Jarkko Moilanen.

JARKKO MOILANEN, PHD

AI CENTERS OF EXCELLENCE FROM AI ACTIVITY TO WORKING SYSTEMS

Operating model, economics, and implementation blueprint

WHITEPAPER

Dr. Jarkko Moilanen | September 2026

jarkkomoilanen.com

STRATEGY | PORTFOLIO | GOVERNANCE | DELIVERY

Contents

This report covers the business case, operating model, team, technology, economics, implementation roadmap, risks, and reusable governance templates for an enterprise AI Center of Excellence.

Executive summary

An AI Center of Excellence, or AI CoE, is an enterprise operating structure that concentrates AI strategy, standards, expertise, governance, platforms, reusable assets, and enablement while connecting AI investment to business outcomes. IBM defines it as a structure dedicated to organization-wide AI adoption, optimization, and governance. AWS describes a similar model that links business strategy to AI delivery, reusable capabilities, governance, and enterprise-scale platforms.

The strongest case for an AI CoE is not that AI requires another central technology department. It is that decentralized AI adoption creates recurring enterprise problems: duplicated pilots, inconsistent architecture, fragmented model procurement, unmanaged data and security risks, scarce specialist talent, weak paths from prototype to production, and poor visibility into value and cost. IBM and AWS both identify these problems as core reasons for establishing a CoE. BCG's AI research also finds a large gap between experimentation and material financial value, with only a small minority of surveyed organizations reaching substantial value at scale.

For an industry-unspecified enterprise, the recommended long-term model is usually a federated or hybrid CoE:

Central CoE: strategy, portfolio management, standards, common platforms, security patterns, responsible AI, evaluation methods, MLOps/LLMOps, reusable components, training, and enterprise vendor management.

Business units: business cases, process redesign, domain expertise, adoption, product ownership, and benefits realization.

Joint product squads: delivery and operation of production AI products.

Independent risk, legal, privacy, security, and audit functions: challenge and oversight according to risk level.

This follows the centralized-enablement, distributed-execution pattern described by IBM, AWS, and the FinOps Foundation. The FinOps Foundation's 2026 survey is useful adjacent evidence: 81% of responding FinOps organizations use centralized-enablement or hub-and-spoke structures, illustrating how a small standards-and-platform team scales through distributed ownership rather than centralizing every activity.

A sound AI CoE should be measured as a business-value system, not by model count or pilot count. Its scorecard should include realized financial value, benefit realization versus approved business cases, time to production, pilot-to-production conversion, adoption, workflow penetration, reuse, model/application quality, risk events, compliance coverage, platform reliability, and unit economics such as AI cost per completed business transaction. AWS explicitly recommends quantifiable KPIs such as cost savings, revenue gains, user satisfaction, time savings, and time to market.

For planning purposes, this report estimates dedicated AI CoE capacity at roughly 8 to 15 FTE-equivalents for a small enterprise, 25 to 45 for a medium enterprise, and 60 to 120 for a large enterprise, including dedicated domain capacity but allowing legal, risk, security, and business specialists to remain organizationally outside the CoE. These are planning ranges, not published market benchmarks. They reflect the multidisciplinary responsibilities identified by IBM and AWS.

Indicative first-year budgets, assuming U.S.-like enterprise labor economics and no large-scale foundation-model pretraining, are approximately $2.5 million to $6 million for a small organization, $8 million to $20 million for a medium organization, and $22 million to $65 million or more for a large enterprise. These are scenario estimates rather than industry benchmarks. U.S. BLS data puts the May 2025 median wage at $120,230 for data scientists and $135,980 for software developers, while private-industry benefits represented roughly 30% of employer compensation in June 2026. Specialized architecture, security, legal, leadership, and AI engineering roles push blended enterprise costs higher.

The first 90 days should produce a charter, executive sponsor, portfolio governance process, AI inventory, risk taxonomy, reference architecture, approved technology path, baseline value scorecard, initial training, and two or three tightly selected lighthouse use cases. The first year should move beyond pilots: productionize the strongest use cases, establish automated MLOps/LLMOps and monitoring, institutionalize risk gates, establish FinOps for AI, create reusable services, and begin shifting delivery ownership toward domain teams. NIST's Govern, Map, Measure, Manage structure and ISO/IEC 42001 provide strong foundations for this governance system.

The central management principle is:

> Treat the AI CoE as a temporary concentration of scarce capabilities that evolves into an enterprise-wide operating system for AI, rather than a permanent factory that builds every AI solution itself.

AWS explicitly describes the CoE's role as including reusable guidance and capabilities, while IBM distinguishes centralized, federated, and hybrid models. BCG similarly argues that centralized AI hubs should evolve as business-unit capability matures.

Definition, business case, and value model

An AI CoE sits between strategy and execution. It is broader than a data-science team and narrower than the entire technology organization.

