Understanding Industry Practitioners' Experiences in Generative AI Governance

Hyo Jin Do
IBM Research
Cambridge, Massachusetts, USA
hjdo@ibm.com

Swati Babbar
IBM India Software Labs
Kochi, India
swati.swati5@ibm.com

Wenjing Li
IBM
Austin, Texas, USA
liw@us.ibm.com

Laura Walks
IBM
Austin, Texas, USA
laura.walks@ibm.com

Shayenna Misko
IBM Software
Böblingen, Germany
shayenna.misko@ibm.com

Abstract

AI governance has become critical, especially as generative AI technology introduces new complexities and uncertainties that require robust risk management. While the need for frameworks and solutions to support AI governance is widely recognized, understanding and addressing the real-world needs of AI practitioners in *operationalizing* governance remains underexplored. To bridge this gap, we conducted semi-structured interviews using a design probe with AI governance practitioners across various industry sectors. Our findings provide insights into the experiences and pain points of industry practitioners in AI governance, highlighting key challenges in achieving performance goals, assessing societal impact, securing user data, and navigating technical difficulties. We also identified their technical and explainability needs, including practical guidance on addressing violations, as well as more detailed explanations of AI models, data, and evaluation. We discuss design guidelines for AI governance tools that effectively support practitioners' needs.

CCS Concepts

• **Human-centered computing** → **Usability testing**; **Empirical studies in HCI**; **User studies**.

Keywords

AI Governance, Industry, Explainability

ACM Reference Format:
Hyo Jin Do, Swati Babbar, Wenjing Li, Laura Walks, and Shayenna Misko. 2025. Understanding Industry Practitioners' Experiences in Generative AI Governance. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3720275

1 Introduction

Generative artificial intelligence (AI) technology opens up new opportunities with impressive capabilities but also introduces unique challenges and risks. For example, generative language models have been observed to "hallucinate" information, producing outputs that are plausible but factually incorrect or unfaithful to the source, which can lead to overreliance or automation bias [13]. They are also found to encode harmful biases, which propagate discrimination and stereotyping in society [19]. Addressing these challenges of generative AI models is difficult due to novel and complex model capabilities and behaviors, its massive and opaque architectures, proprietary technology, rapid evolution, and complex applications that may have multiple agents interact with each other [21].

To tackle these issues, enhanced governance for generative AI models is essential. Government and institutional regulators have proposed numerous frameworks and policies to govern generative AI development and usage including the AI Risk Management Framework [33], Responsible AI Safety and Effectiveness (RAISE) AI Benchmark [28], the EU AI Act [7], and South Korea AI Basic Act [35]. Additionally, researchers have proposed various governance tools and solutions to support practitioners (e.g. [2, 6, 9, 26]). However, many regulations in AI governance remain theoretical or abstract [11, 17, 24, 27], relying on the practitioners' interpretation of what constitutes AI governance in their contexts. There is also a knowledge gap in understanding how current governance tools are perceived and used by AI governance practitioners. To bridge the gaps, **we aim to engage AI practitioners in the conversations to inform the development of effective and practical governance guidelines and tools.** This study focuses on the following research questions (RQs): 1) What are the practitioners' **goals** for generative AI governance? 2) What **challenges** do they face in achieving those goals? 3) What do practitioners **need** for AI governance? 4) How might **tools** support their needs?

We conducted a semi-structured interview study with industry practitioners involved in generative AI governance. The interview consisted of two parts: the first explored current governance practices, and the second used our design probe to gather in-depth insights on governance tool designs. This paper contributes to the literature on AI governance by presenting empirical insights from AI governance practitioners. Our findings reveal unique experiences, challenges, and needs of AI governance practitioners that have not been fully addressed or considered by current tools and frameworks. We discuss practical design guidelines for future AI governance tools that assist governance in practice.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720275

2 Related Work

2.1 Generative AI Governance Guidelines and Frameworks

While generative AI has been rapidly adopted in various domains, it also poses unique exacerbated forms of risks [3, 12, 34]. To manage these risks, regulators and industry leaders are working together for AI governance, establishing policies and principles for the development, deployment, and use of responsible AI [6, 22, 28, 33, 39]. For example, National Institute of Standards and Technology (NIST) has developed the AI Risk Management Framework (AI RMF) [33] which provides a conceptual guidance for characterizing responsible AI and identifying risks. EU AI Act [7] will come into force by 2025, which requires transparency of general-purpose AI models, such as providing technical documentation and instructions for use, complying with the Copyright Directive, and publishing a summary about the content used for training. Along with the evolving landscape of AI governance with various stakeholders, Corrêa et al. [8] reviewed 200 governance policies worldwide and found that the most prevalent principle included transparency and explainability. The principles support the idea that AI systems should be transparent for all interested stakeholders and such information should be understandable.

