Development of an Artificial Intelligence (AI) Action Plan

Optimizing AI Development - Eliminating Ambiguity that Leads to Lack of Innovation

Author:
Tim Klawa
Head of Product F8F

1 The Need for a Quantitative AI Action Plan

Artificial intelligence is poised to redefine decision-making, automate complex tasks, and expand the capabilities of government agencies. Yet, for all its potential, AI development remains plagued by inefficiencies that slow progress, increase costs, and underperform for government programs. Too often, AI programs are launched with broad ambitions, but without a structured, quantitative framework to measure success in individual components of AI development. Investments in data collection, model training, and evaluation are made without a clear understanding of how each contributes to the final system's effectiveness and the impacts changes to one element of the AI lifecycle has on another. The result is a cycle of uncertainty, where agencies continue refining AI based on best guesses without ever knowing the true impact that a particular lever has in driving AI innovation and ultimately AI superiority.

The core challenge is not a lack of funding, but it's the absence of a rigorous approach to assessing AI performance at every stage of development. AI innovation is not reliant on a single aspect or core element, but on a system composed of interdependent parts. Data must be carefully curated, labels must be applied consistently with rigorous quality controls against particular mission needs, models must learn the right patterns, and test and evaluation must expose failures early enough to correct them. If any of these elements are misaligned, the entire system suffers. Yet, many AI programs operate components of the AI lifecycle in rigid, bureaucratic silos, treating each component as an isolated task rather than part of a continuous interconnected feedback loop. This fragmented approach makes it difficult to diagnose problems when they arise and harder to determine whether resources are being used effectively. To compound the issue, a single vendor often controls critical steps in the AI pipeline. This structure makes rapid pivots to accelerate innovation challenging and frequently leads to bottlenecks that hinder AI progress. Additionally, the lack of quantitative metrics and a clear understanding of their interrelationships within the AI pipeline creates ambiguity. As a result, decisions about steering AI development are often based on best guesses and hunches rather than a comprehensive operational picture, making it difficult to allocate funding effectively to maximize AI innovation.

Decisions about AI development must be grounded in measurable impact, not in assumptions. Without structured performance metrics, program managers are often left guessing whether they should invest in more data, improve annotation quality processes, change model architectures, or refine evaluation methods. Some AI failures stem from insufficient training data, while others result from poor label quality or rigid ontologies that fail to capture the complexity of real-world operational scenarios. Treating AI failures as singular events rather than the result of systemic weaknesses in individual elements of an

AI program obscures the true cause of underperformance. A mislabeled dataset might degrade model accuracy just as much as a flawed algorithm, but if agencies lack the tools to measure these effects, they will struggle to implement meaningful solutions.

The key to breaking this cycle and unleashing AI innovation for federal agencies is to embed measurement at every stage of the AI lifecycle and drive for competition across vendors in each stage. Programs should not wait until operational deployment of an AI system to determine whether a system is effective. At that point, millions or billions of dollars may have been wasted in a potentially misdirected effort. Instead, it's critical to establish clear evaluation criteria early and track performance continuously allowing program managers within government to efficiently allocate funding to the levers that will drive maximum optimization for a particular model. This requires a shift from intuition-driven decision-making to a structured, data-driven approach where every investment is assessed for its contribution to the final outcome. If agencies can define what success looks like in quantifiable terms, whether in terms of data quality, annotation consistency, model generalization, or real-world adaptability—they can build AI systems that improve over time rather than stagnate under the weight of inefficiencies.

2 The AI Lifecycle: Where Current Approaches Fail

Artificial intelligence is not a singular achievement but a process—one that must be continuously refined, measured, and adapted. Every AI system is shaped by the quality of its data, the precision of its labels, the effectiveness of its training, and the rigor of its evaluation. These components are deeply interconnected, yet many AI programs treat them as separate, isolated tasks. This fragmented approach makes it difficult to trace the root causes of failure, leading to inefficiencies that persist through multiple development cycles. A model that underperforms is often assumed to need more data, but if that data is inconsistently labeled or misaligned with real-world conditions, more of it will not necessarily solve the problem. Similarly, a well-labeled dataset may still produce unreliable outputs if the model is trained on outdated assumptions or if test and evaluation occur too late to course-correct. Without a structured framework that ties each phase of development to measurable performance indicators, agencies are left diagnosing problems after they have already caused cascading failures.