Its purpose is to make AI repeatable, governable, economically rational, and scalable.

A mature CoE normally owns or coordinates six functions:

Function	AI CoE responsibility	Primary outcome
Strategy and portfolio	Translate corporate priorities into an AI portfolio, rank opportunities, stop low-value initiatives	Capital directed toward high-value use cases
Engineering standards	Reference architectures, development patterns, MLOps/LLMOps, evaluation methods	Faster, repeatable production delivery
Platforms	Shared data/model access, model gateways, registries, pipelines, observability	Reduced duplication and lower marginal delivery cost
Governance	Risk classification, evaluation standards, documentation, monitoring, responsible-AI controls	Controlled operational and regulatory risk
Talent and enablement	Training, communities of practice, mentoring, playbooks	Broader organizational AI competence
Value management	Business cases, baselines, benefits tracking, AI cost allocation	Visible ROI and accountability

This scope aligns closely with IBM's AI CoE model and AWS's definition of CoE outputs as both guidance, such as best practices and lessons learned, and reusable capabilities, such as people, tools, solutions, and templates.

The CoE should not normally own all enterprise AI forever. Business functions must retain accountability for their workflows, economics, customer outcomes, and adoption. Centralizing these responsibilities in an AI team creates an "AI supplier versus business customer" relationship that slows transformation.

Business drivers

The principal business drivers are:

Enterprise problem	Why it occurs	AI CoE response
Too many disconnected pilots	Low entry barriers and weak portfolio governance	Common intake, scoring, funding, stage gates
Poor production conversion	Prototype engineering differs from production requirements	Product engineering and MLOps standards
Repeated work	Teams independently build RAG stacks, connectors, evaluation tools and controls	Reusable services and reference architectures
Scarce expertise	AI, data, security and governance skills are distributed unevenly	Central expert pool plus embedded domain teams
Technology proliferation	Individual teams purchase models and AI tools independently	Model/vendor catalogue and architecture governance
Regulatory and reputational risk	Controls arrive after development	Risk classification and control design at intake
Uncontrolled AI expenditure	Tokens, GPUs, SaaS AI and data costs scale with consumption	AI FinOps and unit economics
Weak adoption	Technical teams optimize models rather than workflows	Product management and change management
Unclear ROI	Productivity gains are reported without financial baselines	Finance-approved benefit measurement
Slow organizational learning	Lessons stay inside isolated teams	Repositories, patterns, training and communities

IBM specifically identifies duplicated work, uneven governance, poor data quality, and limited business impact as consequences of scattered AI experimentation. AWS lists unclear business cases, inability to scale past proof of concept, governance gaps, and scarce talent among the problems a CoE addresses.

BCG's work adds an important operating-model point. Its widely used "10-20-70" transformation heuristic assigns only 10% of transformation effort to algorithms, 20% to technology and data, and 70% to people, processes, and organizational change. This is a consulting heuristic, not a scientific allocation formula, but it correctly warns against treating AI transformation as an engineering program alone.

Expected gains

An AI CoE should create value through four mechanisms.

First, direct use-case economics: revenue increase, reduced operating cost, avoided external spend, lower losses, higher asset utilization, or increased capacity.

Second, portfolio economics: stop weak projects earlier and direct scarce talent and compute toward higher-value initiatives.

Third, reuse economics: common model gateways, data products, evaluation libraries, security patterns, orchestration components, and monitoring reduce the marginal cost of each new AI application.

Fourth, risk economics: fewer security incidents, compliance failures, unsuitable model deployments, uncontrolled vendors, data leaks, and production failures.

AWS cites faster time to market, improved ROI, optimized risk management, structured upskilling, standardized scaling, and improved innovation prioritization as CoE benefits.

KPI and ROI framework

Do not make "number of AI projects" a primary success metric. It rewards activity.

A stronger scorecard follows.

