete information. This gives equality a concrete place in evaluation and connects it to the purpose of the contract.

The tender can then distinguish what every acceptable offer must provide from improvements that deserve additional marks. A workable correction process might be an essential requirement. Among offers that meet it, buyers could assess the time and effort needed to make a correction and whether it reliably changes subsequent recommendations. The criteria and assessment method should be clear in advance. A tested capability should receive a different assessment from an unsubstantiated undertaking to develop one. EIGE's guidance similarly emphasises the connection between award criteria and the contract being purchased.[9] The World Bank's rated-criteria guidance also calls for project-specific quality factors, supporting evidence and clear weights and evaluation methods.[12]

Count the full cost of delivery

Low-cost AI presents a particular challenge because the price of access can conceal the cost of responsible use. A tool may produce a summary in seconds but require several minutes of verification. It may reduce data entry while creating new work to investigate errors, review updates or respond to challenges. Those costs belong in the business case, alongside integration, staff training and eventual transition. Any burden transferred to citizens who must correct the system's account of their circumstances also matters when judging whether the service has improved.

A free feature added to an existing platform can create these responsibilities without a new purchasing decision. Its activation should therefore bring the service owner, digital team and procurement colleagues together to assess the proposed use. Where a limited tool demonstrably saves work and can be supervised adequately, the saving should count. Where the saving depends on removing the staff needed to check consequential outputs, the proposal needs to change.

Selection should leave the authority with a clear account of what has been demonstrated, which limitations remain and what must happen before deployment. Material promises made during competition must carry into the agreement.[9] The next stage is to turn that evidence into a service the authority can supervise and the public can challenge when it goes wrong.

07 Contracting, deployment and monitoring

Make commitments enforceable

The contract must turn the reasons for choosing the system into obligations the authority can enforce. If a supplier wins because its tool recognises skills after a career break, that claim must survive beyond the demonstration. The agreement should state how it will be tested, what evidence the supplier will provide during use and what must happen if performance falls short. AI makes this especially important because an update can change the behaviour on which the purchasing decision relied.

The EU model contractual clauses for AI offer a drafting resource, with a full version for high-risk systems, a lighter version for other systems and accompanying commentary.[10] They need to be adapted to the service and used within the wider agreement. Their practical value depends on securing information and cooperation that the authority can actually use.

The agreement should describe what the system is allowed to do and the limits on its use. In the employment example, recommending vacancies does not by itself authorise the system to determine eligibility for benefits or training. If the authority later wants to use the same score for those purposes, it should assess the new use before proceeding. The convenience of reusing a tool must not decide a change in public policy.

Test the service before launch

Delivery does not by itself justify deployment. Before launch, the authority should test the whole service under realistic conditions and ordinary workloads. Can an applicant see and correct a missing qualification? Does the correction change the recommendations? Can a caseworker recognise an unsuitable result and offer another route? Testing these steps establishes whether the promised protection works when someone needs it. A successful technical installation alone cannot establish whether people can obtain an explanation, correct their information or secure reconsideration.

Make explanation and correction usable

People need enough information to understand and challenge how AI affected their treatment. The authority should tell an applicant when an automated recommendation influenced the service they received, identify the relevant information used and explain how staff relied on the output. For example, someone directed towards unsuitable work should be able to establish whether the system overlooked a qualification or treated a career break as a lack of experience. They should also be told how to request reconsideration.

The contract should therefore secure records that connect the relevant inputs, output and subsequent staff action, together with supplier assistance to investigate disputed results. An explanation should be checked against those records. Asking the same AI tool to generate a

plausible account of its answer does not establish why it produced that result. Where the authority cannot explain the system's influence well enough to support a meaningful challenge, it should limit that influence or reconsider the use.[6]

A challenge must reach someone who can change the outcome. Applicants should have an accessible route to a person with the authority to reconsider their case, including when they cannot use the digital channel. Correcting the applicant's record should also correct the information used for subsequent recommendations. If the application cannot support that process, the problem should be resolved before it becomes a routine part of the service.