However, operationalizing the abstract governance principles (e.g. transparency, explainability) into practice can be a challenge [32]. To provide more practical guidelines, NIST has published a companion resources [3, 4] that suggests actions to manage different AI risks. Lu et al. [22] proposed an responsible AI (RAI) pattern catalog based on a literature review to understand patterns that stakeholders can undertake to implement RAI in practice. Our work complements this research by adopting empirical approach that involves listening to and understanding practitioners' current challenges of translating governance into their practice. We conducted interviews with industry practitioners in AI governance roles to explicate their perspectives and experiences in governance tasks and identify gaps in current policies that fail to account for real-world needs. While governance requires collaboration across numerous stakeholders throughout the entire AI lifecycle [2, 22, 31, 38], we scope our work on team-level stakeholders [22], governing AI models through developing, validating, monitoring, and refining AI models, including data scientists, developers, testers, and operators.

2.2 AI Governance Tools

We join a growing group of researchers designing human-centered tools to assist users in understanding and managing behaviors and risks associated with generative AI models. There are numerous AI governance tools, toolkits, and solutions, focusing on documentation [2, 6, 10, 26, 29], impact assessment [14, 30, 37], and development [15] of generative AI systems (e.g. survey papers [22, 40]). Several researchers took a participatory design approach to incorporate stakeholders' needs within the tool design. Kawakami et al. [15] created Situate AI Guidebook toolkit that scaffolds early-stage deliberation questions around whether to develop and deploy an AI model, informed by public sector agencies and community advocacy groups. Through interactions with a civil rights organization and community organizations, Algorithmic Equity Toolkit [14] was invented to question about potential impacts of AI systems. Leiser et al. [18] conducted a participatory design study with workshop participants and machine learning experts which identified desired features for governance tools to detect and mitigate hallucinations of LLM. Wang et al. [37] conducted a user study with AI prototypers to assess how effectively their tool helps in anticipating potential harms in their prototypes. While most of these studies focus on a specific problem, goal, or resource of AI governance, we take a step back to provide a more comprehensive understanding of the complex challenges and needs of AI governance practitioners. Therefore, our research results provide novel insights into the areas where practitioners prioritize, put the most effort, or face difficulties across various governance tasks. Through a design probe and a question bank, we also contribute to the ideation of practical AI governance tool designs to support practitioners' needs.

3 Method

3.1 Interview Protocol

The interviews were held through a video conferencing platform, which lasted about 45-60 minutes. The interview was divided into two phases. In the first phase, we asked open-ended questions to understand their current experiences in AI governance practice such as their goals, tasks, and challenges. We also asked whether they currently use AI governance tools or solutions and how well the tools support their needs, if any. In the second phase, we shared a link to our low-fidelity interface designs of a governance tool as a design probe. Participants were asked to share their screens and think aloud their thoughts while navigating each screen. We asked design-focused questions in this phase such as what they liked or disliked about each screen, and what questions they had. After they completed going through all the screens, we wrapped up the interview by asking final questions such as their overall experience using the design probe and how well it might help them to conduct their governance tasks at work. The complete list of leading questions is listed in Appendix B.

3.2 Design Probe and User Scenario

We introduced a hypothetical AI governance scenario with low-fidelity interface designs using Figma software[1]. We designed the interfaces with the following design goals, which were inspired by literature and existing solutions: 1) The interface should show a user-friendly dashboard that summarizes key evaluation metrics of an AI model to assess performance and compliance violations; 2) The interface should incorporate interactive visualizations to help practitioners understand the data; 3) The interface should allow users to evaluate individual transactions. Based on these goals, we designed interfaces as shown in Figure 1.