KPI	Formula or definition	Why it matters	Illustrative first-year target
Realized annual value	Verified revenue contribution + verified savings + avoided loss	Core economic outcome	Set portfolio-specific
Benefit realization	Realized benefit ÷ approved business-case benefit	Detects optimistic business cases	>70% after stabilization
ROI	(Realized benefits - total AI cost) ÷ total AI cost	Investment return	Positive for mature production portfolio
NPV	Discounted benefits - discounted lifecycle costs	Captures multi-year economics	>0
Payback period	Time until cumulative benefits exceed cumulative cost	Capital discipline	Use-case specific
Cost per successful transaction	AI platform/model/data cost ÷ successful business tasks	GenAI/agent unit economics	Declining quarter over quarter
Time to production	Approved use case to production release	Delivery velocity	30% to 50% improvement from baseline
Pilot-to-production rate	Production deployments ÷ completed pilots	Detects pilot factories	>50% for funded lighthouse portfolio
Reuse ratio	Solutions using approved shared assets ÷ new solutions	Platform leverage	>60% where applicable
User adoption	Active target users ÷ eligible users	Adoption	>60% for mature internal applications
Workflow penetration	AI-assisted transactions ÷ eligible transactions	Stronger than logins alone	Use-case specific
Quality	Domain-specific accuracy, task success or acceptance rate	Business fitness	Predefined acceptance threshold
Human override	Human corrections ÷ AI recommendations/actions	Failure and trust signal	Trend downward after stabilization
Safety incident rate	Material AI incidents per defined usage unit	Risk control	Zero severe incidents
Governance coverage	Production systems with inventory, risk classification, owner and monitoring ÷ total	Control completeness	100%
Availability	Successful service time ÷ required service time	Production reliability	Based on business criticality
Model/data drift response	Time from material drift to triage/remediation	Operational resilience	Hours or days, not months
AI cost variance	Actual cost - forecast cost	Financial control	<10% to 15% for mature workloads
Training completion	Required staff completing role-specific AI training	Organizational readiness	>90% for governed user groups

AWS recommends business-value KPIs including cost savings, revenue gains, user satisfaction, time savings, and time to market. NIST's risk-management model supports measuring and managing AI risks continuously rather than treating governance as a one-time approval.

The target values above are planning examples, not universal benchmarks. Establish baselines before deployment and set targets against the economics and risk of each workflow.

A finance-grade ROI model should use:

```
Net AI benefit = incremental contribution margin + verified cost reduction + capacity value actually redeployed + expected loss avoided - incremental operating costs
```

```
ROI = net AI benefit / total AI investment
```

For productivity use cases, distinguish three values:

Time saved measures efficiency.

Capacity released measures work no longer required.

Financial value measures capacity eliminated, redeployed to economically valuable work, or converted into incremental output.

Counting every reported hour saved as cash savings usually overstates ROI.

This distinction has become important as enterprises scale AI spending. The FinOps Foundation's 2026 survey found 98% of respondents now manage AI expenditure, up from 63% in 2025 and 31% in 2024. Respondents cite visibility, business-unit allocation, and AI value/ROI as major difficulties.

Operating model, governance, and RACI

Organizational structures

Structure	How it works	Advantages	Main failure mode	Best fit
Centralized	One central AI organization owns most expertise and delivery	Fast standardization, concentrated talent, strong controls	Queue bottlenecks, weak domain ownership	Small organizations, early maturity, highly constrained environments
Federated, hub-and-spoke	Central hub owns platform, standards and enablement. Domain teams deliver	Scales domain ownership while preserving standards	Standards ignored if hub lacks authority	Medium and large enterprises
Hybrid, risk-tiered	Central team directly controls shared/high-risk capabilities while domains own lower-risk delivery	Balances autonomy and control	Complex decision rights	Large or regulated enterprises
Virtual/community model	Small coordination function plus practitioners distributed across existing departments	Low structural cost	Weak enforcement and limited delivery capacity	Early discovery or small organizations

IBM explicitly identifies centralized, federated, and hybrid AI CoE models, each involving a different balance between consistency and local autonomy. AWS similarly treats the CoE as potentially centralized or federated.

For an unspecified organization, a practical progression is:

Months 0 to 6: relatively centralized.

Months 6 to 18: hub-and-spoke.

Beyond that: risk-tiered federation, with the CoE increasingly responsible for platform, governance, specialist expertise, reusable intellectual property, portfolio oversight, and enablement rather than every delivery task.

Governance architecture

The governance model should separate business ownership, technical assurance, and independent risk oversight.

A strong design has five layers:

Corporate or AI steering committee: portfolio priorities, investment, risk appetite, major exceptions.

AI CoE: standards, architecture, intake, portfolio analytics, evaluation methodology, reusable services.

Product/domain owners: workflow redesign, business requirements, adoption, benefit realization.

Platform/security/risk/legal: technology controls and specialist approvals.

Internal audit or equivalent independent assurance: tests whether the governance system works.

NIST's AI Risk Management Framework organizes activities around Govern, Map, Measure, and Manage, with governance acting across the lifecycle. NIST describes the framework as voluntary and is revising AI RMF 1.0 as of 2026.

ISO/IEC 42001:2023 provides a complementary management-system approach for establishing, implementing, maintaining, and continually improving an organization-wide AI management system. It applies across organization sizes and sectors.

For organizations operating in the European Union, regulatory design now matters directly. The EU AI Act entered into force on August 1, 2024, and its main application date was August 2, 2026, subject to phased provisions and subsequent transition arrangements. A global CoE should therefore maintain a jurisdictional regulatory overlay rather than treating one enterprise policy as sufficient everywhere.