Monitor outcomes and investigate failures

Once the system is in use, monitoring should return to the original public purpose. Is it helping people reach work appropriate to their capabilities? Are applicants with career breaks still receiving poorer recommendations? Are staff spending more time correcting profiles than the tool saves? Availability and response speed remain useful measures, but they cannot establish whether the service is working for people. The authority should agree in advance what evidence it will examine and who will act on the findings.

Citizens and staff should remain part of that assessment. Applicants can identify recommendations that appear relevant in the data but are unusable in practice. Frontline staff can notice a recurring omission before it appears in an aggregate report. Sector experts can help distinguish a system problem from a change in the labour market. The service owner should bring these findings together and decide whether to change the tool, the process or both. If the obstacle lies in qualification recognition or access to training, the relevant department needs to act; the software supplier cannot resolve every failure of the wider service.

Complaints alone will miss some failures. A person may never know which opportunities were withheld or that AI influenced the options they saw. The authority should therefore examine a suitable sample of cases and investigate patterns of poorer outcomes, rather than waiting for affected people to discover and prove the problem. The scale and frequency of review should reflect the seriousness of the consequences and the volume of use.

Investigation requires a record of what happened. For the cases under review, the authority needs to establish which system version was used, what relevant information it received, what it produced and how staff acted on it. The contract should specify which records the supplier retains or provides, response times and appropriate privacy protections. Recording every detail indefinitely is unnecessary; retaining too little to examine a consequential decision is equally inadequate.

Control changes and prepare to stop

Updates need particular attention. Suppose a new model version makes summaries shorter but begins to omit information about caring responsibilities. The feature may appear faster and cleaner while becoming less useful. The agreement should require notice of changes that could affect the intended use and provide an opportunity to test them before they reach the live service. Where the supplier cannot offer that control, the authority must decide whether the use is compatible with the uncertainty. Changes in the authority's own working practices need review too.

When a serious problem appears, a named service owner must be able to restrict or suspend the affected function. Staff need an escalation route, and people must still be able to obtain the service. For an employment tool, this might mean returning temporarily to caseworker-led recommendations. That alternative needs staffing and preparation; a contractual right to switch off the tool is of little practical use if no one can continue the work.

Correcting the software may leave earlier harm unresolved. If the system repeatedly overlooked applicants with career breaks, the authority may need to identify affected cases and offer a fresh assessment. The supplier's obligations should support that investigation. Repairing the product and remedying its effects on people are distinct responsibilities.

The authority should also be able to leave. Relevant records must remain accessible, information needed for transition should be available in usable formats and the agreement should provide for necessary supplier assistance. Renewal should then be a decision based on evidence: whether the service has improved, whether the costs remain justified and whether effective oversight is still possible. A relationship should not continue simply because dependence has made reconsideration difficult.

08 Making responsible procurement workable

Assign responsibility across government

Governments must organise and fund the capacity they expect public buyers to exercise. A procurement official cannot independently verify a model, assess its effects across a population, negotiate with several technology providers and sustain participation after deployment. Without institutional support, the quality of scrutiny will depend on the resources of the individual authority. AI's reach makes that an inadequate basis for public provision.

Responsibility should sit with the service, supported by a clear division of work. The service owner is accountable for the intended improvement and decisions about continued use. Procurement colleagues translate requirements into the competition and contract. Technical specialists examine performance and dependencies; legal, equality and data-protection specialists examine the proposed use and safeguards. Operational teams establish whether staff can carry out the necessary checks. This work must begin while the authority can still change its approach.

Other departments need to participate where their decisions affect the result. In the employment example, an applicant may need a qualification recognised by the responsible education body before a matching tool can use it accurately. Acting on a recommendation may also depend on a training place or childcare available through another service. The employment authority therefore needs colleagues who can resolve those practical barriers, whether or not their systems share data directly. The lead authority needs an agreed route for resolving problems across those boundaries. An applicant should not have to establish which institution owns an incorrect record before anyone takes responsibility for correcting it.

