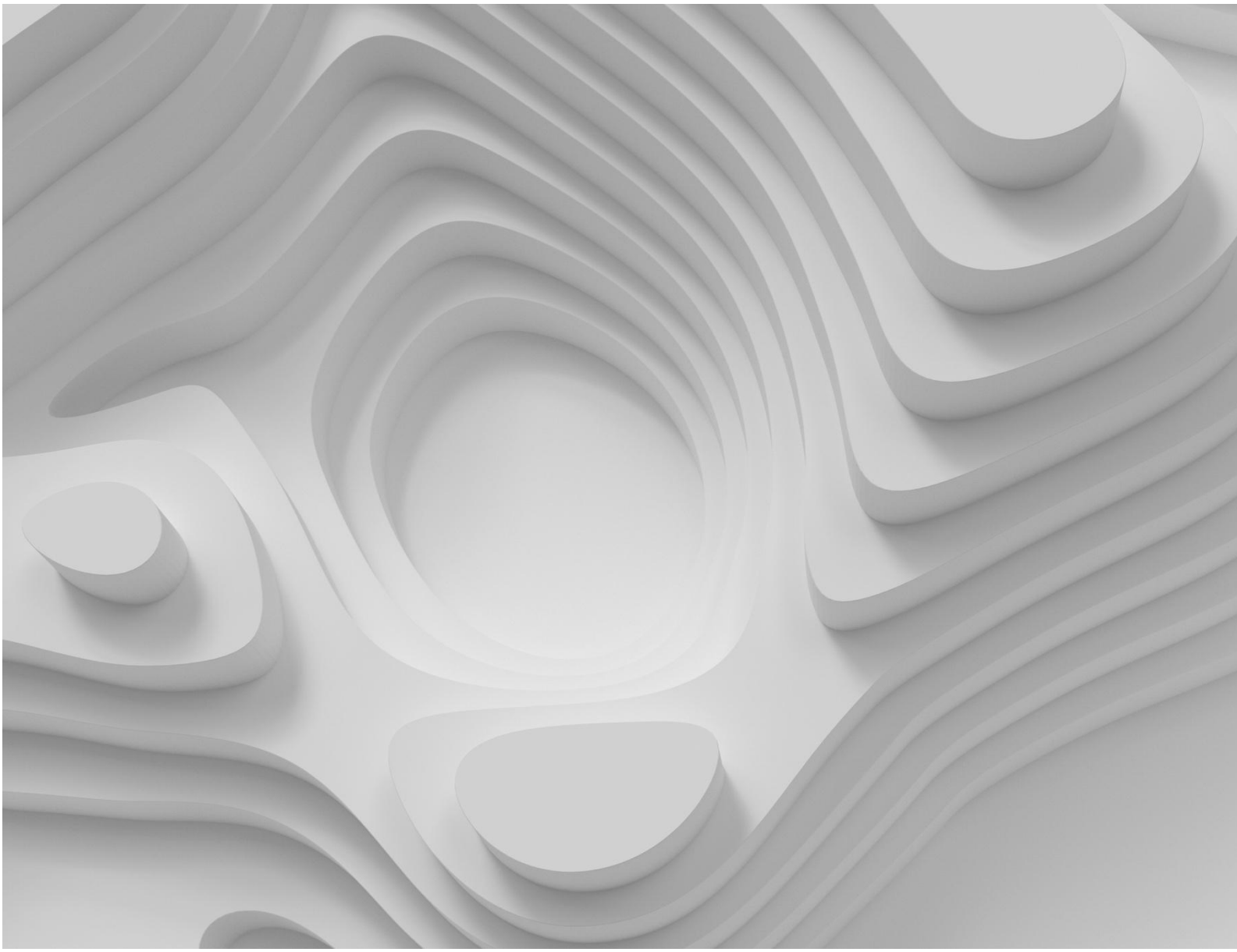
Palantir

Governing AI Agents

Palantir's Perspectives on Agentic Governance



Introduction

Artificial Intelligence (AI) Agents are AI systems built on top of Large Language Models (LLMs) that have access to tools which enable them to take actions on behalf of users. Whereas LLMs alone tend to work in a reactive mode responding to prompt inputs, the direct and intentional tool access of AI Agents enables novel Agentic abilities and actions which introduce novel challenges. Agents have numerous applications in the enterprise software context and have the potential to supercharge productivity and create new opportunities that previously did not exist. Already today, Agents are being embedded in software across institutions to analyze data, execute transactions, and recommend actions to humans at a scale that could have only been imagined even a year ago.