The absence of a unified AI lifecycle framework is one of the biggest obstacles to scalability. In an ideal development process, each phase of AI training should reinforce the next, creating a system that continuously improves through iteration. Instead, most programs operate in a way that makes iteration slow and expensive. Data is collected with limited foresight into how it will be labeled. Annotation teams define categories based on

static ontologies that fail to capture operational nuances. Model developers train AI systems to optimize test conditions rather than real-world unpredictability. Finally, evaluation is treated as a checkpoint rather than a continuous process, revealing issues only when they have become deeply embedded in the system. This linear workflow forces teams into a reactive mode, where failures are discovered too late to be fixed without significant retraining. The result is an AI lifecycle that does not adapt, does not evolve, and does not scale efficiently.

The consequences of this misalignment are often misunderstood. AI failures do not happen at a single point; they accumulate over time. A mislabeled dataset does not immediately break a model, but it introduces a subtle distortion in how the system interprets information. A rigid classification schema does not cause catastrophic failure overnight, but it limits the AI's ability to generalize when faced with edge cases. These problems remain hidden until they surface in operational environments, at which point they are no longer minor inefficiencies—they're fundamental weaknesses. These problems remain hidden until they surface in operational environments, at which point they are no longer minor inefficiencies—they're fundamental weaknesses. AI programs that fail in the field often do so not because of a single, obvious flaw, but because of a long history of small, compounding errors that were never detected or addressed.

A structured AI lifecycle must be designed to prevent these cascading failures by embedding continuous evaluation and refinement at every stage. Data sourcing should not be treated as a one-time acquisition process but as an evolving effort to ensure representativeness and diversity. Labeling should not be based on rigid taxonomies that assume perfect clarity but should incorporate validation mechanisms that account for uncertainty. Model training should not optimize performance on a static dataset but should be structured to detect and adapt to changes in data distributions. Test and evaluation should not be an afterthought but a real-time feedback loop that informs earlier stages of development.

For AI to scale effectively, agencies must move beyond the traditional model of linear development and instead adopt a lifecycle approach that prioritizes adaptability. This means shifting from static assessments to dynamic performance tracking, where data quality, annotation accuracy, and model effectiveness are continuously monitored. It requires building AI systems that are not just measured at deployment but throughout their operational lifespan, ensuring that any degradation in performance is detected and addressed before it becomes a mission-critical failure.

The challenge is not simply in fixing what is broken but in restructuring AI development so that inefficiencies do not become embedded in the first place. By recognizing that AI is not

a single-step process but an evolving system, agencies can move beyond reactive problem-solving and toward a more proactive, scalable approach. The goal is not just to create AI that works but to create AI that continues to work—adapting, improving, and delivering value long after the initial deployment.

3 Moving from Reactive to Proactive AI Risk Mitigation

Artificial intelligence is often treated as an innovation problem when, in reality, it is a risk management challenge. The most pressing issues in AI development are not about whether a model can be built but whether it will perform as expected when deployed. Failures do not always take the form of outright system crashes or misclassifications; more often, they manifest as subtle performance degradation, hidden biases, or a gradual misalignment between AI outputs and mission objectives. The problem is not simply that these risks exist—every complex system carries inherent risks—but that many AI programs lack the mechanisms to detect and correct them before they accumulate into full-scale failures. Instead of embedding proactive safeguards throughout the AI lifecycle, risk mitigation is often treated as a reactive exercise, with failures only becoming apparent after they have already impacted operations.

The traditional approach to AI risk management assumes that if a model passes a final evaluation phase, it is ready for deployment. This mindset is borrowed from conventional software development, where functionality can often be verified through discrete testing stages. But AI does not function like traditional software; it is probabilistic rather than deterministic, meaning that its behavior is influenced not just by its initial design but by the data it encounters over time. A model that performs well on a controlled test set can still break down when exposed to real-world variability. A training dataset that seems representative at one point in time may become outdated as conditions evolve. The challenge is that many AI programs treat risk as a box to check at the end of development rather than a continuous process that requires ongoing monitoring and adjustment.