To help participants understand the design probe, a researcher walked them through the interactions of the design probe using a hypothetical scenario involving an AI practitioner named Derek. Derek is an app developer in an AI governance team. Derek's role involves running evaluations to monitor how the deployed model performs against the metric thresholds set by the risk and compliance owner. One critical evaluation he performs is whether the

[1]www.figma.com

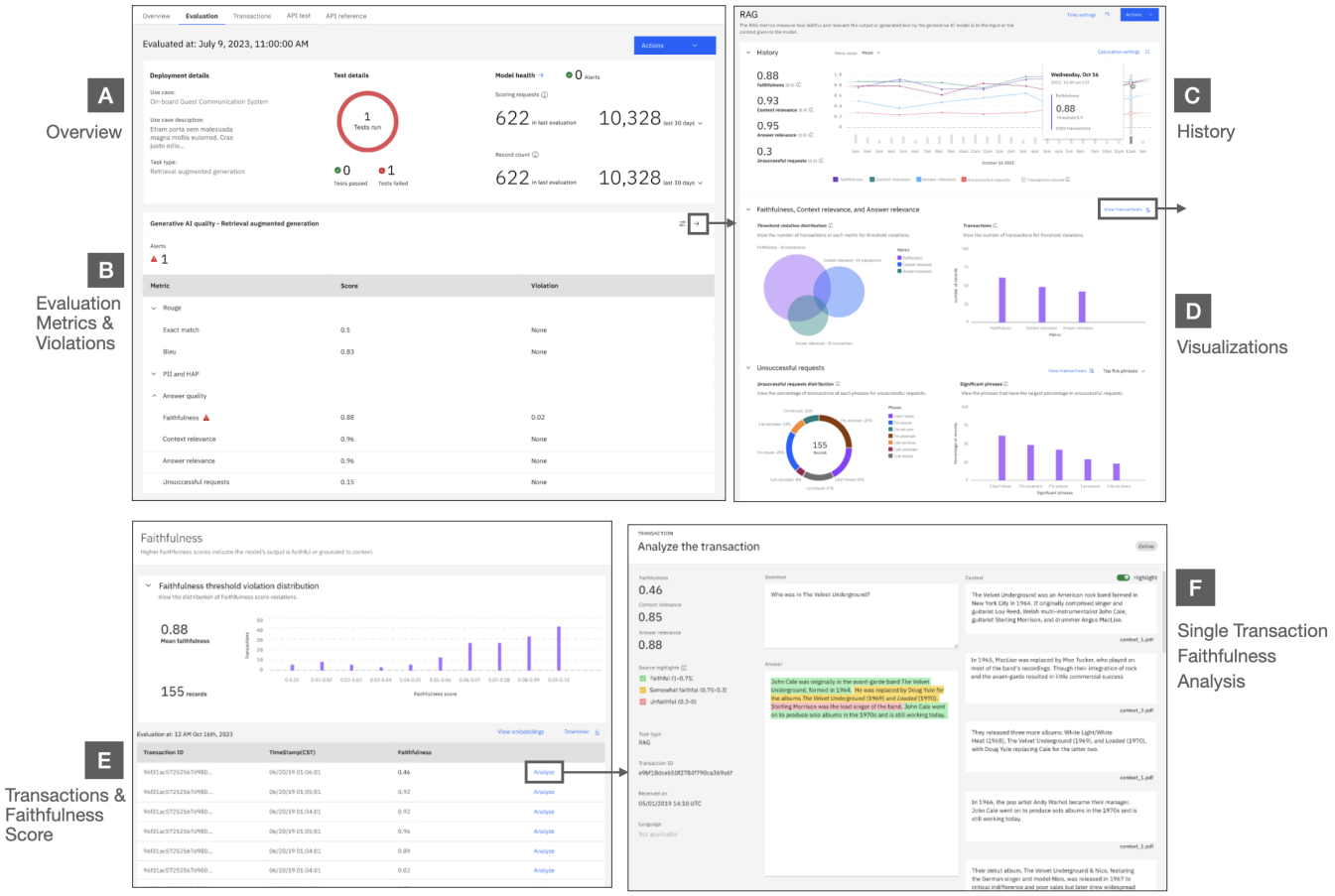

Figure 1: The first two screens consist of a dashboard that summarizes evaluation metrics (A, B), followed by interactive visualizations to understand data (C, D). The last two screens contain a list of user transactions (E) sorted by an evaluation metric (e.g. faithfulness), allowing practitioners to view an individual transaction to analyze problematic outcomes (F).

