



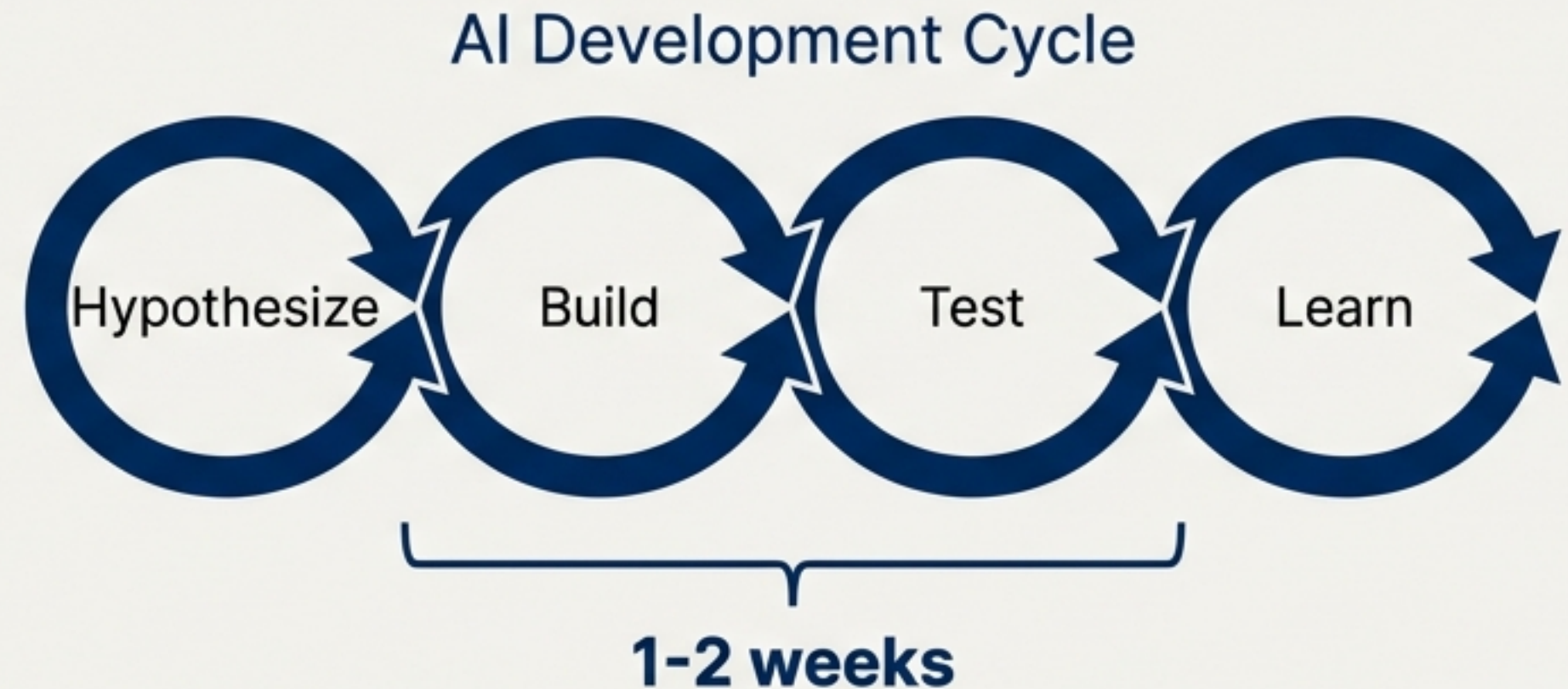
# The AI Scaling Paradox

How to Unleash the Power of Small Teams  
Without Sacrificing Enterprise Coherence



# The Pace of AI Has Broken Traditional Structures

- Generative AI and foundation models have compressed innovation cycles from years to months.
- Unlike previous technology waves, AI demands continuous experimentation and tight feedback loops.
- Rigid structures—functional silos, waterfall processes, and multi-layer approval gates—are now a competitive liability. They were designed for stability, not speed.





# The Emergent Playbook: Small, High-Agency Teams

The most effective response is to empower small, cross-functional teams with the authority to execute.

## Key Characteristics of High-Agency AI Teams



### Compositional Autonomy

3-7 members blending senior ML/AI engineers, domain experts, and product managers.



### Decision Rights

Authority to select models, define metrics, and deploy to users without routine external approval gates.



