Establishing Guardrails for AI

Effective AI Governance for Testing, Deployment, and Use

Introduction

Generative AI has made AI much more useful for typical computing tasks.

This is Generative AI's year. Rather than being a specialized, "behind the scenes" technology as was previously the case with AI, generative AI platforms like ChatGPT and Bard have put AI front and center—with powerful results.

ChatGPT and Bard can rapidly and competently complete many time-intensive tasks typically associated with white-collar employees, including things like:

- Document analysis and summaries
- Data synthesis and reporting
- Letter writing and other kinds of writing
- Simple online research and data extraction
- Data format conversions
- Proposal and plan generation from background and context information

Now that tasks like these can be handed off to AI, they can be completed in minutes rather than hours, with some skill (even if in a few cases the results are inaccurate on close examination).

This is a compelling technology.

AI in Business

The productivity incentives inherent in AI use are evident in adoption rates. ChatGPT had only been available for a matter of weeks when it became the fastest software product in history to reach 100 million users.

In a survey of 1,000 business leaders in early 2023, half said that ChatGPT was already in use inside their environment in some way—and over 90 percent expected this use to continue to expand.[1]

The AI Tug-of-War

Transformative technologies can have complex and unexpected effects.

This is true for any organization, both for good and for bad—and this uncertainty is evident in a tug-of-war playing out in many companies today.

Business Teams and AI

The business or operational functions of many organizations rightly see in AI the potential for step-function productivity increases. They're eager to accomplish two AI-related tasks:

▶ AI experimentation.

Certain that AI can provide significant productivity gains or cost savings, business teams are eager to experiment with AI and develop expertise about how to leverage it in their organization.

▶ AI deployment or rollout.

Once these uses are found, defined, and documented, business teams are eager to realize the gains—to formulate a rollout strategy and see AI use become standard operating procedure, with

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

updated KPIs and other metrics, and in some cases even certain roles replaced with AI entirely.

The dollars that they see in these productivity gains are easy to covet, particularly when it's assumed that competitors will also benefit from AI. In many companies, the questions from business teams are "How soon can we do this?" and "Why isn't it done yet?"

Legal or Compliance Teams and AI

As business teams press for rapid AI adoption and use, legal and compliance teams push back. Their suspicions about AI are equally justified.

Several highly recognizable organizations have seen "incidents" related to AI use or become concerned about it, and have introduced restrictions or outright bans as a temporary measure until key problems and risks can be assessed and addressed.[2,3]

AI Data Disclosure Problems

The problems and risks in question are significant, and revolve mainly around the fundamental problem of inappropriate (and possibly catastrophic) data disclosure.

▶ Regulatory and compliance.

Most businesses and government agencies are subject to regulations around data disclosure—which data may (or may not) be disclosed, under what conditions, and what the penalties are for violations. Whether governed by HIPAA, PIPEDA, PCI-DSS, CCPA, or some other requirement, costs for data leakage are high—as high as four percent (4%) of revenue in the case of GDPR.

This risk haunts many of the tasks that employees can assign to AI, from summarizing reports to drafting HR documents to converting bulk data between formats. In asking AI to perform these tasks, employees may intentionally or inadvertently disclose confidential data to the AI platform in violation of compliance requirements.

▶ IP loss.

In industries such as technology or manufacturing, protection of intellectual property (IP) is a serious concern. Trade secrets, proprietary code, and other forms of IP may represent a significant portion of a company's value.

This risk is illustrated by the engineer who asks ChatGPT to debug a particularly frustrating piece of code, or the line worker who asks ChatGPT to convert weights and measures while following secret recipes or processes.

▶ AI account theft.

Data stored in AI accounts can also be stolen. Phishing and spear-phshing are well-known problems that have proven difficult to solve, and hundreds of thousands of AI accounts are already known to be available on the dark web.[4]

Platforms like ChatGPT store a history of a user's interactions with the AI—the prompts provided and the answers returned. If data has been inappropriately disclosed to the AI, account theft makes this data available to malicious actors.

▶ AI training.

There has been much obfuscation and discussion about whether AI platforms "learn" from the prompts that users provide to them.

