AI Implementation Considerations

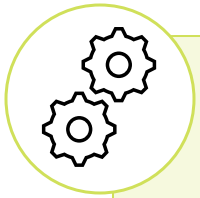
A decision framework for public sector decision-makers

June 2026

cenfri

Policymakers face two overarching questions when assessing AI adoption

How can AI be implemented?



What are the factors to consider when implementing AI?

What methodology of implementation can be considered?

*This requires **understanding** implementation **options***

Should AI be implemented?

What are the costs, benefits, risks and feasibility of different use cases?

Are they worth implementing?

*This requires **evaluating** implementation **options** against specific criteria.*

This presentation focus on the first question

cenfri

5 factors influencing how AI is implemented

These 5 factors guide a country's decision-making process when considering how to implement AI.

	Factor	Factor explanation
How AI is used	**AI integration**	**Describes how integrated the AI tool is with local systems.** This can range from little or no integration (e.g. using ChatGPT) to tools embedded into workflows or coordinating across systems.
How AI is used	**AI autonomy**	**Describes the level of agency of the AI and its authority to act.** Ranges from low autonomy (inform or propose for human review) to high autonomy (AI executes within guardrails.)
How AI is deployed	**Deployment model**	**Describes where the data is located and where processing happens.** Ranging from in the cloud, to in-country servers, or in-country on distributed devices.
How AI is deployed	**Local adaptation**	**Describes the degree of local customization of AI tools.** Ranges from no customization (just using ChatGPT as is, for example) to deep customization (training an existing model on local languages).
How AI is deployed	**Need for compute**	**Describes the level of computing power and infrastructure required**, from basic office IT to clusters of specialised chips and graphics cards for hosting large models.

*Each AI use-case can potentially be **implemented at different ends** of the spectrum for each of these factors. **But the option chosen has implications for the associated value, risks and costs.***

cenfri

The Cenfri AI Implementation Framework

How AI is used

AI integration

Standalone tools
Ad-hoc use of AI tools outside of core system (e.g an individual using ChatGPT).

Workflow embedded
AI is integrated into specific systems to support operations, e.g document review, call centre.

System-of-record embedded
AI is built into authoritative systems. (e.g tax benefits). It substantively supports decisions and outcomes are kept on record.

Cross-system orchestration
AI connects and coordinates multiple different systems to handle an entire process from start to finish.

AI autonomy

Inform
AI provides information (summarise, search, explain) but does not propose actions or outputs to be sent/committed.

Propose
AI suggests actions and/or drafts outputs for a human to review and edit; nothing is executed automatically.

Execute with approval
AI can take actions (send, submit, update, route) only after an explicit human confirmation step.

Execute with guardrails.
AI takes actions automatically within predefined limits (policies, thresholds, budgets) and escalates exceptions for human handling.

How AI is deployed

Deployment model

External API
Applications call an AI model over the internet. The provider handles all technical aspects and scaling.

Hybrid (API + Local data)
An external AI model is used, but local systems manage the data, adding context, applying safety rules, and maintaining logs.

In-country hosted
The AI runs on servers physically located within the country with greater control over data and latency.

Distributed Edge
Model runs on local devices or edge servers near the user (e.g., phones, clinic servers), reducing reliance on central services and connectivity.

Local adaptation

No localisation
The use of off-the-shelf AI services or APIs, like ChatGPT.

Local augmentation
Simple consumption of off-the-shelf tools, but with some local context included like guardrails.

Local tuning
The use of an existing model which has received some extra training on specific local information.

Extensive local training
Building of a new model through substantial additional training on specific local information.

Need for compute

CPU Native
Compute needs are met on local CPUs, no need for dedicated GPUs.

GPU Assisted
GPUs are beneficial for parts of the workload, but only a small pool are needed for some fine-tuning work.

GPU dependent
GPUs are required for core workloads, implying dedicated accelerator capacity.

GPU Cluster scale
The use case requires cluster-scale compute (e.g large training runs).

