# Organizational AI Compute Governance Checklist

**Type:** Checklist

**Target Audience:** IT Infrastructure, FinOps, AI Platform Teams

This checklist helps organizations manage the physical and financial risks associated with high-performance AI hardware, including GPU clusters, cloud compute resources, and specialized AI accelerators.

## 1. Resource Inventory

Track GPU and specialized chip usage across the organization.

- ☐ Maintain a centralized registry of all AI compute resources (on-prem and cloud)
- ☐ Document GPU types, quantities, and locations (e.g., NVIDIA H100, A100, TPUs)
- ☐ Track utilization rates for each resource pool (target: >70% for cost efficiency)
- ☐ Map resources to projects/teams for chargeback and accountability
- ☐ Document capacity limits and scaling thresholds

## 2. Cost Monitoring & FinOps

Establish budgets and automated alerts for cloud compute costs.

- ☐ Set monthly/quarterly compute budgets per team and project
- ☐ Configure automated alerts at 50%, 75%, and 90% budget thresholds
- ☐ Implement auto-shutdown policies for idle GPU instances
- ☐ Review spot/preemptible instance usage for non-critical workloads
- ☐ Conduct monthly cost anomaly reviews (identify "runaway" training jobs)
- ☐ Establish reserved instance strategy for predictable workloads

## 3. Access Controls

Define who can provision high-end GPU resources and for what purpose.

- ☐ Implement role-based access control (RBAC) for compute provisioning
- ☐ Require manager approval for GPU instances above defined tier (e.g., 8+ GPUs)
- ☐ Maintain audit logs of all compute resource provisioning

- [ ] Define approved use cases for high-performance compute (training vs. inference)
- [ ] Implement quotas per user/team to prevent resource hoarding

## 4. Supply Chain Assessment

Identify dependencies on hardware vendors or geographic regions.

- ☐ Document primary GPU/chip suppliers and lead times
- ☐ Assess geographic concentration risk (e.g., single data center region)
- ☐ Identify backup cloud providers for critical AI workloads
- ☐ Review export control implications for AI hardware procurement
- ☐ Maintain relationships with multiple hardware vendors
- ☐ Plan for hardware refresh cycles (typically 3-5 years for AI accelerators)

## 5. Sustainability Tracking

Monitor energy consumption and carbon footprint of compute clusters.

- ☐ Track power usage effectiveness (PUE) for on-premises data centers
- ☐ Monitor kWh consumption per training run / model
- ☐ Calculate carbon footprint using cloud provider sustainability dashboards
- ☐ Set energy efficiency targets for AI workloads
- ☐ Evaluate renewable energy options for data center power
- ☐ Report AI compute environmental impact in ESG disclosures

## Sign-Off

**Reviewed By:** _____ **Date:** _____

**Approved By:** _____ **Date:** _____