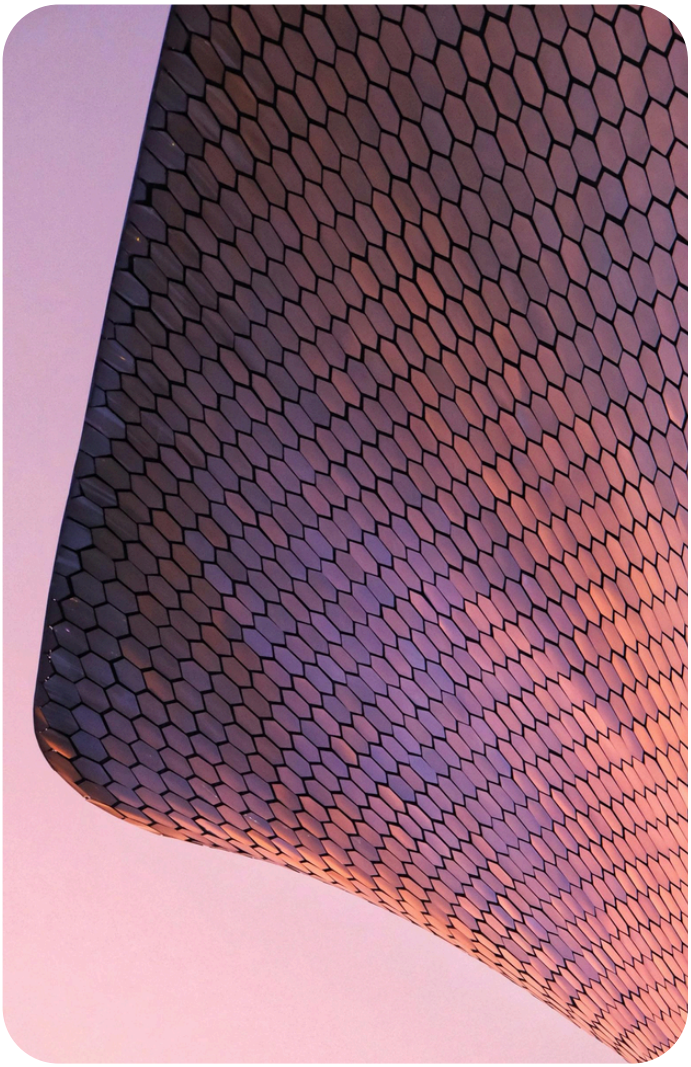
TRUSTIBLE™

AI Governance Triggers

When to Act and Why It Matters



Executive Summary

The rapid evolution of artificial intelligence—with continuous advancements in models, policies, and regulations—presents a growing challenge for AI governance teams. Organizations often struggle to determine when governance intervention is necessary in order to balance risk oversight without imposing excessive compliance burdens. This eBook introduces the concept of "AI Governance Triggers" to provide clarity on the specific AI model events that should prompt governance activities.

An AI Governance Trigger is an event that has the potential to impact an AI system and necessitate a governance response. These triggers may originate internally, such as proposing a new AI use case, or externally, including the enactment of new AI regulations. Understanding and categorizing these triggers is essential for effective AI governance. In this eBook we'll cover the key dimensions of AI Governance Triggers, including:

- **Descriptions** – Each trigger includes a clear definition and context to ensure a shared understanding of its significance.
- **Frequency** – Triggers vary in occurrence. Triggers like customer feedback are constant, while others, such as system decommissioning, may happen infrequently. Infrequent events are those that may happen only a few times per year at irregular intervals, while 'Constant' and 'Highly Frequent' events may happen on a daily, or weekly basis for AI focused organizations.
- **Key Stakeholder** – Triggers can arise from within an organization or from external sources. Internal triggers require proactive communication by the responsible team, whereas external triggers demand continuous monitoring. For internal events, it's important to identify who the key stakeholder is that will oversee any response, or kick-off governance activities.
- **Likely Impact** – The significance of a trigger is determined by its potential to alter an AI system's benefits, risks, or costs. Minor model adjustments typically result in minimal deviation, whereas major incidents—such as a high-profile AI failure—can lead to legal, reputational, or operational consequences, requiring extensive governance action.