A practical intake process should classify each use case by:

business criticality, data sensitivity, affected persons, degree of autonomy, financial/legal consequences, model type and external provider, customer-facing versus internal use, required human oversight, regulatory classification.

Low-risk internal productivity tools should take a lighter route than systems making consequential decisions.

Example RACI

```
R = Responsible. A = Accountable. C = Consulted. I = Informed.
```

Activity	Executive / AI Steering Committee	AI CoE	Business / Product Owner	Platform / IT	Risk, Legal, Privacy, Security
Enterprise AI strategy	A	R	C	C	C
Use-case business case	I	C	A/R	C	C
Portfolio prioritization	A	R	C	C	C
Reference architecture	I	A	C	R	C
Approved models/vendors	I	A/R	C	C	C
Data readiness	I	C	A	R	C
Risk classification	I	R	C	C	A
Model/application evaluation	I	A/R	C	C	C
Privacy/security assessment	I	C	C	R	A
Business acceptance	I	C	A/R	C	C
Production deployment	I	C	A	R	C
Risk exception	I	C	C	C	A
Production monitoring	I	A	C	R	C
Benefits realization	I	C	A/R	C	I
Enterprise AI training	I	A/R	R for local adoption	C	C

The central rule is to keep business outcome accountability with the business owner. The CoE should not declare its own projects economically successful.

For high-risk systems, introduce separate technical approval, business acceptance, and independent risk approval rather than forcing all accountability into one RACI cell.

Organization, roles, and hiring plan

IBM and AWS both describe an AI CoE as inherently multidisciplinary, spanning leadership, data science, AI/ML engineering, data, product, IT, security, legal, privacy, risk, compliance, business-domain specialists, and training.

Headcount should therefore reflect the number of concurrent production products and shared-platform responsibilities, not employee count alone.

The following ranges represent dedicated FTE-equivalent capacity. Some legal, privacy, security, audit, and business-liaison capacity will normally remain in existing departments.

Role family	Small enterprise	Medium enterprise	Large enterprise	Main responsibilities
CoE leadership and portfolio	1-2	2-4	4-8	Strategy, funding, portfolio, executive governance
Data engineering	1-2	3-6	8-15	Data products, pipelines, quality, integration
Data science / AI / ML engineering	2-4	5-10	12-25	Models, GenAI applications, agents, evaluation
MLOps / LLMOps	1	2-4	5-10	CI/CD, registries, deployment, monitoring
Platform / infrastructure / SRE	1-2	3-6	8-15	AI platform, compute, reliability, identity
Product / program management	1	2-4	5-10	Product discovery, roadmap, business cases
Responsible AI / model risk / governance	0.5-1	1-3	3-6	Risk classification, evaluation, policy, inventory
Legal / privacy / compliance capacity	0.2-0.5	0.5-1.5	1-3	Contracts, regulation, privacy, IP
Change management / training	0.5-1	1-3	3-6	Adoption, workflow redesign, training
Business liaisons / domain leads	1-2	3-8	8-20	Domain requirements and value ownership
Approximate total	8-15	25-45	60-120	Dedicated capacity, including embedded domain roles

These are implementation estimates for this report, not published benchmarks.

A small organization should avoid reproducing every specialist role as a separate job. One platform engineer might cover MLOps and infrastructure. Legal/privacy specialists may contribute part time. A senior product manager might own portfolio management.

A large organization should do the opposite: keep the central team lean enough to avoid becoming the sole delivery factory, then establish domain AI product squads.

A practical production squad often contains roughly six to nine dedicated people:

Squad capability	Typical dedicated capacity
Product manager / domain product owner	1
Data engineer	1-2
AI/ML engineer or data scientist	2-3
Software/MLOps engineer	1-2
UX/process/change specialist	0.5-1
Shared platform, security, legal and governance	Fractional/shared

This is a planning heuristic. Actual staffing changes substantially for computer vision, industrial edge AI, regulated decision systems, research-intensive ML, or large agentic applications.

Production engineering deserves substantial capacity. Google's MLOps guidance emphasizes continuous integration, continuous delivery, continuous training, deployment, and ongoing monitoring rather than treating a trained model as the finished product. The classic Sculley et al. paper on hidden technical debt also shows why production ML accumulates complexity through data dependencies, pipeline coupling, feedback loops, configuration, and monitoring requirements.