### Protected Time

Dedicated, full-time allocation, not part-time matrix assignments.



### Resource Access

Direct budget control for compute, APIs, and necessary tooling.

## Evidence of Success

### Ship in Weeks, Not Quarters

Organizations report shipping production AI applications in weeks, not quarters.

### Higher Production Adoption

Moving from proof-of-concept to operational use at a significantly higher rate.

### ROI Driven by Speed

Reducing time-to-value from 12 months to 3 months is the primary driver.



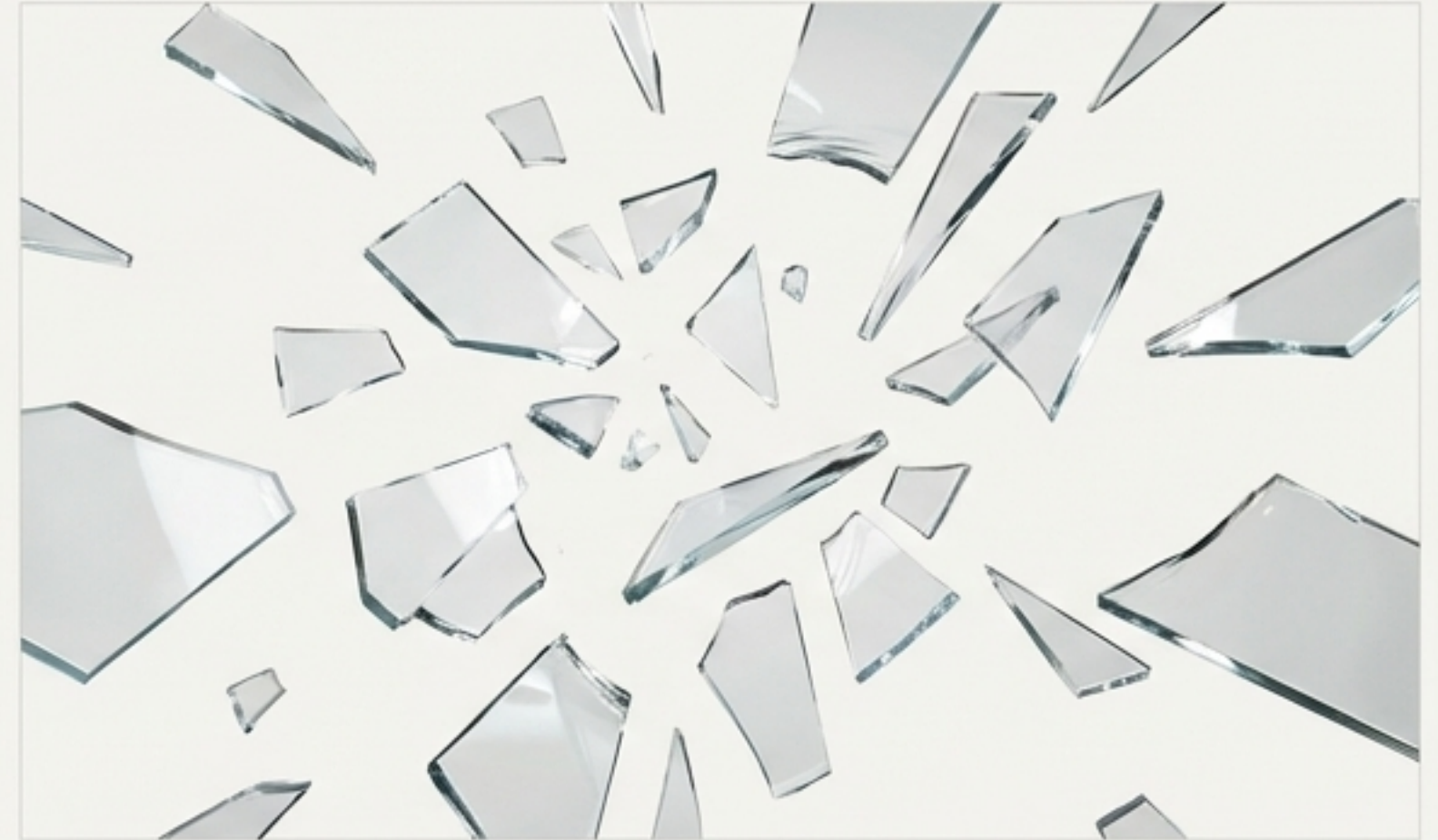
# Scaling Autonomy — Creates a Dangerous Dilemma

## Bureaucratic Stagnation



Reverting to command-and-control structures kills the very agility that made initial projects successful.