This reactive approach creates a dangerous blind spot. When risks are only assessed at the point of deployment, agencies have little visibility into how failures emerge over time. A model that begins drifting away from its intended function may not trigger an immediate failure, but over weeks or months, its predictions can become unreliable. If an AI system is misclassifying objects at a rate of five percent, that may seem like an acceptable error margin—until it is applied at scale, where thousands of decisions are made based on those incorrect classifications. The absence of continuous monitoring means that by the time these issues are recognized, they are no longer isolated problems but systemic weaknesses that require costly rework.

A more effective approach to AI risk mitigation is one that shifts from post-deployment troubleshooting to early detection and real-time course correction. Instead of treating evaluation as a static phase, agencies must integrate dynamic risk tracking mechanisms that assess AI performance as models are trained, refined, and deployed. This requires embedding automated monitoring tools that detect anomalies in decision patterns, track shifts in data distributions, and flag inconsistencies in model predictions. It also means rethinking how AI failures are diagnosed—not as singular events but as signals of deeper structural weaknesses in the development pipeline.

One of the most overlooked aspects of AI risk is the impact of feedback loops between different components of the system. AI models are only as good as the data they are trained on—— if the data pipeline itself is flawed, then errors will propagate throughout the entire lifecycle. A mislabeled dataset does not just degrade model performance; it skews evaluation metrics, leading teams to believe that their models are performing better than they actually are. A misaligned annotation schema does not just affect one training iteration; it establishes a precedent that carries forward into future models, reinforcing a distorted understanding of the problem space. Without a system for detecting and addressing these misalignments early, AI programs risk optimizing flawed assumptions rather than real-world effectiveness.

To transition to a proactive risk mitigation strategy, agencies must embrace an iterative approach to AI oversight. This means designing AI systems with built-in fail-safes, where performance is continuously assessed against evolving benchmarks rather than static accuracy thresholds. It also means integrating real-world validation into AI pipelines, where models are not just tested on controlled datasets but are regularly exposed to new conditions to ensure adaptability. Agencies must establish clear accountability mechanisms, where performance degradation is not treated as an inevitable byproduct of AI deployment but as a preventable outcome that can be corrected before it leads to mission failure.

The goal of AI risk mitigation is not to eliminate uncertainty—AI, by its nature, will always carry some level of unpredictability. The goal is to ensure that uncertainty does not become unmanageable. By shifting from reactive crisis management to a proactive, structured approach to risk assessment, agencies can build AI systems that are not only more resilient but also more transparent, auditable, and aligned with real-world mission needs. Risk is not the enemy of AI innovation—it is the price of admission. The agencies that learn to manage it effectively will be the ones that achieve AI acceleration without compromising reliability.

4 Measuring AI Performance: Creating a Data-Driven Framework

AI is only as effective as the system that governs its development. Without clear, quantifiable measures of success, AI programs risk becoming expensive experiments with no way to determine whether they are progressing or merely consuming resources. A model that produces seemingly accurate predictions may still be failing in ways that are not immediately visible. A dataset that appears comprehensive may still contain hidden biases or gaps that distort outcomes. The absence of structured performance tracking leaves agencies navigating AI development with intuition rather than empirical validation. To break this cycle, AI programs must be built around a data-driven framework that continuously measures system effectiveness at every stage of development.

Defining success in AI is more complex than simply evaluating accuracy. While accuracy is often treated as the primary metric, it is only one dimension of AI performance. A high-accuracy model can still be unreliable if it lacks robustness, misclassifies edge cases, or fails to generalize beyond its training environment. AI systems must be evaluated across multiple performance indicators, including how well they adapt to changing conditions, how consistently they produce reliable outputs, and how effectively they integrate new information without degradation. These factors cannot be assessed through a single evaluation checkpoint at the end of development but must instead be measured continuously, ensuring that performance is not just optimized for one dataset but persists across real-world scenarios.