Hiring sequence

Period	Priority hires	Objective
Months 0-3	CoE lead, portfolio/product lead, AI architect, senior AI/ML engineer, MLOps/platform lead, data lead, governance lead, change/training lead	Establish minimum viable leadership and architecture
Months 4-6	Additional AI engineers, data engineers, product managers, platform engineers, evaluation/safety expertise, business liaisons	Deliver lighthouse products
Months 7-12	Domain pods, SRE/operations, FinOps, training scale-up, support, specialist governance	Industrialize and federate
Year 2 onward	Add staff based on production backlog and service demand, not technology enthusiasm	Scale economically

A hiring-plan template should track:

Field	Example
Role	Senior MLOps Engineer
Why needed	Reduce manual production deployment and monitoring
Linked capability	Deployment platform
Linked use cases	Customer assistant, demand forecast, claims model
Start date	Q2
Employment model	Permanent
Fully loaded annual cost	Finance estimate
Success measure	Deployment lead time, platform adoption
Dependency	Cloud landing zone and security architecture
Sunset/redeployment trigger	Domain engineering teams achieve target maturity

This prevents headcount plans from becoming disconnected from capabilities and business demand.

Technology, infrastructure, and budget

The CoE should define an enterprise AI platform as a set of capabilities, not as one vendor product.

Required technology stack

Layer	Required capabilities
Data foundation	Lake/lakehouse/warehouse access, data quality, catalog, lineage, metadata, access controls
Model access	Approved foundation models and ML frameworks, model gateway, routing, quotas
Development	Managed workspaces, notebooks/IDEs, repositories, secrets, reproducible environments
Knowledge / RAG	Document ingestion, chunking/indexing, search/vector services, source controls
Agent/tool layer	Tool registry, orchestration, permission controls, sandboxing
Evaluation	Offline test sets, task success, hallucination/grounding evaluation, adversarial tests, fairness where relevant
MLOps / LLMOps	CI/CD, model/prompt/version registries, automated testing, deployment and rollback
Governance	AI inventory, risk classification, approvals, documentation, evidence collection
Security	IAM, network controls, secrets, content/data controls, logging, vulnerability management
Observability	Application traces, quality, drift, latency, errors, token/compute consumption
FinOps	Cost allocation, tags, budgets, forecasts, anomaly detection, unit economics
Reuse catalogue	APIs, approved patterns, templates, reference implementations, training material

AWS, Google Cloud, and Microsoft all document production AI architectures that combine development, deployment, monitoring, security, governance, and lifecycle management rather than limiting the platform to model training.

Examples of integrated commercial ecosystems include AWS AI/ML services, Microsoft's Azure AI/ML stack, and Google Cloud Vertex AI. A large enterprise should still preserve architectural separation between business applications and model providers where practical. This reduces migration friction and supports model choice as price, performance, and risk profiles change.

Cloud, on-premises, and hybrid

Criterion	Public cloud	Private / on-premises	Hybrid
Initial deployment speed	Highest	Lowest	Medium
Upfront capital	Low	High	Medium
Elasticity	Strong	Capacity constrained	Strong if well orchestrated
Access to managed AI services	Broad	More limited	Broad for eligible workloads
Infrastructure operations	Provider-heavy	Enterprise-heavy	Shared
Data-sovereignty control	Depends on region/service	Highest direct control	High with workload placement
Hardware lifecycle risk	Lower for customer	Customer-owned	Mixed
Cost at unpredictable demand	Often attractive	Underutilization risk	Flexible
Cost at sustained high utilization	Depends heavily on workload and contracts	Potentially attractive after capital case	Workload-specific
Network/edge latency	Dependent on connectivity	Strong for local systems	Strong where placement is optimized
Vendor concentration	Higher without abstraction	Lower at model-infrastructure layer	Manageable
Power/cooling requirement	Provider responsibility	Enterprise responsibility	Partial
Best default	Early-stage and variable workloads	Sovereignty, edge or justified steady-state cases	Large enterprises with mixed constraints

This table is an analytical synthesis, not a universal cost ranking.

On-premises AI at meaningful scale requires more than buying GPUs. NVIDIA's enterprise AI infrastructure guidance emphasizes high-bandwidth networking, storage, accelerator

density, power and cooling, capacity planning, and cluster operations. These requirements make a private AI platform a capital and operations program.

The recommended default for an industry-unspecified organization is managed cloud for initial experimentation and shared services, or hybrid for a large enterprise with significant existing private infrastructure. Move workloads on-premises only when data sovereignty, latency, security architecture, workload economics, or operating constraints support a clear business case.

Budget scenarios

Industry is unspecified. The figures below are therefore planning ballparks, stated in 2026 U.S. dollars. They assume enterprise AI adoption using existing foundation models and standard ML rather than training frontier foundation models from scratch.

They also assume U.S.-like labor economics. BLS reports median 2025 annual pay of approximately $120,230 for data scientists and $135,980 for software developers. Private-sector employee benefits represented about 30% of employer compensation in June 2026, before adding recruitment, equipment, management overhead, contractors, and the salary premium often attached to senior specialist positions.

