



Behavior Problems

Governance as the superpower for Applied AI

By Demian Entrekin and James Word

Let's face it: the AI industry has spent the last few years putting LLM chat windows into areas where mistakes won't compromise company assets or break mission-critical systems.

But when we consider mission-critical systems, where mistakes can disrupt operations, and leaks can jeopardize the entire business, we expect predictability and reliability. We have zero tolerance for random WTF behavior. Let's take payroll processing systems: we cannot have an electronic funds transfer system changing how it works every time it sends out our paycheck. So where does this leave us in the era of Applied AI?

Applications built using Applied AI present us with a new kind of digital solution. Applied AI systems are probabilistic, which is a fancy way of saying that they create solutions that have a "high likelihood" of being right and useful. But "high likelihood" does not mean safe and reliable. As probabilistic engines, AI applications can produce materially different results with the same inputs, and sometimes those results are disastrous. Traditional systems, however, are designed, built, tested, and deployed to do the same thing over and over. Predictably and reliably. Like payroll.

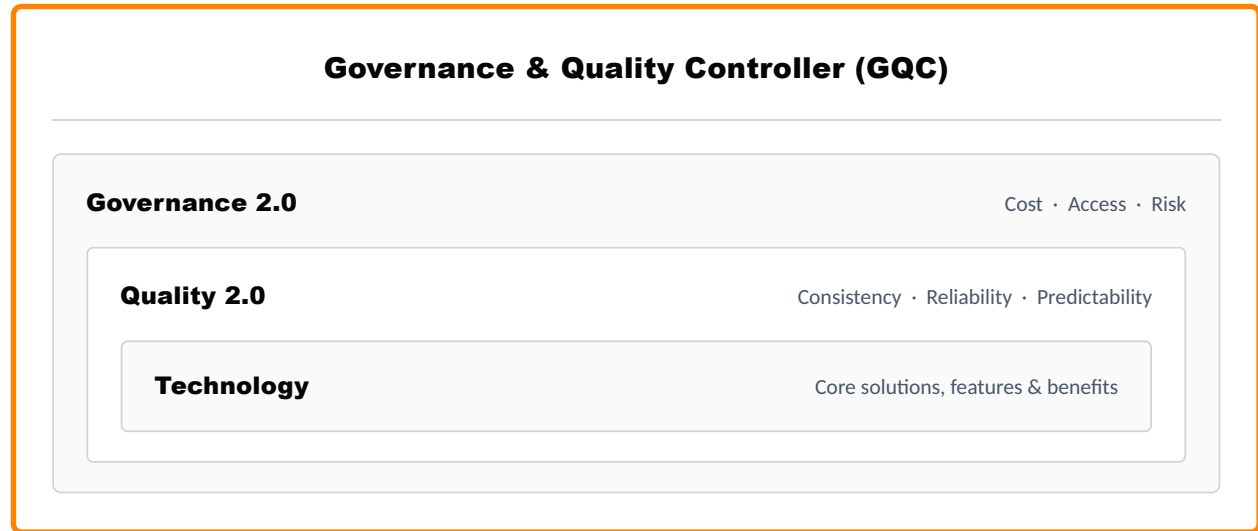
The productive power of generative AI systems is also their greatest weakness because generative AI is not built for predictability and reliability. It is probabilistically creative. Yes, we want fast and inexpensive, but we also demand predictability and reliability. The Applied AI era forces us to reconsider how we create digital solutions: how can we get the massive productivity benefits of Applied AI and protect ourselves against the reliability challenges? What we need is the generative power of Applied AI solutions AND the predictability and reliability of traditional deterministic systems.

To meet this challenge, the market requires a layer of control for AI behavior: the **Governance & Quality Controller (GQC)**. It replaces the manual handoffs between governance, technology, and quality with continuous verification built into the system itself. We call this shift **Governance 2.0**, and it contains within it the boundary of **Quality 2.0**. These two containers hold the whole system to a measurable envelope across every input, process, and impact.