Data quality is the foundation of AI performance, yet it is often assessed in superficial terms. Many programs focus on dataset size rather than representativeness, assuming that more data will always improve model outcomes. But if the data does not accurately reflect the conditions under which the model will operate, its size is irrelevant. A dataset can be large but still lack coverage of key edge cases. It can contain thousands of labeled examples but still be biased toward certain patterns while neglecting others. A truly effective AI framework must measure not just how much data is collected but how well it aligns with the mission-specific requirements of the system. This means establishing structured assessments that track data distribution, label consistency, and the presence of rare but operationally critical cases that the AI must learn to recognize.

Annotation quality is another crucial, yet frequently overlooked factor in AI performance. Poorly labeled data does not just introduce random noise, it actively misleads the model, teaching it incorrect associations that persist throughout training. Many AI failures are not due to a lack of data but to inconsistent labeling that creates ambiguity in how the system interprets its inputs. If annotation guidelines are vague or vary between labelers, models trained on that data will inherit those inconsistencies, leading to unpredictable results.

Measuring annotation quality requires more than just checking for errors; it demands a structured review process that tracks inter-labeler agreement, detects areas of cognitive uncertainty, and applies automated consistency checks to flag potential misclassifications before they compromise model learning.

Model performance itself must be evaluated beyond standard accuracy benchmarks. Traditional AI assessments often focus on how well a model performs on a static test set, but this provides an incomplete picture. A model that performs well in controlled conditions may still fail under real-world variability, adversarial inputs, or domain shifts that occur when new data distributions emerge. AI performance should be assessed not just by how well a model classifies known examples but by how well it handles uncertainty, adapts to novel conditions, and maintains robustness across different operating environments. This requires metrics that track model drift, decision confidence, and error patterns over time, allowing for early detection of performance degradation before it impacts mission-critical decisions.

One of the biggest limitations in AI measurement today is the delayed integration of test and evaluation. Too often, AI systems are validated only after they have already gone through extensive training cycles, meaning that misalignments between model assumptions and real-world conditions are not discovered until late in development. By that point, correcting those issues requires expensive retraining or, in the worst cases, an entirely new dataset. A more effective approach is to embed evaluation throughout the AI pipeline, conducting iterative assessments at multiple stages rather than waiting until the model is near completion. Early testing allows teams to identify weaknesses in the data, detect annotation inconsistencies, and refine model architectures before errors become deeply embedded in the system.

For AI programs to be successful, agencies must transition from one-time validation toward continuous performance monitoring. AI systems do not remain static after deployment; they interact with new data, encounter unexpected conditions, and require ongoing updates to maintain effectiveness. A model that performs well today may degrade over time if new operational variables are introduced that were not present in the training data. Agencies need structured monitoring systems that track AI performance in real-world applications, detecting when accuracy begins to slip, when classification errors become more frequent, or when decision patterns shift in unintended ways. This ensures that AI remains reliable not just at the moment of deployment but throughout its entire lifecycle.

A well-designed AI measurement framework does not just track success, it enables continuous improvement. By integrating real-time assessments of data quality, annotation consistency, model adaptability, and long-term performance trends, agencies can move

beyond reactive troubleshooting and instead adopt a proactive approach to AI optimization. AI is not a static product; it is a system that must be monitored, refined, and adjusted as conditions evolve. Programs that fail to embed these measurement principles will struggle with inefficiencies, while those that embrace them will build AI systems that are not only effective today but continuously improving for the challenges of tomorrow.

5 Breaking the Cycle of AI Stagnation

AI programs often fail to reach their full potential not because of a lack of ambition or investment but because they become trapped in cycles of stagnation. These cycles emerge when development efforts focus on expanding rather than improving, where additional data, more complex models, and increased computational power are treated as the primary solutions to underperformance. Without structured assessments that measure the effectiveness of these changes, agencies risk scaling inefficiencies rather than addressing the underlying limitations of their AI systems. The result is an AI system that grows in size but not in capability, consuming more resources without delivering proportionate improvements.

One of the primary drivers of stagnation is the assumption that more data inherently leads to better AI. While data is the foundation of any machine learning system, its quality and representativeness matter far more than its volume. Collecting more of the same type of data does not solve the problem if the model's failures stem from gaps in coverage or labeling inconsistencies. A system trained on an imbalanced dataset will not suddenly generalize better because additional examples reinforce the same biases. Programs that rely on indiscriminate data accumulation without evaluating whether it is meaningfully expanding the model's understanding risk spending years increasing dataset size without improving performance in any operationally relevant way.