Scenario	Dedicated capacity	Labor and related cost	Platform/cloud/software	Consulting/training/change	Indicative first-year total
Small enterprise	8-15 FTE-equivalent	$1.5M-$3.5M	$0.3M-$1.5M	$0.2M-$1.0M	$2.5M-$6M
Medium enterprise	25-45	$5M-$10M	$1.5M-$6M	$0.5M-$2M	$8M-$20M
Large enterprise	60-120	$12M-$27M	$5M-$25M+	$2M-$8M+	$22M-$65M+

These are estimates produced for planning, not surveyed AI CoE budget benchmarks. They exclude acquisitions, major data-center construction, large private GPU clusters, frontier-model pretraining, and unusual sector-specific certification.

The largest cost drivers are:

Cost driver	Why it matters
Number of concurrent production use cases	Drives product, engineering, data and support capacity
Model choice	Model size and pricing strongly affect inference economics
User/request volume	GenAI costs often scale with usage
Context and agent complexity	Long context, tool calls and multi-step agents multiply consumption
Data readiness	Poor data creates engineering and governance work
Availability requirements	Higher resilience increases infrastructure and SRE cost
Risk level	High-risk systems require more testing, documentation and oversight
Private infrastructure	GPUs, networking, storage, power, cooling and operations
Vendor fragmentation	Duplicate platforms increase licensing and integration costs
External consulting	Speeds initial deployment but raises first-year cost
Geographic labor mix	Large effect on labor expense
Change effort	Workflow redesign, training and adoption often exceed technical effort

FinOps should start with the CoE rather than follow it later. In the FinOps Foundation's 2026 survey, AI cost management ranked as the top skill area teams wanted to develop, and respondents specifically requested granular monitoring of token usage, LLM requests, GPU utilization, pre-deployment costing, and business-value attribution.

Every production AI application should therefore carry a cost center, product/use-case identifier, business owner, environment, model/provider, budget, and unit-economic measure.

Ramp-up roadmap and action plans

A CoE should deliver value while it establishes governance. A year spent designing the perfect operating model before deploying anything usually loses executive support. The opposite extreme, shipping uncontrolled pilots, creates debt that later slows the enterprise.

A balanced first-year sequence is:

First-year delivery sequence

01	**Mobilize** Weeks 0-4	Sponsor, charter, inventory
02	**Design** Weeks 4-8	Governance, architecture, portfolio
03	**Lighthouse delivery** Weeks 8-16	Two or three priority use cases
04	**Industrialize** Months 4-8	MLOps, controls, reuse
05	**Scale and federate** Months 9-12	Domain pods, training, FinOps
06	**Optimize** Year 2+	Distributed ownership and improvement

AWS's CoE guidance follows a similar logic: executive sponsorship and mission, people and enablement, governance, infrastructure, security, prioritization, and operational excellence. NIST supports continuous governance and risk management across the AI lifecycle rather than a single compliance checkpoint.

Phase	Timing	Main work	Milestone	Estimated cross-functional effort
Mobilize	Weeks 0-4	Sponsor, charter, leadership, current-state inventory, funding	CoE formally mandated	3-6 person-months
Design	Weeks 4-8	Operating model, RACI, risk tiers, intake, architecture, KPI baseline	Governance and reference architecture approved	5-10 person-months
Lighthouse delivery	Weeks 8-16	Build 2-3 high-value use cases, evaluate, redesign workflows	At least one production-ready use case	16-40 person-months
Industrialize	Months 4-8	CI/CD, monitoring, model gateway, reusable services, documentation	Repeatable path to production	40-120 person-months
Scale and federate	Months 9-12	More domain squads, training, FinOps, service catalogue	Multiple domains delivering on shared standards	80-240 person-months
Optimize	Year 2 onward	Benchmark value, rationalize platforms, transfer capabilities	CoE shifts from builder to enterprise enabler	Ongoing

The effort figures are implementation planning estimates. They exclude broad employee training hours and part-time executive/business participation.

First 90-day action plan

Time	Required output
Days 0-30	Name executive sponsor and CoE leader. Approve mandate and provisional funding. Inventory production AI, pilots, SaaS AI, model providers, key data assets and existing skills.
Days 15-45	Establish steering committee. Define RACI, intake process, risk taxonomy, minimum controls, procurement policy and initial AI inventory.
Days 30-60	Approve reference architecture and preferred platform strategy. Select model/provider approach. Set cloud and data security patterns. Establish cost tagging.
Days 30-60	Create portfolio scoring model based on strategic value, feasibility, data readiness, adoption complexity, risk and expected economics.
Days 45-75	Select two or three lighthouse initiatives, each with a named executive/business owner and measurable baseline.
Days 45-90	Build evaluation datasets, technical prototypes, workflow redesign and adoption plan. Establish role-based AI training.
By day 90	Review results. Stop weak initiatives. Approve production funding for successful cases. Publish CoE standards, service catalogue, KPI dashboard and first-year hiring plan.