How well these Agents can perform in an enterprise context, how safe they are, and how reliable they are, are all still open questions. Capability testing of models on Agentic workflows from frontier labs is promising. Models perform increasingly well on a wide array of Agentic benchmarks which provide a baseline indication of how Agents can perform routine functions in standard settings. Those benchmarks, however, do not necessarily translate directly into how Agents perform in complex, dynamic, often uncertain enterprise settings. Safety measures applied to models also do not predict safe model behavior downstream. In-built model safeguards, which frontier labs take great care to introduce, quickly fall away when domain-specific context is introduced through fine-tuning, red teaming and other methods of fitting model applications to specialized environments. Research into the reliability of Agents suggests that current Agents are not consistent enough for safety-critical applications and that benchmark measurements do not give us a good understanding of how they will perform in a real-world environment.

To understand how Agents perform in operational contexts, robust infrastructure for in- and post-deployment evaluation and governance, not just model safeguards or upstream benchmarking, is integral for managing and monitoring Agents. Governing AI Agents requires a solid foundation to be in place – one which starts with controls for effective governance and is finished with tools for governing workflows that provide legibility into Agents' functionality over time.

Capability improvements do not predict real world performance, safety, or reliability

Model and Agentic benchmarks are useful for testing capabilities in a lab or controlled setting, but do not predict real world operational performance, safety, or reliability of Agents in open world environments.

Operational Performance

Testing frontier models in a lab does not effectively mirror real world conditions of how AI capabilities perform inside of a broader AI system and enterprise context. When exposed to data, tools, and contexts outside of the testbed, models and Agents embedded within systems perform differently than in a lab. Investments in experiments and research are starting to demonstrate this gap, but less attention has been paid to the operational performance of Agents in the enterprise context to date. The friction of effectively integrating Agents into production environments is demonstrated in a 2026 Snowflake report on the ROI of Generative AI, which points to 96% of users having problems implementing generative AI into the enterprise context, citing data quality, employee skills gap, integration with existing systems, and scalability and performance.

Palantir's experience embedding and testing models from nearly all vendors has highlighted the same gap. Working with customers to integrate frontier models across different enterprise contexts has provided us with direct insights into the limitations of laboratory model performance translating into the operational setting. We have seen, for example, how models and more recently Agents coming out of the box have under-performed expectations and often require substantial modifications and testing to meet desired business needs. This trend is not new, we have spent years demonstrating these learnings through our deployment of Classical Machine Learning Models and our findings have only strengthened with the adoption of LLMs and the deployment of Agents built on those models.

Safety

Mechanisms at the model level to ensure that AI Agents meet safety standards are insufficient. Implementing and maintaining those guardrails once models are deployed in the enterprise context are necessary. We have consistently seen this in our experiences deploying machine learning systems in production. The research community has started to draw the same conclusion for LLMs and have shown that even small amounts of fine-tuning of a model abandons any intentionally designed safety guardrails and that red-teaming models and Agents follow a similar trend. These issues that researchers are just starting to appreciate are all the more pronounced for models and Agents in complex production settings. This reaffirms our understanding that safety testing in a laboratory is not a substitute for downstream evaluation in production. Other engineering disciplines take it for granted that safety testing in a lab is useful but insufficient and must involve an operational component. The more LLMs are used in production, the clearer it becomes that Agents built on top of those models are no different.