model produces faithful outcomes that are consistent with factual sources or contexts (i.e. faithfulness). Derek uses a governance tool for configuring monitors that evaluate models against thresholds. Derek accesses a dashboard to monitor the real-time performance of the deployed model, which summarizes the performance of multiple prompts over time (A). Derek observes that the model's faithfulness metric for retrieval augmented generation (RAG) tasks falls below a specified threshold (B). Derek investigates the RAG case further to explore how the faithfulness metric varies over time (C). Additionally, Derek checks various charts to understand evaluation results (D). Derek views the list of RAG transactions (E). Derek notices a recent transaction with a particularly low faithfulness score. Derek clicks 'Analyze' button for that single transaction. Derek examines that transaction closely to understand the problem behind the low faithfulness score (F).

3.3 Participants

We targeted practitioners who are employees with experience in developing, validating, and/or monitoring generative AI models for governance. We distributed a screening survey using two user research recruiting platforms, User Interviews[2] and Respondent[3], which are known to connect with professionals in this field. The screening survey included questions about their experiences related to AI governance, which we described in Appendix A. After the screening, we invited 11 participants. One participant's data was removed after the interview as the participants were not able to provide detailed explanations about their role. Table 1 summarizes their job titles, industry domain, demographics, and example governance tasks. We compensated $55 for participation. All participants provided written informed consent and were treated following the guidelines for the ethical treatment of human subjects.

3.4 Analysis

The first and second authors, who have expertise in UI/UX research in AI software, conducted a thematic analysis following the steps outlined by Braun and Clarke [5] to analyze interview data using

[2]https://www.userinterviews.com/
[3]https://www.respondent.io/

Table 1: We analyzed 10 participants' data who are involved in developing, validating, and/or monitoring generative AI models.

ID	Job title	Industry Domain	Org. Size	Gender	Ethnicity	Governance Task Example
P1	Product Manager	Information Technology	10,001+	Male	White or Caucasian	Develop/validate conversational LLM search
P2	Technology Consultant	Medical Insurance	10,001+	Female	Asian or Pacific Islander	Validate LLMs via real-time monitoring
P3	Technical Specialist	Healthcare	5001-10,000	Male	American Indian/Alaskan Native	Train/validate healthcare cost prediction AI model
P4	AI Engineer	Information Technology	501-1000	Male	Asian or Pacific Islander	Implement risk assessment frameworks
P5	AI Prompt Engineer	Information Technology	10,001+	Female	Asian or Pacific Islander	Run AI model security and safety tests via prompting
P6	Data scientist/Data Engineer	Financial Services	5001-10,000	Female	Asian or Pacific Islander	Develop tools and frameworks for bias detection
P7	Application Development Senior Analyst	Information Technology	501-1000	Female	Develop AI Asian or Pacific Islander	Analyze large-scale data and build AI solutions
P8	Sr. Solution Architect	Financial service	10,001+	Male	Asian or Pacific Islander	Develop AI chatbot models for bank
P9	Senior Site Reliability Engineer	Information Technology	10,001+	Male	Asian or Pacific Islander	Monitor/validate AI models for sentiment analysis
P10	Data Science Lead	Telecommunications	10,001+	Female	Asian or Pacific Islander	Analyze clients' data in AI chatbots

affinity diagramming [23]. The data consisted of interview transcripts and detailed notes taken during the interview. The data was then divided into idea units and copied to a Mural board[4], an online collaboration tool. The researchers independently coded these idea units and collated the codes into potential high-level themes. Through an iterative process, researchers discussed the themes until reaching a consensus, then created and refined the final set of themes. A sample of the data (approx. 10%) was coded by both researchers, achieving a Krippendorff's alpha of 0.77, indicating substantial agreement [16]. The researchers labeled all the data using the final set of themes and ensured perfect agreement in the label assignments.

4 Results

4.1 Goals and Challenges in AI Governance Practice

In this section, we highlight four primary goals discussed by industry practitioners, as well as their challenges in achieving those goals. Notably, many of reported challenges necessitate human intervention to resolve.