GOVERNANCE 1.0 / MANUAL HANDOFFS



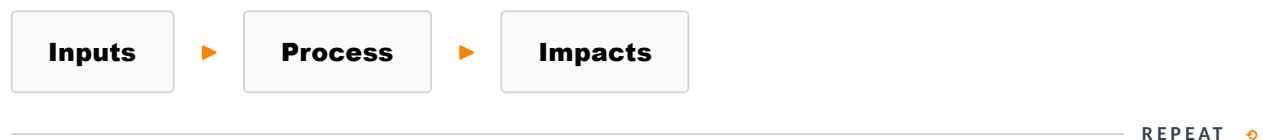
GOVERNANCE 2.0 / BUILT-IN, REAL-TIME



THE NEXT FOCUS GOVERNANCE & QUALITY CONTROLLER: FROM MANUAL HANDOFFS TO A GOVERNANCE LAYER BUILT INTO THE SYSTEM, IN REAL TIME

THE GOVERNANCE 2.0 ERA

The Governance 2.0 era is here, and quality operates within the governance boundary. We must contain Applied AI to achieve predictability and reliability. Governance 2.0 establishes testable boundaries around the system (access controlled, reach contained, cost managed), and it continuously verifies that the system operates inside them.



THE GOVERNANCE LOOP: EVERY CYCLE RE-VERIFIES THE SYSTEM AGAINST ITS BOUNDARIES

It's one thing to use generative AI in a "chat mode," where users develop outputs based on conversations with AI. It's quite another thing to put generative AI models into the center of production systems where we cannot predict the outputs.

When we use AI in a "chat mode," we download the output from AI (e.g. language, code, images) and then we deploy it into deterministic systems where it is once again subject to change control. In "chat mode," AI remains firmly outside of the systems infrastructure. Once the output has been extracted from the "chat," it is no longer subject to the whims of generative AI. In "chat mode" scenarios, the end user remains personally accountable for the stability of these outputs.

But if we are going to use Applied AI to run core enterprise systems, we must be able to control it. Random behavior must be curtailed. Applied AI solutions must behave with predictability and reliability.

As AI systems move into the center of our production stack, we must change the way we think about governance and quality. We can no longer operate in a world where governance, quality, and technology are managed through analog, manual handoffs. Applied AI systems change faster than people can inspect them, and no inspection can certify an individual output of a probabilistic system.

This is how every mature discipline governs stochastic processes: manufacturing never inspected its way to quality; it verified that the process stayed within control limits. Applied AI is the newest stochastic system, and it arrived without its control layer. The governance layer is not a policy gate bolted onto the system. It is the control envelope, the only place where every input, invocation, cost, and output can be observed, managed, and verified.

The quality envelope has three axes. **Consistency:** does it behave tomorrow the way it behaved today? **Reliability:** does the system perform and remain available under heavy use? **Predictability:** does the system deliver the outputs we expect in the real world? A system inside all three boundaries is a governed system. A system outside any one of them is a liability with good demos.

With 17 patents pending, NextFocus has built the Governance & Quality Controller (GQC) to address the challenges of deploying Applied AI solutions in the real world. The GQC functions as the middleware framework that runs as a hosted offering, and it establishes the Governance and the Quality controls that allow organizations to deploy true AI solutions in a governed framework.

What the GQC governs

The sections that follow map the GQC to the three layers of problems and solutions: Governance 2.0, Quality 2.0, and Technology. Each layer is described as the problems organizations actually hit and the solution the GQC puts in place.

GOVERNANCE 2.0

Safety & security

PROBLEM	SOLUTION
How can we keep users in their business domains?	Domain controls with least-privilege scoping.
How can we keep agents in their business domains?	Agents come with zero permissions by default and are granted rights explicitly. They are principals with permissions, not free actors.
Corporate data leaks	Opt-in / opt-out and Zero Data Retention (ZDR) safety for training data and data-use policy enforced at the control plane, per provider and per request.
Inappropriate user or agent access to LLM models, data and search	LLM access, data access, compliance & encryption

Cost control

PROBLEM	SOLUTION
Managing token costs	Quotas enforced in real time per user, per agent, per use case. Cost is a governed boundary, not a monthly surprise.
Expensive LLM lock-in	Maintaining performance and reliability while taking advantage of cheaper models through model swapping
Lack of dynamic model routing	Route each task to the least-cost model that satisfies its behavioral contract.

Compliance

PROBLEM	SOLUTION
Public standards (HIPAA, SOC2, GDPR)	Controls mapped to the governance policies and evidenced continuously from control-plane telemetry; audit is native and real-time rather than a project.
Private standards	Internal policy expressed as code and enforced at the same governance policies. Examples might be corporate lexicon or branding guidelines and styles.
IT security standards	Encryption algorithms, boundaries and zones. Programmatically tracking agent behavior within desired trust levels.
Scope level boundaries & controls	Ability to classify a scope as public, internal, confidential in order to ensure that behavior can be dialed down to a "need to know" basis.