## Uncoordinated Chaos

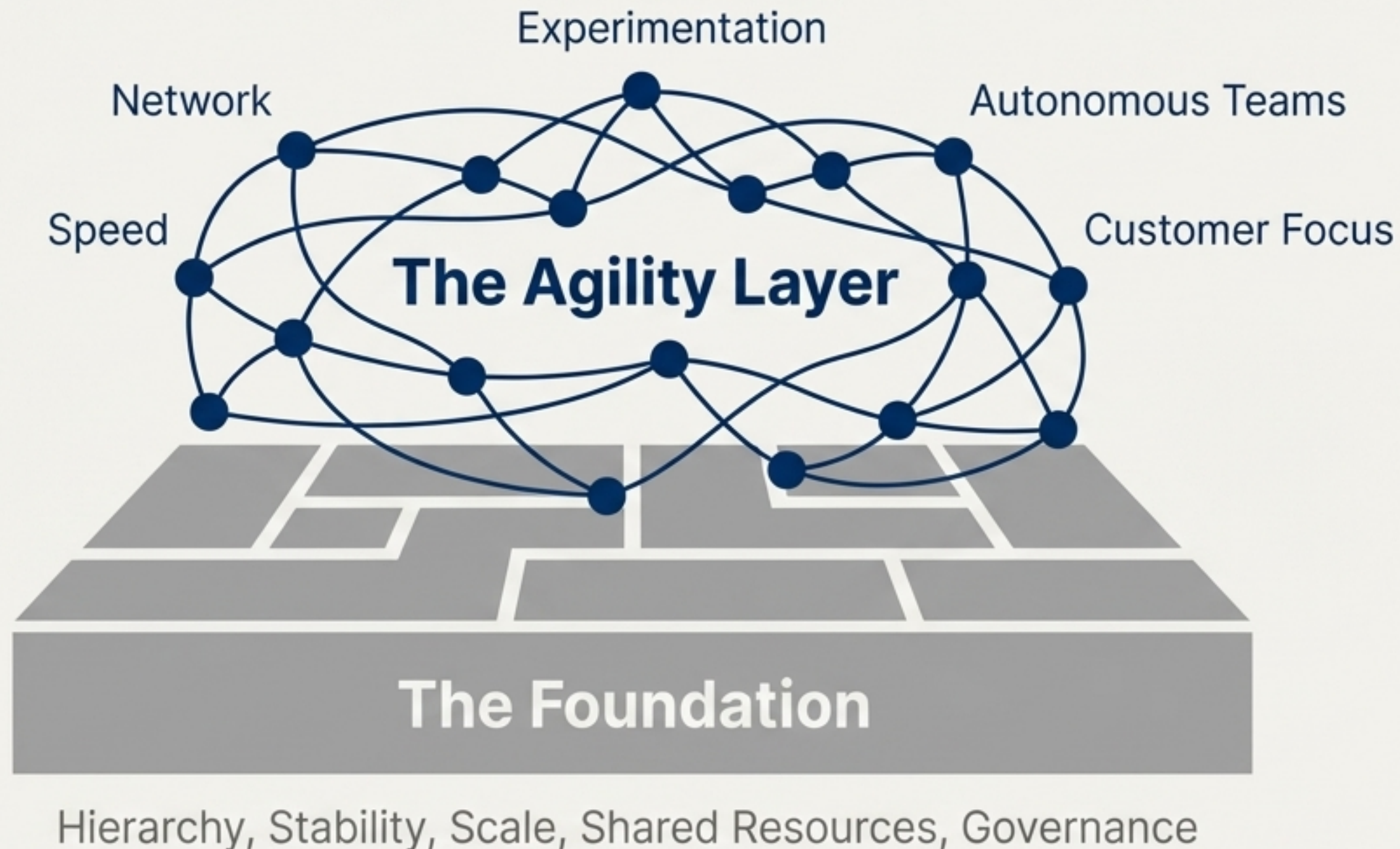


Proliferating teams without coordination leads to duplicated effort, fragmented standards, and misallocated scarce AI expertise.

How do we scale the benefits of autonomy without descending into chaos?



# The Resolution: A Dual Operating System



The solution is not to choose between **speed** and **stability**, but to **design for both**.