The most common goal emphasized by eight participants was **to improve the AI model's quality** by continuously assessing and monitoring the model outputs through various performance metrics including accuracy, biases, and fairness. For example, P7 said, *"I want [AI models] to be as accurate as possible"* and P4 said, *"[AI modls] should not be biased and fair. If I have a thousand test datasets, I want all thousand samples to be classified correctly, and accurately, and no mislabeling."* Relatedly, four participants mentioned **assessing ethical and societal impact** as their goal. They wanted to ensure that their systems are designed and used in ways that align with ethical principles, human rights, and societal values while avoiding potential harms and risks. P1 explained, *"We evaluate the potential ethical implications and society impacts of the LLM outputs including misinformation, harmful content generation, [...] and reinforcement of stereotypes."*

However, participants had difficulty **evaluating and advancing their AI models to meet their target performance** and quality standards, such as *"achieving the best-fit fairness. [P4]"* Evaluating model outcomes can also be complicated and nuanced, and human interventions were sometimes required to identify the risks within the outcomes. P5 explained that when model outcomes fall into a gray area in risk assessment, they exercise their own discretion: *"If a model has generated a response that is violent or if it clearly exhibits some sort of bias, then it's black and white, right? But if it's something that's in between, I'm trying to use my own moral compass to decide whether this is right or wrong."*

Another goal was to ensure that their AI systems **handle and store personal data securely**, and protect individuals' or clients' privacy rights. P10 said: *"We need to make sure the data that we use in this model does not include any sensitive data [...] So we need to set up some policies to prevent using those data."* However, there were concerns about **keeping data private and confidential when using third-party governance platforms**. For example, P6 said: *"There's a big challenge in terms of data protection until I use it like some S3 server or put it in our controlled environment. There's a challenge that we can't just put it out in an open server."* Participants also expressed challenges in validating the AI model's security, as P1 said: *"Ensuring the robustness against virtual attacks, input perturbations, and unseen scenarios are significant challenges."*

The fourth goal was to **ensure compliance** with government and institutional regulations. Three participants described that they work on their AI models to fulfill legal compliance needs and ensure that these are kept up-to-date, synchronized, and effectively managed in their governance systems. For example, P3 described their goal as updating their medical AI system to be validated against the evolving external regulations: *"DRG as a bundle is something that is invented by CMS, the governing body of healthcare in the United States. That definition keeps updating on the portal. [...] We use that as a reference for our AI governance model to treat itself."*

Practitioners encountered challenges in **interpreting the regulations in their specific contexts** due to their ambiguity and the absence of specific guidelines for differing domains, which often necessitates human judgment to practically implement it. For example, P6 explained that policies around disparate impact are often vague, and human discretion is essential to implement policies in their context: *"What's the disparate impact? It can depend on [the size of the company]. Say, the larger the bank, they are governed to more policies, versus we are a FinTech, we're not governed to that much. So what's the level of governance that I need for my model? That's where it requires the discretion of a human."*

There were also other **technical automation challenges** throughout the governance lifecycle, including training, testing, scaling, controlling, and testing the AI models. Participants emphasized that many of these processes require manual work such as the integration of data from different warehouses, policy sources, tools, and models. For example, P3 explained the manual integration work they need to do whenever a new governance policy gets updated: *"It [policy] gets published onto the third party portal and we can't*

[4]https://mural.co/

have the API, due to its sensitivity that we are dealing with. [...] We have to manually integrate anything that is part of the compliance policy"

4.2 Needs and Current Solutions

To practice governance, participants mentioned that they need **evaluation metrics** to assess their models. However, there was a high variability in metrics they needed for evaluation, as they use specific models tailored to their use cases and different models have distinct metrics. For example, P5 explained that appropriate metrics depend on whether the AI model is conversational or not: *"Every project has its own evaluation rubric. [...] It depends on what the model is, what the use case for the model is, and what the client has defined in terms of what they deem acceptable for the model."*

Five participants also needed information about the used **AI models** such as the model's architecture and parameters, training process, and data processing method, which are not often available in proprietary AI models. P8 said: *"now these LLM models are like a black box. You don't know what's going on behind the scenes. So you should have it somewhere, where you can get insight on how it is working, what are the weights, what are the different parameters it is using, and how it is pulling the information from the system, if we need to fine-tune."*