Reliability

Another significant gap between lab and operational environments is the lack of reliability of Agents in an enterprise context. Researchers have used more controlled environments to establish that recent Agentic capability gains have only yielded small improvements in reliability. Unreliable Agents in non-critical settings may be okay. For example, an Agent that fails 1% of the time may be fine for sales emails, but automatically shipping code for production from those Agents would be an unacceptable risk. The same is true of other mission-critical contexts such as the military, financial institutions, and many healthcare settings. Either way, not knowing when your Agent will fail is a significant risk for any organization and highlights the need for understanding Agents in their deployed context.

Agentic Diffusion is Controllable

Given these limitations and growing adoption of Agents, many would start to sound the alarm that these Agents — which have uncertain operational performance, have few safety guardrails, and are unreliable — will soon start to uncontrollably disrupt our institutions and cause widespread harm. Lab capabilities are rapidly improving and organizations of all shapes and sizes are trying to figure out how to use Agentic AI and not get left behind. But if the risk landscape is poorly understood, the approaches to handling those risks are even more so and both industry adopters and those with governance responsibilities are at growing peril of assuming a zero-sum game. Currently, diffusion of AI Agents is much slower than Agentic lab capability improvements and that position is affirmed by real-world adoption and emerging research. This gap between lab capabilities and operational performance allows organizations to decide when to adopt Agents and to dictate how they are used: the gap leaves room for commensurate development and adoption of effective governance of these AI systems.

How Agents will be used to meaningfully improve people's workflows will take time to work out and should be a question for institutions — through the exercise of human discretion — to sort out. Similar to traditional software development, building an initial demonstrative version of an Agent that can be helpful for improving day to day work may not take too long, but building a sustainable robust Agent which reliably produces results that compound on what a user needs is a much more complex process which takes time. This process will determine how Agents are actually built and deployed, and just like in the development of most technologies, creates the space to layer governance on Agentic systems.

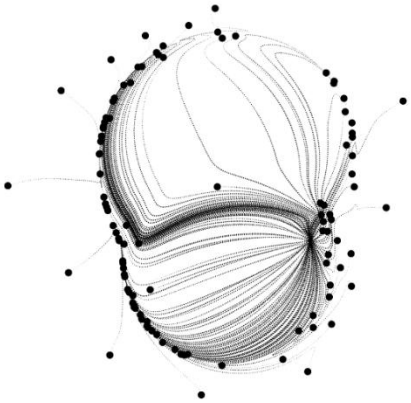
The role of governance extends further to how humans leverage Agents to pursue work as "teams". Humans and Agentic "collaboration" will require adjustments to determine how to most effectively deliver outcomes. Humans leveraging AI can perform better than humans alone but will not always produce strictly better outcomes. Use cases, the type of task that needs to be performed, and how to optimize their collaboration, like working with a colleague or new team, will take time and understanding. Take AI Coding Agents. They can easily produce code which is effective and efficient, but the code is very difficult for human engineers to understand. For example, the coding Agent may deploy unconventional techniques or curate functions differently than a human programmer might and therefore actually pushing that code to production is a process which will take time. As a result, where and how Agents will plug into workflows will be a longer process and is where deliberate governance can steer outcomes.

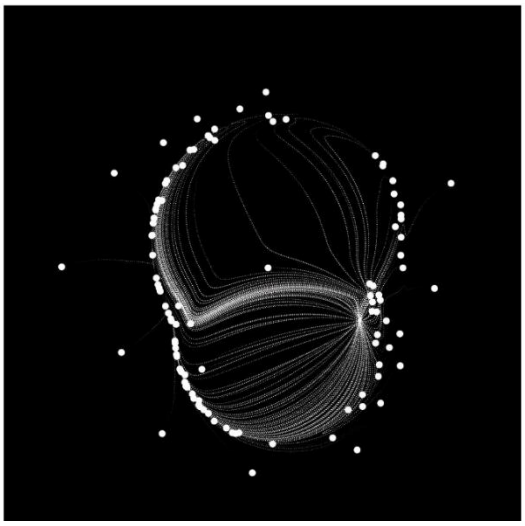
Governance Infrastructure in Practice

The gap between Agentic performance, safety, and reliability in theoretical and real-world settings is significant. This presents an opportunity for organizations to understand Agentic performance in the enterprise context and to monitor when they are not producing expected results. To do so, robust governance infrastructure for downstream evaluation of AI Agents in their deployed settings is necessary. The capabilities which ought to be embedded at the core of the AI system should be essential to the design of the infrastructure and surrounding workflows to ensure that Agents are legible and can be governed.