QUALITY 2.0

Information inputs

How can we make sure the inputs are in excellent condition?

PROBLEM	SOLUTION
Source data ingestion	Built-in, secure integration with external systems such as API, Data Lakes, etc.
Data quality audit	Validate inputs at the boundary before they reach the model; quarantine and trace what fails
Institutional knowledge	Capture the unique "how to" of the business in order to guide the behavior of the systems

Systems process

How can we provide reliability and completeness in the process?

PROBLEM	SOLUTION
Reliable systems	Continuous envelope telemetry: every invocation observed, attributed, and scored against the value, cost, and consistency boundaries.
Solution completeness	Rapid iterations without drift: every change ships as a governed release, tested against the behavioral baseline before promotion.
Total cost of ownership	Cost is instrumented natively at the control plane (per token, per invocation, per use case) making TCO a measured fact, not an estimate.

Business impacts

How can we make sure the system impacts achieve our goals?

PROBLEM	SOLUTION
Demonstrable ROI	By providing a predictable TCO, we can provide the finance team with the ability to prove the ROI of the system investment.
Reliable systems	Separate reliability (the system stays up) from consistency (the system stays the same). The first is an SRE problem. The second is a governance problem (see System Consistency below).

System consistency

What you cannot have is an operational system that keeps changing because the LLM keeps changing it. That system will fail. End users build procedures, training, and trust around how a system behaves. When behavior mutates underneath them (a provider model update, sampling variance, an agent revising its own workflow) trust collapses even when the change is technically an improvement. From the end user's standpoint, consistency is not adjacent to quality. It is a component of quality.

The failure mode is not "change." It is "undeclared change." Governance 2.0 converts model evolution from continuous drift into discrete, verified releases.

PROBLEM	SOLUTION
LLM releases new behavior	Version pinning at the model and prompt level; behavioral regression against golden datasets; staged promotion with rollback.
LLM "hallucinations" cause reliability challenges	Reposition the LLM to drive deterministic mode rather than generative modes, and add layers of validation that binds systems behavior to validated results.
Same input, different output	The current crop of generative LLMs produce different results with the same inputs; the system's behavior must be reliable.
Agents modify their own tools and workflows	Self-modification is a change request, not a runtime event: sandboxed, tested, and gated like any release.

Consistency cannot be guaranteed by LLM providers. They are the source of the change, and their incentive is to ship it. Only a layer standing between them can pin versions, detect drift, and hold the behavioral contract. This is the capability that makes the governance architecture permanent rather than transitional.

Authentication & authorization

PROBLEM	SOLUTION
User authentication	User identities can authenticate, hold scoped credentials, and are audited in real time.
Agent authentication	Agent identities can authenticate, hold scoped credentials, and are audited in real time.
User authorization	Controls in place for what an end user can access and with what privileges once they've been authenticated.
Agents authorization	Controls in place for what an agent can access and with what privileges once they've been authenticated.

External systems integration

PROBLEM	SOLUTION
System-to-system CRUD	All reads and writes transit the governed seam, attributed and reversible.
Domain controls	Each integration is bound to its domain; cross-domain action is an explicit, logged grant.
Data ingestion management	Management of items within the ingestion pipeline for Atomicity, Consistency, Isolation, and Durability (ACID) for transaction level management.
Data integration	Data ingested as native data with full text indexing and graph relationship mapping. Conflict resolution and strategies give you controls to ensure that the results map to the inbound and outbound requirements.
Semantic labeling	Control semantic identification of the source data for consistency within the Applied AI framework.

WHY GOVERNANCE 2.0 TURNS APPLIED AI INTO SUPERPOWER

Governance 2.0 manages Applied AI systems through a risk management framework. Controlling risk is a priceable task: a failure rate, an impact radius, a cost ceiling that a risk officer can sign. Every use case that is too risky to deploy without a governance envelope is a market available only to those who have one. That is not compliance. That is an advantage.

The Applied AI era forces us to build Governance and Quality into the system itself. Governance 2.0 will unlock a new level of predictability and value creation by creating a layer of quality control that was never achievable in the traditional coding and QA business model.

CONTACT

Demian Entrekin

demian.entrekin@nextfocus.ai