The use-case filter should be strict. IBM recommends prioritizing use cases with a clear business owner, defined workflow, sufficient data, measurable goals, and a realistic production path.

A useful prioritization score is:

```
Priority score = strategic value × economic value × adoption probability × technical feasibility × data readiness × risk-adjustment factor
```

This avoids choosing projects solely because an impressive demo is easy to produce.

First-year action plan

By the end of year one, a credible CoE should have accomplished the following:

Outcome area	Year-one state
Strategy	AI portfolio explicitly linked to corporate priorities
Governance	All known production AI inventoried, owned, risk classified and monitored
Delivery	Several use cases operating in production, not only demonstrations
Economics	Business cases baselined and benefits reviewed with finance
Platform	Reusable secure development and production path available
MLOps/LLMOps	Versioning, testing, deployment, monitoring and rollback automated for priority patterns
Cost	AI expenditure allocated to applications/business units with budgets and unit economics
Risk	Evaluation, security, privacy and legal gates integrated into delivery
Reuse	Common components published through a service/reuse catalogue
Talent	Role-based training and practitioner community operating
Federation	Selected business units own delivery while following common controls
Executive governance	Quarterly portfolio review stops, scales or redirects investments

A useful first-year executive milestone is not "100 AI use cases." It is a small set of production systems with verified value, plus a delivery system that makes the next wave faster and safer.

Risks, case studies, and reusable templates

Common failure modes

Failure mode	Early warning	Mitigation
Pilot factory	Many demos, few production users	Production path and business owner required before pilot funding
Technology-first portfolio	Use cases start with a model rather than a business problem	Score business outcome before technical novelty
Central CoE bottleneck	Growing delivery queue	Federate delivery, centralize reusable capabilities
AI ivory tower	Business units treat CoE as external supplier	Embedded product/domain ownership
Shadow AI	Unknown SaaS/model use grows	AI inventory, approved catalogue, procurement and identity controls
Excessive governance	Low-risk experimentation takes months	Risk-tiered controls
Governance too late	Legal/security issues found before launch	Risk classification during intake
Poor data foundation	Teams spend most delivery time finding and repairing data	Data ownership, quality measures, reusable data products
MLOps gap	Manual deployment and no production monitoring	Standard pipeline, registry, observability and rollback
Model quality decay	Rising overrides or complaints	Monitoring, evaluation sets, drift detection and retraining/reconfiguration
No adoption	Good technical metrics, weak workflow usage	Process redesign, manager accountability, training
Fake productivity ROI	Hours "saved" but operating expense unchanged	Finance-approved benefit realization rules
Runaway model spending	Usage rises faster than economic value	Cost tags, budgets, model routing and unit economics
Vendor lock-in	Application logic directly tied to one provider everywhere	Standard interfaces and abstraction where economics justify it
On-prem overbuild	Expensive accelerators sit idle	Benchmark utilization before capital commitment
Key-person dependency	Only a few experts understand systems	Pairing, documentation, rotation and training
Talent empire	CoE headcount rises indefinitely	Capability-transfer objectives and federation milestones

The MLOps risks are well established. Sculley et al. documented how data dependencies, feedback loops, configuration and system interactions create hidden technical debt in ML systems. Modern cloud MLOps guidance addresses the same class of problem through automation, reproducibility, monitoring, and lifecycle management.

On organizational risk, AWS and IBM both emphasize business alignment, governance, cross-functional teams, standardized reusable capabilities, and clear executive sponsorship.

Case evidence

Public CoE case evidence should be treated carefully. Many published examples come from technology vendors and represent customer-reported results rather than independent controlled studies.

Organization	Model / initiative	Reported outcome	Lesson
NTT DATA	AI CoE in the CIO organization using Microsoft Fabric and Azure AI technologies	Microsoft reports at least 50% faster time to market for relevant AI/data-agent work, with applications across HR, operations and sales	A central CoE plus shared data/AI platform speeds reusable delivery
Unieuro	Internal AI CoE using Google Cloud data and AI services	AI use expanded across legal, back office, retail and training. Employee-facing chatbots supported more than 100 employees in the cited case	Start with internal workflows and build a reusable enterprise capability

Organization	Model / initiative	Reported outcome	Lesson
Hearst, adjacent Cloud CoE case	Cloud Center of Excellence used a GenAI self-service support assistant	AWS reports support/guidance requests fell around 70% in the first month and 76% in the following month	CoEs scale better when recurring expert knowledge becomes self-service

The NTT DATA result is documented in a Microsoft customer example. The Unieuro case is described by Google Cloud. The Hearst example is a Cloud CoE rather than a pure AI CoE, so it should be read as an adjacent operating-model lesson.