We break requirements into two layers: governance controls and governance workflows. **Governance controls** provide the foundation to ensure that primitives and tools are in place to facilitate supervision. **Governance workflows** build on top of those mechanisms to enable users to actualize their governance aims.

We bring these examples from our experiences deploying our software across a broad array of institutions for mission-critical outcomes. These practices and policies, which ought to be built on top of existing governance processes of enterprise software systems should guide how policy makers approach oversight and regulation of AI systems.



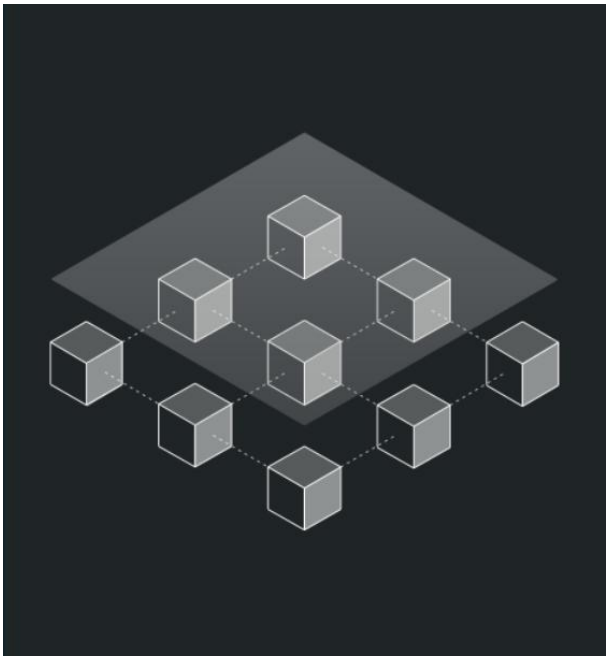


Governance Controls

A strong set of governance controls is a critical foundation for any context in which AI Agents are developed, deployed, or used. Just because Agents are powerful does not mean they should by default have access to all data or be able to take any action in any system. The decisions around their access to data and actions belong to people in the organization like administrators, developers, privacy officers, and other oversight teams. Ensuring that decisions reliably remain in human hands requires governance controls to be built into the system by default, not bolted on afterward.

- Authorization Workflows: Implementing authorization workflows across an organization that define both what a human user and what an Agent can access and what actions they can perform are clear examples of the way organizations can govern Agentic diffusion and protect against reliability concerns. A canonical example of access control is that an employee's compensation data should be visible to their compensation manager and the finance team but should not be visible to their peers. Similar to human users, Agent access ought to be restricted based on organizational norms. So, an Agent operating on behalf of the finance team may have access to compensation data, but other Agents should not. In highly sensitive environments such as a hospital system, it may be acceptable for Agents to take actions to support staffing workflows, but may be inappropriate for an Agent to take actions based on patient data. Authorization workflows may include adhering to classification-based access controls (CBAC), purpose-based access controls that enforce data use limitations based on purpose specified on collection, mandatory access controls that represent categorical data handling categories (PII, PHI, etc.), limiting Agents from taking user-authorized actions or other forms of granular controls which implement privacy-protective principles.
- Bounded Execution: Given that Agentic performance is mixed and that Agents remain operationally unreliable, Agents should only be authorized to take actions within a scoped environment and that starts at the foundational layer. All underlying processes should be guaranteed through mechanisms at the infrastructure layer which can ensure that applications operate based on precisely governed permissions. Similarly, what internal and external tools or APIs an Agent can hit should be restricted by default and any granted permissions should be decided explicitly by developers and an organization's security team. These sorts of constraints guarantee that humans continue to decide what actions Agents can take and are another example of the way in which diffusion will limit Agentic scope in production and ensure effective deployment of Agents.