This eBook provides a structured approach to identifying and responding to key events, ensuring that AI systems remain compliant, effective, and aligned with organizational objectives. In our next piece, we will explore common types of AI governance activities, ranging from automated AI evaluations to formal third-party audits, and share our insights on which governance measures are best suited for different triggers.

NAME	DESCRIPTION	FREQUENCY	KEY STAKEHOLDER	IMPACT LIKELIHOOD
Small Training Data Change	A small change to core training data, or fine-tuning datasets. This could include adding a few new datapoints from existing data sources. These changes nominally do not greatly alter the distribution or characteristics of the dataset.	Very Frequent	Data Engineering	Low
New Training Data Source	Adding a new data source to training or fine-tuning datasets. This may involve new types of data, additional licensed content datasets, or new geographic/demographic sources. The core aspect of this change is that the underlying characteristics of the training data may be altered.	Infrequent	Data Engineering	Medium
Change to Training Data Preprocessing	A change to data preprocessing criteria, logic, or implementation. This may include changes to data deduplication rules, toxic content or PII filtering, or removing copyrighted data for future retrainings.	Very Frequent	Data Engineering	Low
Update to Model Input/ Output Structure	Changes to the input or output modaliaties or structure for the model. This may include new input feature representations, or a change to the structure of model outputs, such as changing from a single output, to a distribution of outputs.	Very Frequent	AI / ML Team	Medium
Minor Model Code Change	A minor change made to underlying model training, or inference code. This may include fixing small bugs, refactoring code, or patching dependencies. These changes are unlikely to cause massive shifts in the outcomes of model training or inference.	Constantly	AI / ML Team	Medium
Major Model Code Change	A major model change that may involve updating a package to a new major version, a large architectural rewrite of the code base, or a shift in the general model architecture. The changes here pose a reasonable risk to shift the model's outputs and characteristics significantly.	Infrequent	AI / ML Team	Medium

NAME	DESCRIPTION	FREQUENCY	KEY STAKEHOLDER	IMPACT LIKELIHOOD
Model Retrained	The model is retrained incorporating updates to training data, model code changes, and new parameters settings. Model training time can vary widely from hours, to day or weeks.	Infrequent	AI / ML Team	Medium
Updated Fine Tuning	Changes to a model through fine-tuning. Fine-tuning may incorporate new data, alter the outputs of a model, or incorporate new safety/alignment rules. Fine-tuning is a less resource intensive task than a full model retraining, even though they are similar in concept and approach.	Frequent	AI / ML Team	Medium
Hyperparameter Change	Changing a hyperparameter, or inference parameter for a model. For example, changing the 'temperature' parameter for an LLM to alter its 'creativity'.	Very Frequent	AI / ML Team	Low
Evaluation/ Testing Updates	Changes to the evaluation or test suite used to automatically evaluate AI/ML models. This may include new examples, edge-cases, or new types of testing to catch previously identified errors.	Very Frequent	QA / Trust & Safety Team	Low
New Safety Rule/ Policy	Creation or update to some set of policy around inputs or outputs of an AI system. For example, this could involve a policy against used a specific phrase, rejecting a generation, or other filter/guardrail updates. This trigger point is primarily focused at changes to a system that integrates with an AI system, but is not handled directly within the model (for example, an 'AI firewall').	Constantly	QA / Trust & Safety Team	Low
New Vendor in Development Pipeline	A new vendor, tool, or other resource is integrated into the AI/ML development pipeline. This is particularly relevant if another set of stakeholders or entities now has access or control over another part of the pipeline.	Infrequent	Business / Product Team	Medium