This model preserves team autonomy for execution (Agility Layer) while building the connective tissue for enterprise-level coherence and capability (The Foundation).



# Build a Platform to Make the Right Path the Easy Path

## Problem

Inconsistent tooling and duplicated infrastructure across teams create friction and waste.

## Solution

A Transparent Platform & Standards Strategy

Provide pre-approved, shared infrastructure that handles undifferentiated heavy lifting (deployment, monitoring, compliance) while preserving team flexibility.

## Key Components



Curated Model Library



Automated Deployment Pipelines  
(with embedded security/compliance)



Federated Data Catalog



Shared Experiment Tracking



Implemented a shared AI platform for multiple autonomous teams.

**“Substantially reduced duplicated infrastructure work while maintaining team velocity.”**



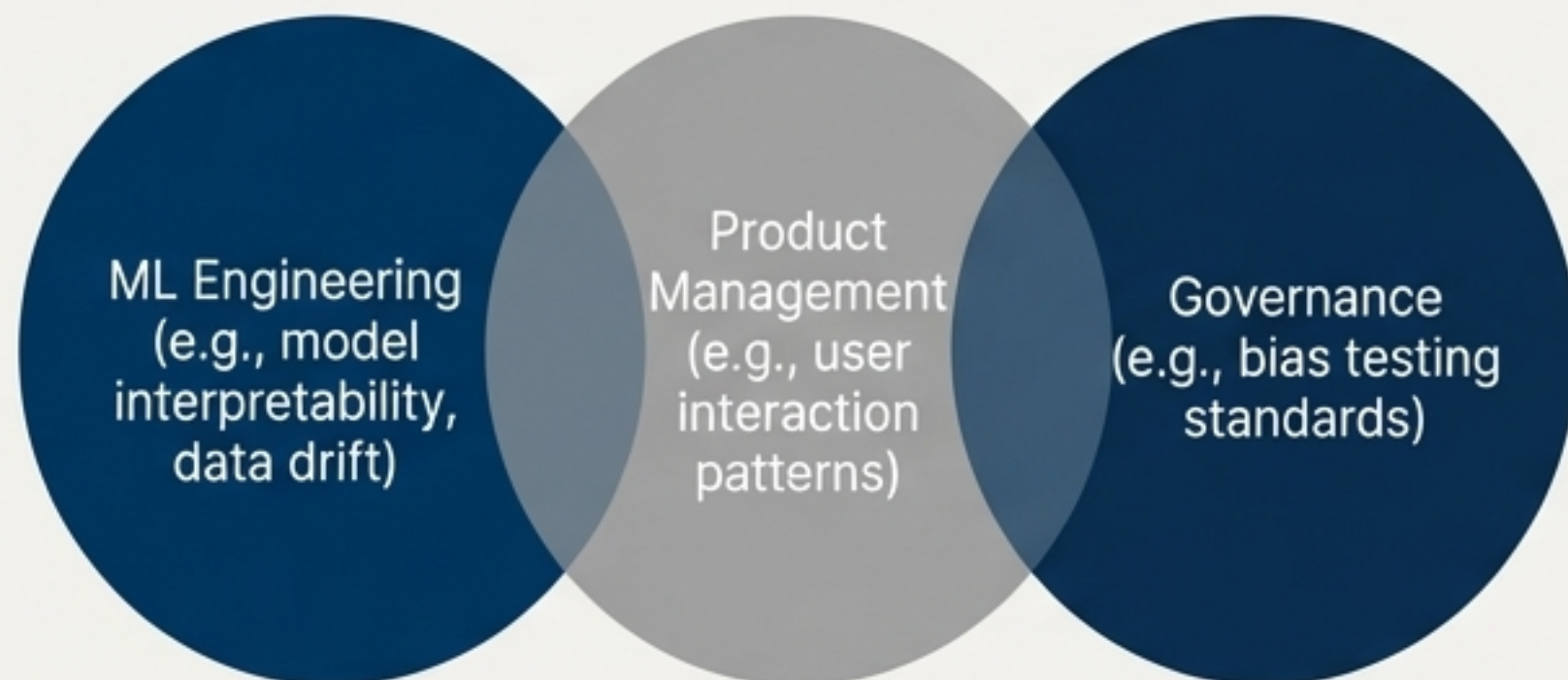
# Foster Distributed Coordination via Guilds

**Problem:** Expertise becomes siloed and mistakes are duplicated when teams operate in isolation.

## **Solution:**

Establish Guilds (Communities of Practice). These are horizontal, voluntary forums that span autonomous teams to facilitate knowledge exchange without creating approval bottlenecks.