Whether or not any particular platform learns from a particular interaction, it is certainly possible that this learning could in theory happen. Even if the practice is ultimately contested or litigated at some point, legal departments recognize that leaving confidential data handling and care to unrelated third parties is unacceptable.

For this reason, legal teams want to closely review fine print, licensing information, and service level selection when AI use is rolled out, so that any

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

terms that are unacceptable from a regulatory perspective can be ruled out.

▶ Company confidential information.

Most organizations also have company confidential information—strategy, future plans, internal figures, competitor research, and so on.

Leakage of this data is also a risk, as is the concern that AI platforms may be able to attribute multiple users to a single company—in the process learning something about the "bigger picture" at that company across interactions with these users.

For example, if one user at a sports manufacturer asks about product launch practices and synonyms for "lightweight," another for top tennis tournaments, and a third for major tennis publications and influencers, an AI could in theory infer that the company is about to launch a new, lightweight tennis product (likely a racket) and a social media campaign to support the launch.

▶ Licensing and liability.

Finally, in a world of intellectual property, legal and compliance teams are also trying to come to grips with concerns about leaking licensed data to AI systems, either in violation of the license or in ways that may incur further licensing costs.

An employee request to have AI to summarize a costly analyst report, for example, could incur new licensing fees and liabilities if discovered—and if the AI were to learn from this data and to begin to repeat it to others in the general public, the potential liability implications remain unclear.

IT in the Middle

IT departments are tasked with enabling work while ensuring that rules are followed.

This fact places IT departments in an uncomfortable bind today. Business teams are suggesting that AI is now required to enable the work—while legal and compliance teams insist that AI itself may break the rules.

A Seemingly Impossible Task

As a result, IT departments need to find a way to deliver AI to employees while ensuring that significant guardrails are placed around this access to mitigate data disclosure risks and concerns.

By their nature, however, generative AI systems don't lend themselves to current access control or policy enforcement techniques. Most access controls are grant-or-deny tools, suitable only for blocking access entirely, and policy enforcement tools generally work to cordon off items in a list of specific capabilities shown in a standardized user experience—which doesn't match how AI works.

Human Factors and BYOD Devices

IT departments are also aware of a key "human factors" reality—the fact that users will generally find the shortest distance between two points.

When a tool can save hours or even days of work, rules alone become far less persuasive. From an incentives point of view, most employees will find a way to adopt and use AI for their work to the extent possible.

This is particularly concerning in today's post-pandemic world, where both remote work and bring-your-own-device (BYOD) work are commonplace. Remote work and personal devices make AI use very tempting even if forbidden, and under concerning circumstances—unsecured networks, without endpoint agents, data loss prevention, or network controls in place.

For these reasons, any guardrails on AI use must be something more than a sternly worded policy—yet must support AI use with such low-friction that users aren't tempted simply to ignore the policy and circumvent it via their own devices.

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

Implementing Guardrails

What are IT departments to do?

What kinds of "guardrails" will enable organizations to plan and execute a low-friction AI rollout strategy while maintaining compliance and mitigating against risks—especially in organizations with some share of remote or BYOD work?

Plurilock recommends a three-tiered approach to this problem that can deliver robust AI usability while maintaining significant technical guardrails to mitigate against data disclosure risks.

The tiers are not necessarily sequential; each should inform the others, as they must align and work cooperatively to establish effective AI governance in the real world.

1. Adopt a strong AI governance policy.

Organizations should adopt a detailed, purpose specific AI use policy that spells out:

- Types of data at risk and subject to the policy
- Conditions under which AI may be used
- Conditions under which AI may not be used
- Acceptable parameters of that use with respect to data and any other concerns
- Consequences for policy violations
- Any processes related to the above items

This policy—which should be added to the employee handbook, called out during onboarding, and subject to signed agreement before employment—can have several general postures.

▶ Prohibition.

Most simply, an organization may simply forbid the use of AI for all employees. For the reasons already outlined, Plurilock does not recommend this solution, as it's unlikely to be effective given the strong incentives in favor of AI use.

▶ Organization-wide acceptable use.

An organization-wide acceptable use policy outlines rules that must be followed during AI use, but with no particular other day-to-day tools or processes in place to govern access or ensure adherence. Plurilock also does not recommend this solution, as it ultimately relies on the honor system and employees' ability—which may vary—to correctly adhere to the policy at all times.