The same flawed logic applies to AI model complexity. There is a persistent belief that larger architectures—deeper neural networks, increased parameter counts, and higher computational loads—automatically translate to better decision-making. While larger models can capture more nuanced patterns, they are also more prone to overfitting, require exponentially greater resources, and often provide diminishing returns if the training data does not support the increased complexity. Simply making a model bigger does not make it smarter. If its training data lacks diversity, its labels are inconsistent, or its evaluation process is flawed, scaling the architecture only amplifies these weaknesses rather than fixing them.

Beyond technical stagnation, many AI programs struggle with structural inertia—the inability to pivot when early investments prove ineffective. Once significant time and

funding have been committed to a particular dataset, labeling methodology, or training approach, there is often resistance to change, even when evidence suggests that a different strategy would yield better results. This reluctance is reinforced by the absence of quantitative decision frameworks that clearly show when continued refinement is producing diminishing returns. Without structured performance tracking, program managers are left second-guessing whether a pivot is necessary or whether they simply need to push forward with existing methods. AI systems that are designed to adapt must be matched by program structures that allow for adaptability at the decision-making level.

Another factor contributing to AI stagnation is the reliance on static ontologies and rigid classification schemas that fail to evolve alongside real-world conditions. Many AI systems are trained with predefined categories and assumptions that reflect a specific moment in time rather than an adaptable understanding of their domain. This rigidity is particularly problematic in high-stakes environments where the nature of the problem itself shifts over time. A model trained to detect specific object types in satellite imagery, for instance, may perform well initially but degrade as adversaries change their tactics, modify camouflage patterns, or introduce new operational variables that were not present in the training data. If AI programs do not have mechanisms to update, refine, and reassess their ontologies as new intelligence emerges, they will struggle to remain relevant.

To break out of these cycles, AI programs must transition from linear, assumption-driven development to iterative, evidence-based refinement. This means moving away from bigger models, larger datasets, and longer contracts as default solutions and instead focusing on targeted improvements that directly impact mission effectiveness. Rather than blindly expanding datasets, agencies should evaluate which specific data gaps are limiting AI performance and fill those gaps with strategically sourced information. Instead of continuously retraining models under the same static assumptions, programs should implement adaptive learning mechanisms that allow AI systems to refine their understanding over time. And rather than treating AI as a one-time development effort, agencies must embrace continuous validation and refinement, ensuring that models are not just deployed but are actively maintained as living systems that evolve alongside operational needs.

Breaking AI stagnation requires more than technical adjustments—it demands a fundamental shift in how AI programs are structured, measured, and managed. Programs must be built with intentional flexibility, where changes are not seen as setbacks but as necessary adaptations to new information. Success should not be defined by the number of models built, the volume of data collected, or the complexity of architecture deployed, but by the measurable impact AI has on decision-making and operational efficiency.

6 AI Procurement Reform: Overcoming Barriers to Innovation

Artificial intelligence is not just a technological challenge; it is a systems challenge, one in which policy and procurement structures play a defining role. No matter how advanced a model or well-curated a dataset may be, the way AI is acquired, funded, and managed can determine whether a program thrives or stagnates. Many AI initiatives fail not because of technical limitations but because their development is constrained by rigid contracting structures that create artificial barriers between the components of the AI pipeline. When procurement strategies do not align with the iterative nature of AI, they can slow innovation, limit competition, and restrict access to emerging breakthroughs.

To accelerate AI progress, we strongly recommend that agencies prioritize multi-award, shorter-duration contracts that foster continuous competition and innovation. AI is a rapidly evolving field where new models, methodologies, and optimization techniques emerge frequently. Long-term, single-award contracts can inadvertently hinder progress by locking agencies into a single vendor for years, reducing their ability to integrate new advancements. Multi-award contracts allow multiple vendors to contribute in parallel, ensuring agencies have access to the latest innovations and fostering a competitive environment where vendors are incentivized to improve performance over time.