Share expertise and strengthen public demand

Smaller authorities will often need access to shared expertise. A central or regional team could help examine supplier evidence, commission specialist testing and adapt contractual provisions. Sector networks could develop realistic cases for evaluating employment or benefits tools. Local authorities would still judge suitability for their populations, but they would have access to capabilities they could not reasonably maintain for every occasional purchase. Shared technical work should reduce duplication while preserving responsibility for the particular service.

Cooperation can also strengthen demand for better products. Public bodies facing the same problem can describe the capabilities they need and signal that credible evidence will matter in future purchasing. Suppliers then have a clearer reason to invest in, for example, recognising qualifications across different formats or supporting usable correction processes. This applies the UNECE charter's emphasis on public demand for innovation to an identifiable service need.[1] Common requirements should leave room for different solutions and be informed by local experience.

Match scrutiny to consequences

Scrutiny should follow the consequences of the use. A tool drafting an internal agenda needs a different level of assessment from one influencing access to housing. The relevant factors include the seriousness and scale of a possible error, who may be affected, the influence of the output and the difficulty of obtaining correction. Contract value alone cannot capture these factors. A low-priced subscription used across a large caseload may require substantial attention.

Authorities also need a straightforward way to make new AI uses visible. A service owner should be identified when a department proposes a pilot, enables an added feature or extends a tool to another task. Digital and procurement teams can then decide what assessment is needed. The process must be simple enough to use in practice and capable of distinguishing a minor administrative aid from a change that affects people's treatment. Otherwise consequential functions can become routine before responsibility has been assigned.

Fund oversight and learn from use

Budgets should cover the whole period of use. Evaluation, staff review, participation, monitoring and correction are costs of delivering the service. They should be considered alongside expected savings when the proposal is approved. Training must give staff enough understanding to recognise limitations, challenge outputs and escalate concerns; it should go beyond learning the interface. If meaningful checking cannot be sustained under normal workloads, the authority should narrow the use or reconsider the purchase.

Participation needs continuing support as well. Affected people should be able to contribute when experience reveals a problem or when a substantial change is proposed. Existing advisory groups and trusted community organisations may help, but officials must also seek perspectives that those arrangements miss. This includes people who abandoned an application or cannot use its digital channel. Reporting what changed as a result gives participation a clear purpose and helps maintain trust.

Governments should preserve and share what they learn. Records of failed tests, unexpected checking costs and reasons for ending pilots can be as useful as accounts of successful deployment. Shared findings can help another authority recognise an unsupported supplier claim or avoid repeating an unsuitable use. Institutions can share lessons while protecting personal information and legitimate confidentiality. They can report, for example, that a tool repeatedly omitted relevant qualifications, describe how the problem was tested and explain the remedy without disclosing identifiable case records or protected commercial details.

Leadership determines whether officials can act on that learning. Targets based mainly on the number of AI tools adopted or the speed of award create pressure to overlook unresolved concerns. A stronger measure is whether the institution has improved the service and can demonstrate the result. Officials need support for a justified decision to restrict or stop a use, alongside recognition for successful innovation. That is how experimentation can strengthen public accountability.

09 Policy recommendations

AI procurement policy should give public authorities the means to shape the services they buy and remain answerable for their effects. The recommendations below address the conditions governments should establish across institutions. Their implementation should reflect the applicable legal framework and the consequences of the proposed use.

1. Make the public outcome the starting point. Require proposals to identify the service problem, the intended improvement and the evidence by which success will be judged. Compare AI with credible alternatives. For an employment service, assess whether people reach suitable opportunities, alongside measures of speed or volume. This prevents a convenient technical measure from quietly replacing the public objective.

2. Bring government, sector expertise and citizens into the definition of need. Establish participation while the purpose and specification can still change. Include the public bodies whose records, rules or services affect the result, specialists who understand the sector, frontline staff and affected people. Give their findings a visible role in requirements and testing, and explain the decisions taken. This can reveal both barriers to access and opportunities for a better service.

3. Assess the AI function wherever it enters government. Apply a proportionate review to consequential uses introduced through existing software, free trials, consultancy and updates as well as dedicated purchases. Maintain a usable record of the function, intended use and responsible service owner. Scrutiny should follow potential impact, including influence on human decisions, rather than the contract label or licence price.