▶ Conditional acceptable use.

A conditional acceptable use policy outlines the conditions that must be met before AI can be used, in addition to outlining acceptable uses. This is Plurilock's recommendation, as it makes clear to the employee that AI use is actively governed. Required conditions may include formal requests for access requiring written approval, participation in training, installation of software or tools, or any combination of these.

A Note About...

Hidden AI in Applications

In recent months, hundreds and hundreds of widely-used applications have adopted AI integrations to provide additional functionality. Writing assistants and grammar checkers, creative and web development software, coding platforms and engineering environments—virtually all of these have thrown their hats quickly into the "AI gold rush" by integrating AI behind the scenes to make their users more productive.

As AI policies and governance are being drafted, it is increasingly important for compliance and IT departments to collaborate on audits of existing software to determine what kinds of AI functionality are present in company software beyond direct interactions with AI SaaS websites.

In most cases, software that "transparently" uses AI to enhance work as employees carry it out—automatically providing writing and grammar suggestions, for example, or automatically helping to write code—is providing much of what an employee works on or even everything than an employee works on to an AI platform.

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

2. Control access in keeping with policy.

The next tier of guardrails that should be implemented is simple access control. As previously discussed, practical restrictions may not easily extend to remote employees or personal devices, but this gap should not lead companies to ignore access control altogether.

Where it can, access control gives a conditional acceptable use policy some "teeth"—with employee access to AI technically managed and governed as outlined in the policy.

▶ Apply firewall or network controls.

Employees that are not approved for AI use should not be have access to AI platforms from within the corporate network. If there are variable conditions attached to access (only in certain offices, only at certain times, only when certain tools are present), network configuration should dynamically reflect these restrictions.

▶ Apply DLP if possible.

AI technology is far too flexible and relies too much on natural language for traditional DLP technologies to be a natural fit, but endpoint DLP should be used to the extent possible to govern activity adjacent to AI.

▶ Balance friction with resources.

For the human factors reasons previously discussed—avoiding the "employee workaround" case—simple blocks should be avoided if possible during access control.

Instead, employees denied access should be automatically provided with the processes to follow to "unlock" AI access and the resources that they need in order to initiate these processes.

▶ Account for application AI integrations.

Note which applications in use have native AI integrations and are thus likely to transmit some or all of an employee's ongoing work to an AI model automatically.

Take a least-privilege approach to these platforms, placing them under the same kinds of controls, referencing them in the policy, and providing processes and requirements for access.

3. Grant selectively with guardrails.

Once access control measures are in place, selectively grant AI access to the extent possible under the policy. The conditions in which grants occur should in part inform the language of the policy and the controls and exclusions implemented.

▶ Review data jurisdiction.

Before granting, consider the employee at issue and the ways in which their data flows, both in terms of your own systems and in terms of AI platforms and AI-integrated applications.

In general precedent has not yet been established around these realities; for example, data about European customers may be governed differently from data on United States customers, so questions about the kinds of data that an employee works with and even where the employee resides may arise during the request process.

In cooperation with legal and compliance in each case, develop an understanding that informs each approval or denial, and as these are developed, roll them back into the policy and into access controls if appropriate.

▶ Grant with guardrails.

Where possible, grant AI access on the condition that technical guardrails are in place during use, rather than simply granting unrestricted access.

Plurilock recommends Plurilock AI PromptGuard, since it is able to detect and redact most confidential data before the AI receives it—while preserving that data in answers the user sees. This "data

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

protection that's invisible" prevents friction—eliminating the incentive to deviate from policy or circumvent guardrails in order to get work done.

▶ Maintain an audit log.

Ensure that whenever possible, employee AI use generates an auditable log. This should include more than just prompts, since generative AI systems are conversational and the meaning or consequences of an interaction may span many prompts and the responses to them, including "advice" or direction that the AI has provided.

We recommend Plurilock AI PromptGuard for this purpose as well. PromptGuard maintains auditable log of the interaction on your behalf and in the interest of your compliance, rather than relying on AI systems to accountably do this for you.

▶ Keep human factors in mind.