Exclusive, long-term contracts often lead to vendor lock-in, where critical AI components—such as model architectures or optimization techniques—become proprietary and difficult to transition away from. This lack of interoperability can create significant barriers to integrating better solutions as they emerge. By contrast, shorter contracts with multiple providers encourage modular development, ensuring that agencies maintain flexibility and control over their AI systems.

Competition is also essential for driving efficiency and quality improvements in every component of AI. When multiple vendors compete, they are motivated to enhance speed, reduce costs, and refine their contributions. A sole-source contract, however, eliminates this competitive pressure, making it more likely that innovation will stagnate. Agencies need the ability to continuously evaluate and select the best available solutions, rather than being constrained by outdated methodologies dictated by a single vendor.

In such instances where an agency decides to pursue a single vendor strategy, we strongly recommend implementing safeguards to ensure continued innovation and transparency. These contracts should include clear, performance-driven metrics that allow agencies to assess a vendor's ability to push the boundaries of AI advancements. Additionally, structured incentives should be built in order to encourage ongoing improvements, rather than allowing the vendor to meet only the minimum contractual requirements and not

focus on continued innovation. Transparency in these agreements is critical—agencies must have visibility into the vendor's methodologies, benchmarks, and performance data to ensure accountability and to make informed decisions about future procurement strategies.

Ultimately, procurement reform is not just an administrative consideration; it is a critical enabler of AI success. Without adaptable, performance-driven contracting models, agencies risk falling behind in a field where technological advancements happen rapidly. By prioritizing flexible, competitive procurement structures, agencies can create AI programs that not only meet today's needs but remain capable of evolving with the innovations of tomorrow.

7 A Roadmap for Efficiency and Accountability

AI will not reach its full potential through abstract strategy documents or broad commitments to innovation. It requires a structured, repeatable approach that ensures every investment, every dataset, and every model improvement contributes to a measurable outcome. Without a clear framework to assess efficiency, agencies risk pouring resources into AI programs that grow in complexity but fail to deliver real impact. The challenge is not simply deploying AI but maintaining its effectiveness over time, adapting to new conditions, refining performance, and ensuring that every iteration improves upon the last. This demands an AI action plan built around measurable impact, continuous refinement, and clear accountability.

The first step in building this roadmap is defining performance baselines across the AI lifecycle. Without an objective starting point, there is no way to determine whether changes to data sourcing, annotation quality, or model training are actually improving system performance or just increasing complexity. AI programs must begin with structured benchmarks that assess data diversity, annotation consistency, and initial model behavior before optimization efforts begin. This baseline evaluation serves as a control, allowing program managers to track whether specific interventions lead to tangible improvements or whether they introduce new inefficiencies that must be addressed.

Establishing a dynamic risk registry is equally critical. AI failures are rarely isolated events; they emerge from compounding weaknesses in the development pipeline. If agencies only assess risk at deployment, they miss opportunities to detect early signs of misalignment between the model and its intended operational environment. A structured risk tracking system should identify not just immediate performance concerns—such as unexpected drops in accuracy—but also longer-term vulnerabilities, like data drift, annotation inconsistencies, or an overreliance on static ontologies. By continuously updating this risk

registry, agencies can move from reactive problem-solving to proactive risk mitigation, preventing small issues from escalating into system-wide failures.

Embedding adaptive quality control into data annotation and model training is another fundamental pillar of an effective AI action plan. Many AI failures originate not from poor model design but from inconsistencies in the training data itself. If annotation rules are loosely defined or change over time without structured oversight, models will learn conflicting patterns that degrade performance. Instead of treating data labeling as a static process, agencies must implement automated and human-in-the-loop validation mechanisms that track label consistency and flag areas of uncertainty. The same principle applies to model training, where incremental improvements should be evaluated not just on accuracy metrics but on their ability to generalize to new conditions. AI action plans must prioritize continuous quality assessments, ensuring that every stage of development is aligned with real-world performance expectations.

Test and evaluation must be fully integrated into the AI pipeline rather than treated as an endpoint. Traditional AI programs often separate T&E from the development process, leading to situations where weaknesses are only identified after significant time and resources have already been invested. By embedding iterative testing cycles throughout model training, agencies can detect misalignments early and refine models before performance issues become embedded in the system. This approach reduces the likelihood of last-minute failures that require expensive retraining and ensures that AI systems remain adaptable as mission requirements evolve.