See Glossary of terms at the end of this slide deck for detailed explanations of each term.

Example AI use cases

Coding assistant for government teams	Automated grading of exam papers	Weather forecasting + multi-channel alerts	AI radiology triage in hospitals	Local language LLM (shared capability)
AI coding assistant to help public-sector developers to speed up and improve government software delivery (code, test, document). Can be individual tool or embedded into shared dev workflows (repos, code review, CI/CD).	AI-assisted marking to reduce exam backlogs and overtime costs by pre-scoring scripts & routing exceptions for human moderation. Focus on faster turnaround with auditability and fairness checks to relieve pressure for end-of year and transition exams.	Machine learning enhanced weather forecasting plus automated alerts via SMS/app and government channels. Deployment can range from using external forecasts with local dissemination to locally calibrated or locally hosted models.	AI radiology tool that analyses imaging (e.g. X-rays/CT scans) to detect fractures, TB, and pneumonia with decision support to clinicians. Can run via cloud, in-country hosting, or on-site/edge in hospitals depending on connectivity and data controls.	Shared local language LLM, locally hosted and fine-tuned on an open-source base for citizen and civil-servant interactions. Positioned as enabling layer for multiple downstream applications (e.g., voice-enabled public services, or other interfaces/use cases).

cenfri

Plotting example use cases across the 5 factors

	Coding assistant for government teams	Automated grading of exam papers	Weather forecasting + multi-channel alerts	AI radiology triage in hospitals	Local language LLM (shared capability)
AI Integration					
Standalone Tools	**x**	**x**	**x**		
Workflow Embedded	x			**x**	
System-of-Record Embedded					
Cross-System Orchestration					**x**
AI Autonomy					
Inform					
Propose	**x**			**x**	
Execute with Approval		**x**			
Execute within Guardrails			**x**		**x**
Deployment Model					
External API	**x**				
Hybrid (API + Local Data)	x	x	x	**x**	
In-Country Hosted		**x**	**x**		**x**
Distributed Edge					
Local Adaptation					
Consume	**x**			**x**	
Augment		**x**	x	x	
Tune			x	x	**x**
Train			**x**		
Need for Compute					
CPU-native	**x**		**x**		
GPU-assisted		**x**			
GPU-dependent				**x**	**x**
GPU cluster scale					

How you use AI

How you deploy AI

cenfri

Thank you

Stefan Steffen

Stefansteffen@cenfri.org

About Cenfri

Cenfri is a global think-tank and non-profit enterprise that bridges the gap between insights and impact in the financial sector. Cenfri's people are driven by a vision of a world where all people live their financial lives optimally to enhance welfare and grow the economy. Its core focus is on generating insights that can inform policymakers, market players and donors who seek to unlock development outcomes through inclusive financial services and the financial sector more broadly.

cenfri

In partnership with

mastercard foundation

Glossary of terms

Term	Full name	Description
GPU	Graphics Processing Unit	A special chip originally designed to handle visuals in computers and gaming. Due to their effectiveness at rapid calculations they got repurposed to run AI simulations.
CPU	Central Processing Unit	The main general-purpose brain of a computer that handles basic tasks and instructions, including running AI software once it's built.
Agentic AI	Agentic Artificial Intelligence	AI systems that can act on their own. They take decisions and actions rather simply than responding to basic prompts from humans.
LLM	Large Language Model	An AI trained on enormous amounts of text so it can understand, summarise, and generate written content.
API	Application Programming Interface	A digital messenger that lets different software Interoperate (speak with each other), so developers can easily plug existing AI services into their own apps.
Guardrails	Guardrails	Rules and limits built into AI systems to keep outputs safe, appropriate, and aligned with human values.
Distributed Edge	Distributed Edge Computing	Processing data locally on a device (like a phone or local box) instead of sending it to a central cloud.
Cluster Scale	Cluster Scale	Linking lots of ordinary computers or GPUs together so they act as one powerful machine capable of large-scale AI work.