Additionally, four participants reported that they needed **user data and use case** information to discover and understand their problems, use cases, and their characteristics (e.g. level of expertise). For example, P9 who works for customer service application said, *"We need to gather those inputs from those people who are deeply engaged with that type of profession, meaning who have their core jobs and responsibilities with respect to customer service."*

However, when we asked how they are addressing those needs currently, many participants expressed that their needs are not fully addressed, and often need to do a lot of manual work for their needs, such as cleaning the data, fine-tuning, evaluating, validating, and retraining their models. For instance, P8, who previously emphasized the need for AI model and data transparency, noted that since this information is not readily available, they face challenges in comprehending the model and enhancing it through manual work: *"It's really manual. You can say validations, which we are doing it, we are testing it, we are running evaluations around it."*

Their current AI governance tools were also limited to support their needs. Seven participants expressed dissatisfaction with the **limited performance evaluations and metrics** their tools have. P7 mentioned that some rare metrics are not supported by their tools so they have to measure it by themselves: *"We don't use any tool to measure the 'data drift'. That is manually done because it's quite a rare thing."* Participants also pointed out that these tools often lack customization for a specific domain such as healthcare. Hence, the tools may not accurately apprehend risk factors relevant to the specific domain. Moreover, participants mentioned they would like the tools to incorporate real-time evaluation of the model outputs before presenting them to the users. P2 stated that their risk assessments are often conducted by gathering user feedback through market research tools such as surveys, which makes it challenging to maintain and track risks and grade logs in real-time.

Another key limitation participants pointed out was about **lacking explainability features** in the tools. Five participants mentioned that their current tools have limited or no explainability features in which they often have to make sense of the outputs and the model's performance manually. P8 elaborated, *"It is just those data points which we collect around accuracy, but there is not a specific tool which explains that like why those points are made. It is all our investigation."*

4.3 Design Probe Feedback

All participants expressed satisfaction with our design probe, saying that the tool can reduce their manual work and increase their work efficiency. Participants appreciated how information was displayed and arranged, such as on overview dashboards and visualizations. In analyzing individual transactions, participants found the interactive explanations (e.g. highlights) useful as these will help them to evaluate faithfulness metrics easily.

When participants were going through the design probe, three suggestions for technical and explainability improvements emerged across all screens. A technical feature that participants desired was a **recommendation feature on how to resolve violations**. P9 explained, *"We are having a violation and, on that score, if there is some sort of recommendation or best practice approach, a popup notification will be going to appear after that."*

A significant number of participants reported challenges in **comprehending terminologies**, particularly the metric names that they are not familiar with, due to a high variability of metrics and evaluation methods they use and a lack of standardization. P6 mentioned, *"There should be a definition section like faithfulness means this and why is it important."*

Participants also expressed a desire for a more in-depth **understanding of the context guiding the evaluation process**. To facilitate a more targeted approach to addressing the underlying issues, they requested global explanations about 1) AI models (e.g. model name, parameters, training data, sources), 2) data and prompts (e.g. the number of prompts evaluated, the types of questions asked) and 3) evaluations (e.g. details on associated violations). P8 mentioned that such information is often very complex in practice, because *"In a company, you have multiple LLMs working, and multiple applications are there. It's good to have all those details."*

We further expanded participants' specific **explainability needs** through the question-driven XAI approach [20], as an explanation is essentially "an answer to a question" [25]. We constructed a question bank (Table 2) based on the questions practitioners asked. It includes global explanations about the AI model, data, context, and evaluation methods; counterfactual explanations for different types of data; local explanations for failed test cases; and example-based explanations about possible solutions.

5 Discussion and Conclusion

Practitioners' primary governance task was to improve the AI model's performance and assess its broader impact. However, they found this challenging due to technical difficulty of improving the model quality and resolving risks. To tackle this, AI governance tools can provide actionable guidelines and real-time monitoring features to enhance model performance and address violations

Table 2: AI governance question categories that emerged in our design probe study. The questions represent the explainability needs that could be supported by AI governance tools.