Throughout the process of selectively granting access with appropriate guardrails, remember that human factors are key; for technologies as compelling to employees as generative AI, if a policy makes work too hard to do, the user will find a way to circumvent prohibitions, adopting a posture that runs counter to good governance.

This is, once again, why Plurilock recommends Plurilock AI PromptGuard, which is designed to provide guardrails and prevent data leakage to the AI without incentivizing work-arounds.

Key Takeaways

Employee AI use is coming to your organization—whether you are ready for it or not.

The best available data says it's more likely than not that AI is already being used inside your organization, perhaps via ChatGPT, perhaps via Google Bard, and almost certainly via any one of the now innumerable applications that have rapidly integrated AI as a core part of their functionality.

Your Organization Has Existential Data

For this reason, the risks are great for organizations that do not have AI governance in place or a strategy for managing the transition to an AI-enabled workplace.

Most organizations, whether in business or in government, have some data whose importance is existential—where inappropriate disclosure threatens either the organization or the organization's fundamental reason for operating.

This may be trade secrets for a manufacturer, IP for an engineering firm, personal or customer data in education, healthcare, or finance, some combination of the above in government, or something else entirely.

Whatever the nature of your organization's existential data, you should assume that without guardrails, this data will be revealed, sooner rather than later, to a third party if you do not achieve reasonable, employee-friendly AI governance quickly.

A Note About...

Plurilock AI PromptGuard

PromptGuard is a Plurilock AI product that provides strong AI guardrails that also serve to enable low-friction AI use. Users interact with AI through PromptGuard just as they normally would, in a back-and-forth exchange.

As they do this, PromptGuard identifies likely confidential information, marks it as important to the user, and then redacts it, either by inserting fictional data or by randomizing or encrypting the values. When the AI returns an answer containing these redacted values, PromptGuard restores them before providing the answer to the user.

This enables common AI tasks like summaries of legal briefs, drafts of HR letters, or data format conversions to return the same copy-pasteable results—without ever delivering sensitive data to the AI platforms, and without creating friction that might cause the user to attempt to circumvent the guardrails that are in place.

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI

You Need Usable Guardrails

Because AI is such a powerful technology and productivity multiplier, you are unlikely to find that all of your users adhere to policy simply because it's the right thing to do.

Saved hours or even days or weeks of busywork are simply too persuasive an incentive to ignore.

For this reason, you should to implement robust guardrails that nonetheless don't incentivize workarounds for AI use, not just on AI rollout but even before then—during the experimentation and evaluation phases.

1. Adopt a strong AI governance policy.

- Establish an AI policy
- Add it to your employee handbook
- Make agreement a condition of employment
- Prefer a conditional acceptable use policy
- Update your policy as AI adoption plays out

2. Control access in keeping with policy.

- Operate on a least-privilege basis
- Implement firewall or network controls for corporate network users
- Add what DLP protections you can
- Prefer "how to" detours to access denials
- Control access to AI-integrated apps in the same way as AI platforms

3. Grant selectively with guardrails.

- Grant access in a way that provides technical guardrails
- Use Plurilock AI PromptGuard to hide data from AI but not from the user
- Use Plurilock AI PromptGuard to maintain an auditable log of all interactions
- Keep human factors in mind at all times, so as not to incentivize work-arounds

We Define Our AI Future

Generative AI has the potential to transform work as we know it.

Business teams are right to want to seize the future quickly and begin to leverage the value that AI promises to deliver.

Employees are one step ahead of that—already seizing the opportunity to turn hours-long tasks into minutes-long conversations.

At the same time, legal and compliance teams' concerns about risks are clear-headed and justified given AI's capabilities.

It is up to us to define the AI future ahead by deploying AI quickly, but with sound governance and good guardrails in place.

If we do it well, the future is bright. ■

References

1. https://www.resumebuilder.com/1-in-4-companies-have-already-replaced-workers-with-chatgpt/
2. https://fortune.com/2023/05/19/apple-restricts-chatgpt-employee-data-leaks-iphone/
3. https://www.sciencealert.com/many-companies-are-banning-chatgpt-this-is-why
4. https://mashable.com/article/chatgpt-stolen-accounts-passwords-dark-web

1.888.282.0696
sales@plurilock.com
www.plurilock.com

PlurilockAI