## Common Guilds in AI



## Proof Points



**Action:** Established AI guilds to coordinate on challenges like regulatory compliance.

**Result:** Prevented architectural fragmentation while preserving decision speed.



**Action:** Uses technical forums for drug discovery AI teams.

**Result:** Reduced duplicated research effort as teams learn from each other's work.



# Define Who Decides What with Explicit Authority Matrices

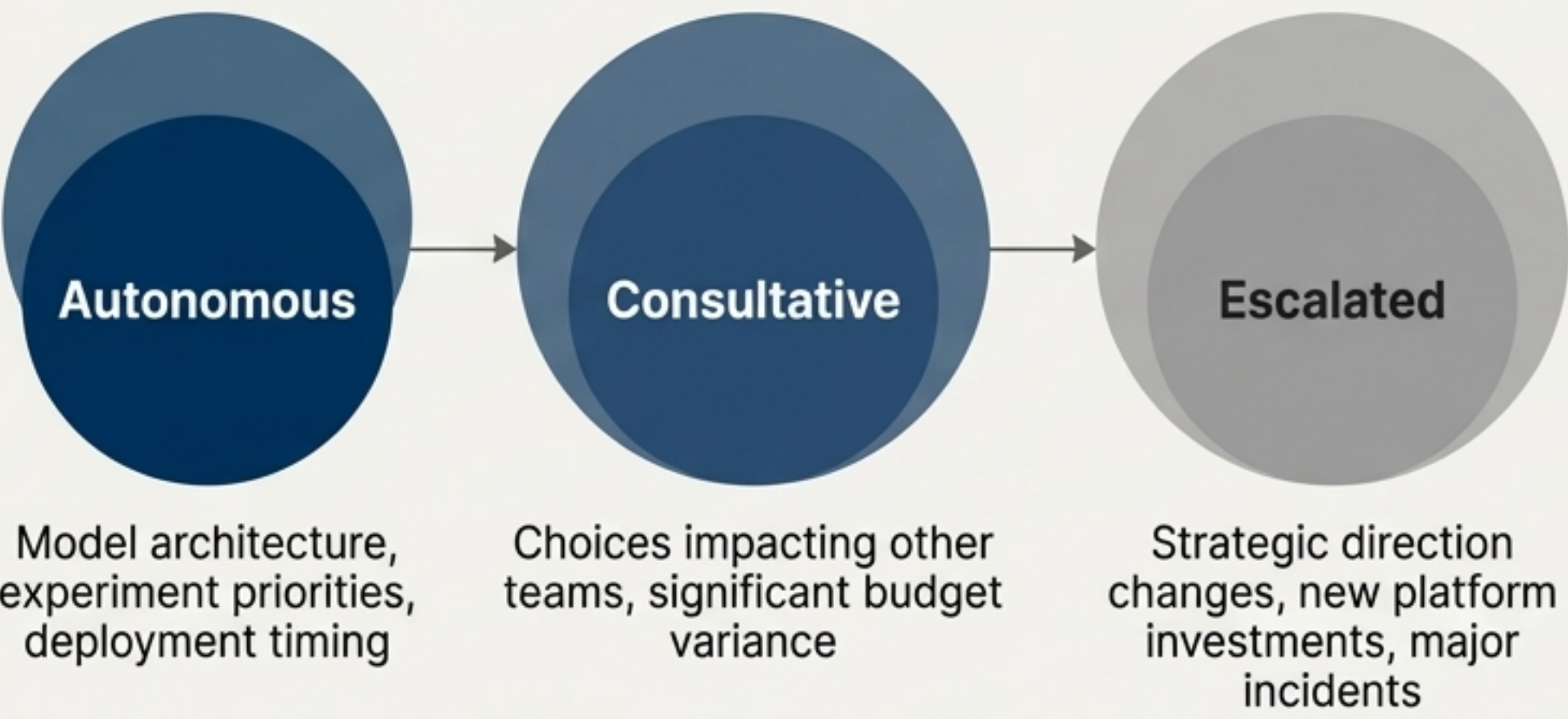
## Problem

Ambiguity about decision rights systematically undermines team performance and creates delays.