Accountability must be embedded at every level of AI program execution. Many AI initiatives struggle not because of technical limitations but because there is no structured way to measure progress and enforce responsibility for results. Success must be defined not by contract fulfillment or model completion but by demonstrated improvements in AI performance over time. Vendors, internal teams, and program managers must all be held accountable for meeting clearly defined performance metrics, ensuring that AI systems are continuously improving rather than stagnating.

A well-designed AI action plan does not just optimize AI development—it transforms how agencies build, manage, and sustain AI capabilities over time. Programs that adopt a structured approach to measurement, refinement, and accountability will be positioned to scale AI effectively while avoiding the inefficiencies that plague traditional development cycles. The goal is not just to deploy AI but to ensure that every iteration is smarter, more efficient, and more aligned with mission needs than the last.

8 The Future of AI Program Management

Artificial intelligence will not succeed in government programs simply because the technology exists. It will succeed when agencies manage AI as an evolving capability rather than a one-time deployment. Many AI programs today operate under outdated models——developing a system, testing it once, and assuming it will function indefinitely. But AI does not behave like traditional software. It learns from data, reacts to changing conditions, and, if left unchecked, can drift away from its intended purpose. Managing AI effectively requires a shift from static oversight to continuous performance tracking, ensuring that AI systems remain relevant, reliable, and aligned with operational goals long after their initial deployment.

For AI to remain effective, it must be managed as a system in motion. A model that performs well today may degrade tomorrow if new variables emerge that were not present in its training data. An annotation schema that worked for an initial dataset may need adjustments as real-world conditions evolve. If agencies treat AI as something that can be "finished" rather than something that must be maintained and refined, they will find themselves repeatedly retraining models without fully understanding why performance declines. The future of AI program management must be centered on adaptability—ensuring that AI systems are not only measured at deployment but continuously evaluated for shifts in accuracy, bias, and decision confidence.

This shift requires breaking away from rigid performance benchmarks that treat AI evaluation as a one-time event. Traditional government programs often define success through initial model accuracy, but a single metric at a single point in time tells only part of the story. An AI system that achieves high accuracy on controlled test data may still fail when exposed to unexpected conditions. A model that works well in one region may struggle when deployed in another with different environmental factors. Instead of relying on static success criteria, agencies must implement dynamic evaluation frameworks that assess AI across different operational contexts, identifying when and where performance begins to drift so that corrections can be made before failures accumulate.

Managing AI effectively also means ensuring that agencies retain control over their systems rather than becoming overly dependent on external vendors. Many AI initiatives are structured in a way that places critical components—data pipelines, labeling standards, model architectures—under the control of contractors, leaving agencies with limited visibility into how their own systems function. This lack of transparency makes it difficult to diagnose failures, introduce improvements, or transition to new approaches when better technologies emerge. The future of AI program management must focus on retaining

institutional knowledge, ensuring that government teams understand not just how to procure AI but how to oversee, refine, and sustain it independently of any single vendor.

Just as AI itself is built on feedback loops, AI program management must incorporate continuous feedback into decision-making. Agencies must move beyond periodic AI assessments and instead develop real-time monitoring capabilities that track system performance under actual operational conditions. This means integrating automated checks for model drift, continuously evaluating how AI decisions align with mission objectives, and ensuring that AI remains a tool for augmentation rather than a black box that operates outside human oversight. AI is only as effective as the systems put in place to manage it, and if agencies fail to implement structured monitoring, they will be unable to detect when their AI begins to diverge from expectations.

A well-managed AI program is not one that simply produces models but one that ensures those models continue to deliver value over time. The agencies that succeed in AI will be those that recognize that measurement is not a final step but a continuous requirement. They will be the ones that prioritize adaptability over rigid benchmarks, maintain visibility over their own AI systems, and build management structures that allow AI to evolve in alignment with real-world challenges. The future of AI program management is not about keeping up with technological advancements, it is about creating systems that are built to adapt, ensuring that AI remains an asset rather than a liability as conditions change.