4. Require evidence relevant to the people and setting. Ask suppliers to substantiate claims against realistic service conditions and cases that could expose failure. Examine differences concealed by average performance, including those associated with gender, disability, language and other relevant circumstances. Combine testing with sector knowledge and the experience of affected people. Where serious uncertainty remains, narrow or defer the use rather than treating missing evidence as reassurance.

5. Make equality part of performance. Apply established gender-responsive procurement practice to the design, evaluation and delivery of AI services. Examine whether the system reproduces barriers in records of work, care, income or access. Keep requirements for supplier participation proportionate, while judging the product on evidence about the service it will deliver. Inclusion among suppliers and fair treatment of service users both matter; neither establishes the other.

6. Secure explanation, correction and control throughout use. Carry material commitments from selection into enforceable terms. Obtain the records and supplier cooperation needed to investigate failures, manage significant changes and give affected people a meaningful account of their treatment. Provide an accessible route to human reconsideration. Establish arrangements for suspension, continuity, remedy and exit, supported by contractual obligations. Test whether these arrangements can work across the supplier chain before relying on them.

7. Fund shared public expertise and continuing oversight. Provide procurement and service teams with access to technical, sectoral, legal and equality expertise. Budget for participation, realistic evaluation, staff review and monitoring throughout use. Where authorities face common suppliers or needs, support joint testing and exchange of lessons. Public capacity should grow alongside the ambition to deploy AI. World Bank analysis likewise stresses continuing ownership, funding and skills to evaluate and maintain AI services beyond a pilot.[13]

8. Judge adoption and renewal by demonstrated public value. Require a named owner to review whether the service is meeting its purpose, including its effects on different groups and its full operating costs. Make it possible to restrict or end a use when evidence warrants that decision, and investigate whether earlier cases need correction. A pilot should earn its continuation through results; deployment should remain open to reconsideration.

Public procurement gives government a practical way to shape what AI is developed and used for. The employment example makes that responsibility tangible: an applicant's experience should be recognised, a caseworker should have reliable information and the authority should be able to correct a failure. Achieving this requires public institutions to define their needs together with the people they serve, demand evidence and remain responsible as systems change. The opportunity is to make better public services the outcome the market is paid to deliver.

Questions for decision-makers

Use these questions to guide discussion at the relevant stage of a purchase or review of an existing service. Each addresses a different decision; the accompanying sentence explains what answering it should establish. Record the evidence, any unresolved issue and who will act on it. Revisit answers when the system or its use changes.

1. The power of procurement

What public improvement should this purchase deliver? Define the change people should experience, so the authority can distinguish a better service from a faster process.

What should our purchasing requirements encourage suppliers to develop? Identify unmet needs the market should address, such as recognising capabilities overlooked in conventional employment records.

2. Understanding the AI landscape

Where will AI influence the service—and with what consequences? Trace how predictions, rankings, summaries or automated actions affect people, including when an official makes the final decision.

What parts of the system can our supplier actually control? Identify dependence on other model providers or platforms, so promises about testing, changes and correction can be assessed realistically.

3. AI readiness and the lifecycle

Do we have the knowledge and capacity to judge this system? Establish which technical, sectoral and human rights expertise is needed before the authority can make an informed purchasing decision.

What do we need to know about the data behind its outputs? Examine both the records used to develop the system and the information it receives during use, because gaps in either can distort its treatment of people.

4. Defining the public need

What is causing the problem—and where could AI help? Separate tasks technology could improve from barriers rooted in policy, staffing or service design, then compare credible alternatives.

Who must help define the solution before we approach the market? Bring affected people, frontline staff, sector experts and relevant government departments into the requirements process while the proposed approach can still change.

5. Gender-responsive AI procurement

Could the system treat unequal circumstances as evidence of lesser need or ability? Examine how it interprets factors such as unpaid care, career breaks or irregular earnings, so existing disadvantage does not become a reason to withhold opportunities.

How will we establish whether the service works fairly for different groups? Define the results that must be examined separately, so average performance does not conceal poorer treatment of women or groups facing overlapping disadvantages.