- Testing & Evaluation: Measuring model and Agent performance and reliability in the enterprise context requires testing infrastructure built for operational settings. Relying solely on infrastructure for and outcomes from frontier model testing will not provide the tools and visibility needed by enterprise organizations. Further, testing Agents in operational setting requires support both for single-turn evaluations but also for multi-turn evaluations where Agents and users take multiple steps to execute outcomes. Testing and evaluation infrastructure ought to be accessible to domain experts and operators in production since they are the ones who understand the constraints of the context and know what "good" looks like.
- Observability: Compiling and maintaining an immutable record of actions taken by an Agent is an essential governance mechanism for administrators and compliance teams that cannot be overstated. Ensuring that human and Agentic data access and actions alike are captured in audit logs, telemetry, action traces, and other records is essential for maintaining organizational control and observability of Agents. Audit logs as the authoritative source of truth in any software system are also instrumental in post-hoc accountability.
- Fail-Safe Modes: As a stop gap for acute failures, organizations should also design and deploy mechanisms for cutting off agentic workflows when identifiable failures or constraints are encountered. Ensuring there is a way to turn "off" specific Agentic workflows is an important lever of control that ensures orchestration of Agentic governance.



Governance Workflows

While governance controls provide the technical foundational elements for enabling control of AI workflows, governance workflows are the methods, associated processes and policies within an organization, and applications that allow users to implement their plans and continuously monitor them. These workflows which technically grounded still require a culture which reinforces these norms to ensure they are a meaningful component of a sophisticated governance regime.

- Downstream Evaluation: Evaluation cannot be a one-time activity prior to deployment. Agents must be continuously evaluated over time against the specific use cases for which they are deployed, particularly where models have been fine-tuned, to understand if Agents are performing safely and reliably enough for a given context. Human feedback to agent outputs is the highest-quality signal available for understanding where an agent is failing, and this signal should be systematically captured and leveraged. Feedback capture should be embedded in the operational workflow where AI Agents are used, not in a separate tool, so that the transaction cost of providing feedback is low enough that operators actually use it. Evaluation processes are most effective when they are accessible to all stakeholders in the same environment where Agents are deployed.
- Human and Agentic Collaboration: As organizations measure Agentic performance and reliability in an operational context, they can begin to decide how much autonomy to assign to Agents. Agents by default should be configured to suggest actions to users until their role and utility is more clearly understood. As an Agent suggests an increasing number of actions, users, administrators and developers can determine if a certain set of actions would be appropriate for the Agent to execute loop automatically. Similar to data access, some actions that are less consequential may be appropriate for Agents to automate. Others may not be. The decision may also depend on if the action an Agent would take is reversible and if that is important given the context. Part of determining the right balance will be users' learning how to best leverage Agents for their own organization in a secure and transparent environment where potential operational pitfalls are most acutely understood.".
- AI Lifecycle Development: As organizations deploy more Agents, users, developers, governance teams, and an organization's leadership will need tools to manage AI use cases at scale. Understanding where AI is being used is essential because organizations cannot govern what they cannot see. Capabilities for use case management can include methods for inventorying or cataloging use cases, tracking use case progress along an AI Lifecycle, understanding adherence of a use case to AI policy requirements, or managing documentation related to use cases.

Conclusion

AI Agents present a genuine opportunity to transform how organizations operate – but only for those who understand what Agents can do in practice, not just in a lab. This moment of innovation calls for sophisticated thinking around operational testing and downstream evaluation of AI Agents to understand where desired outcomes are generated versus where AI may be introducing inefficiencies or worse. If policymakers want to really get a picture of how Agents are shaping our world, they should look to the enterprise context to get a better understanding of what is real and what is AI hype. Furthermore, building broader policies, legislation, and regulations around this framing – that downstream evaluations are a better measure of performance, safety, and reliability – will help ensure that protections address the most salient areas of concern and risk, but also have more enduring value as the technology landscape continues to evolve.