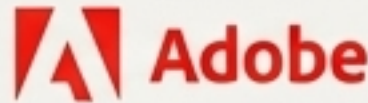
## Solution

**Establish Clear Decision Rights Frameworks.**  
Establish Clear Decision Rights Frameworks.  
Formally define team authority boundaries and escalation paths using written team charters.

## Decision Tiers (Example)



## Proof Point



**Action:**  
Implemented detailed decision rights with defined risk thresholds.

**Result:**  
"Substantially reduced approval cycle time while maintaining appropriate oversight."



# Allocate Scarce Resources Through Transparent Mechanisms

**Problem:** As teams scale, competition for talent, compute, and data intensifies, leading to political battles and inefficiency.

**Solution:** Implement Resource Brokering & Market-Inspired Allocation. Move away from committee-based planning to transparent systems.

## Effective Patterns



Transparent capacity dashboards (for compute, expert availability)



Time-boxed resource allocation (e.g., guaranteed access for defined sprints)



Internal service marketplaces

## Proof Point



**Action:** Manages compute resources via a combination of reserved and spot capacity.

**Result:** Surfaced the highest-value use cases, leading to more efficient allocation than previous committee-based approaches.



# Turn Individual Lessons into Institutional Knowledge

## Problem

Valuable experience and lessons learned remain trapped within individual teams, preventing organizational learning.

## Solution

Design Deliberate Knowledge Circulation Mechanisms. Combine “push” (easy sharing) and “pull” (discoverable repositories) approaches.

## Practices That Scale

- Rotation programs between teams
- Internal tech talks and paper reviews (recorded and archived)
- Shared, searchable incident databases and post-mortems
- “Office hours” with platform and domain experts

## Proof Point



**Action:** Implemented monthly “AI Showcase” sessions where teams present both successes and failures.

**Result:** Ideas from showcases frequently influence the direction of subsequent projects across the organization.



# Evolve from Gate-Based Reviews to Federated Governance

## The Challenge

Traditional, pre-deployment governance is insufficient for AI models that learn and drift over time.

## The New Model: Continuous Compliance

- **Automate Guardrails:**  
Embed compliance, bias, and fairness tests directly into deployment pipelines.
- **Monitor Outcomes, Not Activities:**  
Use real-time dashboards to track model behavior in production and intervene by exception.
- **Tier Governance by Risk:**  
Apply rigorous oversight to high-stakes applications (e.g., credit decisions) and minimal review to low-stakes ones.

JP.Morgan Chase 

### Action:

Implemented a federated AI governance framework with risk tiers and automated tooling.

**Result:** Maintained compliance standards while "significantly reducing average approval cycle time for medium-risk models."



# Shift from Annual Plans to Distributed Strategic Coherence

## The Challenge

Annual planning cycles are too slow for AI's rapid evolution; a strategy set in January may be obsolete by June.

## The New Model: Continuous Strategy

- **Quarterly Strategic Themes:** Set broad organizational priorities (e.g., "improve customer retention") that guide, but don't prescribe, team projects.
- **Focus on Outcomes, Not Outputs:** Teams align to business outcomes (e.g., "reduce churn by 15%") and have freedom to find the best AI solution.
- **Use Teams as Strategic Sensors:** Create channels for upward-flowing intelligence from teams about new opportunities and threats.

**intuit.**

**Action:** Leadership sets quarterly strategic priorities; teams self-organize and propose initiatives to address them.

**Result:** Maintains strategic coherence while preserving team autonomy by operating at different levels of abstraction.



# Treat Architecture and Data as Products with Active Stewardship

## The Challenge

Uncoordinated teams will naturally fragment technical architectures and data practices.

## The New Model: Attract, Don't Mandate



**Evolutionary Architecture:** Platform teams build attractive, reusable components (e.g., feature stores, model serving infrastructure). De facto standards emerge through voluntary adoption.



**Data Mesh Principles:** Treat data as a product owned by the domain team that creates it. That team is responsible for publishing it in a usable, documented, and discoverable form for others.