6. Approaching the market and choosing a supplier

What must a supplier demonstrate before we select its system? Require evidence from realistic cases, including difficult ones, and establish which failures would rule out the proposed use.

Are our tender requirements unnecessarily narrowing the field? Check whether financial thresholds, experience requirements and documentation exclude capable smaller or women-led suppliers without improving assurance.

7. Contracting, deployment and monitoring

What must the contract let us inspect, challenge and change? Secure access to relevant records, cooperation in investigations and control over significant updates, so the authority can enforce the commitments made during selection.

What happens to people when the system gets something wrong? Establish how errors will be detected, decisions reconsidered and harm remedied, including for people who may not know that AI affected their treatment.

8. Making responsible procurement workable

Who will oversee the system, and can they do the work in practice? Assign responsibility and fund the expertise, staff time and continuing participation needed to review performance and act on problems.

Can we suspend or replace the system while keeping the service running? Prepare staffing, access to records and transition arrangements, so the authority can exercise its right to stop rather than remain dependent.

9. Policy recommendations

What should government organise once rather than leave every buyer to solve alone? Identify opportunities for shared expertise, joint testing and common requirements, strengthening public capacity and influence over suppliers.

What results should justify continued investment in AI? Make adoption and renewal depend on demonstrated service improvements, effects on equality and full operating costs, rather than the number of systems deployed.

Sources and notes

The employment and benefits examples are hypothetical illustrations used to explain procurement choices; they are not reported case studies.

[1] UNECE. Transformative Innovation Charter, adopted in 2025 (Annex II of ECE/CECI/2025/2). See paragraphs 5 and 29 on demand and procurement, and paragraphs 21 and 30 on experimentation and learning. Charter PDF.

[2] European Institute for Gender Equality. Gender-responsive Public Procurement: Step-by-step toolkit. Toolkit PDF.

[3] European Investment Bank. Gender-responsive public procurement EIB guidance note. 2026. The publication describes voluntary measures and an eleven-step roadmap within applicable legal frameworks. Source.

[4] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile. NIST AI 600-1, July 2024. See the discussion of confabulation. Source.

[5] Kallina, E., Kypraiou, S. and Kraft-Buchman, C. Integrating Human Rights Considerations Along the AI Lifecycle (December 2024). White paper PDF.

[6] Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225 (2024), especially Articles 8, 10 and 14–16. Treaty text. The Convention is cited as an international framework; the procurement recommendations here do not assume identical legal obligations in every jurisdiction. Applicable duties depend on the relevant treaty and domestic legal position.

[7] Council of Europe. HUDERIA: risk and impact assessment of AI systems. The Methodology and the Model: Context-Based Risk Analysis (COBRA) are standalone, non-binding resources. Methodology and supporting resources.

[8] EIGE. Selection criteria. Guidance on proportionate requirements and barriers affecting smaller or newer organisations. Source.

[9] EIGE. Award criteria. Guidance on the link to the subject matter of the contract, meaningful scoring and carrying commitments into the agreement. The legal discussion concerns the EU framework described by EIGE. Source.

[10] European Commission Public Buyers Community. Updated EU AI model contractual clauses, 5 March 2025, and accompanying commentary. These are drafting resources rather than a complete agreement. Source.

[11] World Bank. GovTech. Programme approach linking reform across government with citizen-centred services, accessibility and citizen engagement. Source. Accessed 21 October 2026.

[12] World Bank. Rated Criteria. Guidance on evaluating relevant non-price attributes, evidence, weighting and value for money in World Bank-financed procurement. Cited as an established procurement approach, not a rule applying to all public authorities. Source. Accessed 21 October 2026.

[13] Nayyar, G., Peixoto, T. C. and Ramos-Maqueda, M. From AI pilots to scale: Improving the delivery of public services. World Bank, 24 September 2026. Institutional analysis on problem definition, evaluation, procurement and sustained public capacity. Source.

This publication has been created and published with the support of and in cooperation with the Friedrich Naumann Foundation for Freedom Human Rights Hub.