NAME	DESCRIPTION	FREQUENCY	KEY STAKEHOLDER	IMPACT LIKELIHOOD
Increased Scope/ Deployment	The scope of an AI system is increased in a significant way. This may include releasing the system in a new legal jurisdiction, expanding out of an initial pilot/beta period, or a change to the deployment method of the model such as going from an API-only access model to open sourcing the model weights.	Very Frequent	Business / Product Team	High
New Team Members/ Leadership Change	Significant turnovers in key team personnel supporting an AI system that likely impacts model development, monitoring, or access. This could include replacement of key AI engineers, new corporate management that could shift model strategies or ethical practices, or scaling up team members with access to key details.	Frequent	Business / Product Team	High
New Vulnerability/ Best Practice Identified	A new ML vulnerability, best practice, or other recommended change is identified from academic literature, or other sources that necessitate an update to an ML system. Some risk management frameworks expect that known best practices for AI are always applied, but these requirements themselves shift over time.	Infrequent	External Event	Medium
Model Drift hits a threshold	Many ML systems may degrade in quality or accuracy over time as a consequence of shifting trends, events, or other external factors. If this model 'drift' reaches a certain point, it may post excessive risks and the system may need to be updated or retired.	Very Frequent	External Event	Medium
Adversarial Cyber Attack	An adversarial attack is launched against an AI system. This could be a denial of service attack, a model evasion attack, form of poisoning attack, or other AI specific cyber threats. Different kinds of threats and attacks may be easier to detect than others, and the treatment for each kind will differ.	Frequent	Cybersecurity	High
New Regulation	A new regulation, judicial ruling, or regulatory guidance is issues that is relevant to a specific AI system. The impacts of a new legal AI policy will have to be carefully evaluated for each AI system.	Very Frequent	External Event	High
New Proposed Use Case	An AI system, likely a general purpose AI system, is integrated into a new product or service and is being used in a net-new way. The benefits and risks of this use case will be determined on a case-by-case basis.	Frequent	Business / Product Team	High

NAME	DESCRIPTION	FREQUENCY	KEY STAKEHOLDER	IMPACT LIKELIHOOD
Model Retirement	A model is retired or decommissioned after previously having been used in a production context. The extent to which a model can be truly decommissioned depends on its deployment methods, with open source model being very difficult to truly remove from all sources.	Very Frequent	AI / ML Team	Medium
Incident Reported	A user, customer, or member of the public reports a significant unintentional failure of an AI system. This may include cases where the model did function as expected, but yielded a harmful outcome.	Very Frequent	External Event	Medium
New Documentation From Upstream Model Provider/Vendor	A model provider or vendor publishes new documentation that may be clarify or disclose new risks. This may include the publication of a vendors's red-teaming exercise, risk assessment, or other technical report that changes an organization's understanding of the model risks.	Infrequent	External Event	Low
Regulator/ Legal Investigation Launched	A legal proceeding related to an AI system is launched against your organization. This may take the form of a regulatory investigation, customer lawsuit, or other proceeding that may require an audit, or documentation transfer.	Very Frequent	External Event	High
Data Takedown Notice	A user or member of the public exercises their rights under data privacy regulations to remove personal information about them from a system, which may impact the AI systems built with it, or using it.	Frequent	External Event	Low
Major Technological Breakthrough	A significant technological breakthrough is made related to an AI system that could quickly impact the opportunities, costs, or risks of existing AI uses.	Very Frequent	External Event	High
Fixed Period of Time Passes	Even if an AI system is never updated, and there aren't signs of errors, it is still recommended to periodically review AI systems, especially high risk ones, to examine if there has been able significant model drift, whether the system is meeting its goals, and to evaluate if there are new recommended risk mitigations that should be put in place.	Infrequent	External Event	Low

TRUSTIBLE™

Website www.trustible.ai

E-mail contact@trustible.ai

HQ address 1201 Wilson Blvd, Floor 27
Arlington, VA, USA 22209