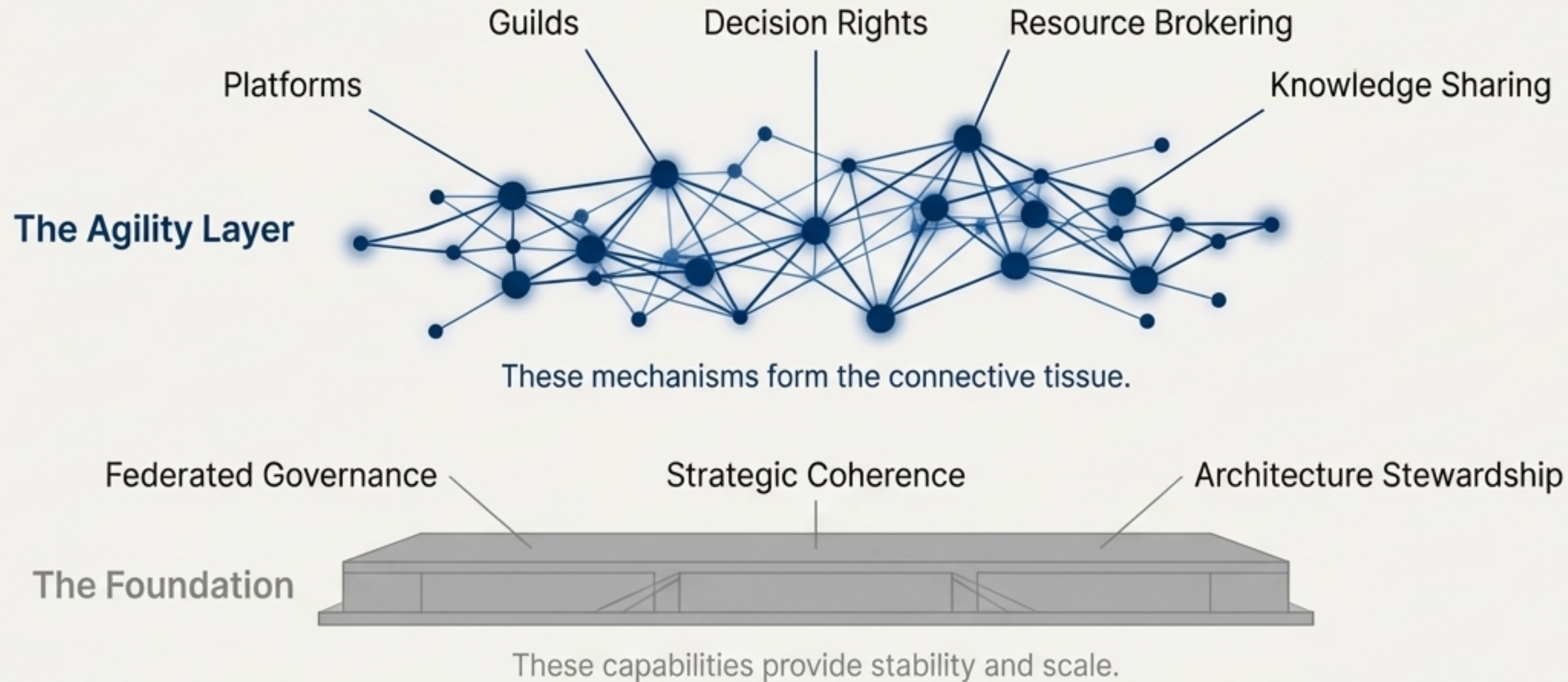


**Action:** The central AI platform team (Einstein) actively evolves its offerings based on feedback from autonomous product teams.

**Result:** High adoption of shared infrastructure through choice, not mandate, creating a virtuous cycle of improvement.



# The Dual Operating System: Agility and Coherence in Action



**Key Takeaway:** Preserve team-level speed and creativity while building the institutional capabilities for enterprise-wide coordination.



# The Leader's Imperatives for the AI Era

Mastering this balance requires a deliberate shift in leadership focus. Four imperatives are critical:

- 1. Invest in Platforms Before Scaling Teams.** The shared infrastructure that enables autonomy without chaos must precede or accompany team proliferation.
- 2. Make Coordination Opt-In and Valuable.** Design guilds, forums, and standards that teams use because they help them move faster, not because they are mandated.
- 3. Evolve Decision Rights as Context Changes.** The boundary between team autonomy and escalation is not static. Review and adjust decision frameworks quarterly.
- 4. Measure Coordination Quality, Not Just Team Velocity.** Track reuse of shared components and duplication rates. An organization can suboptimize while individual teams appear productive.

***“The window for getting this right is measured in quarters, not years.”***