The broader lesson is stronger than any single vendor case: reusable architecture, common data, product ownership, governance, and self-service mechanisms create leverage. A CoE that repeatedly performs bespoke project work does not.

AI CoE charter template

Charter element	Template
Purpose	"Accelerate measurable and responsible enterprise value from AI by providing common strategy, platforms, standards, specialist expertise and enablement."
Executive sponsor	Named C-suite executive
Scope	ML, GenAI, AI agents, third-party AI, AI embedded in enterprise applications
Out of scope	Business-process ownership and ordinary software delivery unless explicitly assigned
Customers	Business units, product teams, technology teams, control functions
Core services	Portfolio intake, architecture, AI platform, evaluation, MLOps/LLMOps, governance, training, reusable components
Decision rights	Standards, approved model/provider catalogue, risk classification method, architecture exceptions
Business accountability	Domain/product owner
Risk accountability	Relevant independent risk/legal/security authority
Funding model	Central platform budget + use-case/domain funding
Measures	Realized value, adoption, production conversion, delivery time, risk, reuse, platform reliability, unit economics
Cadence	Monthly portfolio review, quarterly steering review
Evolution criterion	Transfer repeatable delivery capabilities to domains as maturity increases

The charter should explicitly state what the CoE does not own. Without that boundary, it tends to absorb unrelated analytics, data engineering, automation, innovation, and IT responsibilities.

KPI scorecard template

Use one row per KPI:

Metric	Definition	Baseline	Target	Current	Owner	Data source	Frequency	Corrective action
Realized value	Finance-verified annualized benefit	$0	$X	$Y	Business owner	Finance	Quarterly	Revisit workflow
Time to production	Approval to production	180 days	90	105	CoE/product	DevOps	Monthly	Remove bottleneck
Adoption	Active ÷ eligible users	0%	70%	52%	Product owner	App analytics	Monthly	Training/process change
Cost per task	Total run cost ÷ successful tasks	$X	$Y	$Z	Product/FinOps	Billing + logs	Weekly	Routing/right-sizing
Severe AI incidents	Defined severity threshold	0	0	0	Risk	Incident system	Monthly	Stop/remediate
Governance coverage	Governed production apps ÷ all production apps	35%	100%	82%	AI governance	AI inventory	Monthly	Close inventory gaps

Every metric needs a named owner, denominator, data source, measurement cadence, and management action. Otherwise it is reporting rather than governance.

Recommended target-state blueprint

For most medium and large enterprises with no specified industry constraint:

Dimension	Recommended default
Structure	Federated hub-and-spoke
Executive home	CIO/CTO/CDO/CAIO-level, with strong business sponsorship
Funding	Central platform/governance funding plus domain use-case funding
Central team	Lean specialist hub
Delivery	Cross-functional domain product squads
Governance	Risk-tiered, based on NIST/ISO principles plus jurisdiction-specific requirements
Platform	Managed cloud or hybrid with standardized model/data interfaces
AI models	Multi-model where justified, centrally approved
Data	Governed enterprise data products and access patterns
Operations	Automated MLOps/LLMOps, evaluation, observability and incident management
Economics	Finance-verified benefits plus workload-level AI FinOps
Talent	Central experts, distributed champions, systematic training
Evolution	Central build role declines as domain maturity rises

This target aligns with IBM's federated/hybrid structures, AWS's emphasis on shared guidance and reusable capabilities, NIST's lifecycle risk-management approach, ISO/IEC 42001's management-system model, and current FinOps evidence favoring centralized enablement with distributed execution.

The strategic test for the CoE after its first year is simple: new AI products should become faster to launch, cheaper to operate, easier to govern, easier to reuse, and more likely to produce verified business value than they were before the CoE existed. If the organization instead has more committees, more pilots, a larger central AI team, and no finance-verified portfolio return, the CoE has optimized activity rather than outcomes.

Selected public references

The source report cited IBM, AWS, BCG, the FinOps Foundation, NIST, ISO, the European Commission, major cloud providers, public case studies, and U.S. labor data. The links below provide direct access to several core public sources. Internal research citation markers from the source file were removed because they do not function outside ChatGPT.

1. IBM, AI center of excellence
2. NIST, AI Risk Management Framework
3. ISO, ISO/IEC 42001 AI management systems
4. European Commission, regulatory framework for AI
5. Google Cloud, MLOps continuous delivery and automation pipelines
6. Sculley et al., Hidden Technical Debt in Machine Learning Systems
7. U.S. Bureau of Labor Statistics, Data Scientists
8. U.S. Bureau of Labor Statistics, Software Developers

Planning note: Headcount, effort, and budget figures in this report are scenario estimates, not market benchmarks. Validate them against organizational size, geography, risk profile, delivery scope, and existing platform maturity before investment approval.