Category	Question Type	Sample Questions
AI Model	What/How (Global)	What AI models are used? What are the model's parameters? What is the use case of the model? How was the model trained/fine-tuned? How does the model handle sensitive data? How well does the model perform? How frequently was the model fine-tuned?
Data/ Context	What/How (Global)	What training/testing data was used? What is the input/output type? How is the data quality?
	What If (Counterfactual)	What if the task is open-ended? What if the task has multiple answers? What if the output is shorter/longer?
Evaluations/ Violations	What/How (Global)	What is the evaluation criteria? What are the metrics and their meaning? How is the metrics calculated? How many transactions were tested? How might the average scores trend over time? What are the min/max values of the metrics? What is the impact of the violation? What are the risks associated with the violations?
	Why (Local)	Why did the test fail? What specific metric of interest was violated?
	How to be that (Example based)	How to fix the problem or improve it?

easily and promptly, going beyond merely presenting evaluation results [1]. While many existing tools help users anticipate risks from AI systems [14, 37], our research highlights the challenge practitioners face in interpreting this information and exercising their discretion to make final assessments, especially in nuanced scenarios. The tools may assess risks through reinforcement learning, which could learn from user feedback and eventually automate the process. Moreover, participants found that current governance tools are limited in their evaluation capabilities and metrics specific to their domain, which aligns with the limitations of many existing tools [36]. Our research demonstrates the importance of customization features to tailor evaluations to users' domains and contexts. Open-sourcing the tool can also facilitate community-driven development and enable adaptability to diverse use cases.

Keeping AI models compliant with governance and institutional regulations is also crucial for practitioners. However, our results revealed hurdles in manually integrating external policies into their governance platforms. Our findings reaffirm existing criticisms of governance policies for their lack of specificity [11, 22, 24, 27]. Additionally, we also discovered that these policies are constantly evolving, further complicating practitioners' ability to review and implement changes promptly, offering empirical insights into the temporal gap between technological advancements and legal adaptations [17]. Therefore, we suggest that regulatory bodies should collaborate with companies and business leaders to automate the integration process, define specific guidelines for practical implementation, and alert or train practitioners about major changes in policies.

AI governance tools should enhance explainability features that provide clear insights into the AI model, data, context, evaluations, and violations. When exposed to our governance design probe for the first time, participants often required clarification on key terminologies and details about the evaluation context. To address this need, we propose our question bank (Table 2) as a practical checklist or deliberation prompts when deciding appropriate explanations to offer within a governance tool. We also suggest incorporating interactive features, such as tooltips, and detailed documentation that offer more explanations in the tool's terminologies, features, and functionalities.

Data protection is another major concern that practitioners have, and it is essential to ensure that the governance tool stores and transfers data in a controlled, safe environment. The tool should implement robust security measures that align with their organization's data protection policies and requirements, including encryption methods to protect data, and provide users with transparency and control over their data, as well as resilience to malicious attacks and unintentional user errors through user authentication mechanisms and regular security audits and training.

As to limitations, we focused on model developers and validators among various roles in AI governance. While this allowed us to gain in-depth insights into that particular roles, it leaves room for future research to explore other key stakeholders, including regulators and end-users. To enhance the generalizability of our findings, we recommend that future studies expand the sample size and scope, including smaller organizations and exploring other domains.

To conclude, our interview study reveals that current generative AI governance imposes a significant burden on practitioners in integrating and implementing the policies into their governance tasks, which often demands manual work and human discernment. To effectively support practitioners, we suggest technical features for governance tools to improve evaluation and impact assessment, regulatory compliance, and security. Furthermore, we propose the question bank for governance tool developers and designers to leverage for enhanced explainability of their tool.

References

[1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[2] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[3] National Institute of Standards Artificial Intelligence (AI) and Technology (NIST). 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

[4] National Institute of Standards Artificial Intelligence (AI) and Technology (NIST). 2024. NIST AI RMF Playbook. Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook (last accessed: 2025/01/23).

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Kasia Chmielinski, Sarah Newman, Chris N. Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, Mary L. Gray, Audrey Chang, , and Maui Hudson. 2024. The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers. *The Shorenstein Center on Media, Politics and Public Policy* (2024).

[7] European Commission. [n. d.]. laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj

[8] Nicholas Kluge Corrêa, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza

Galvão, Edmund Terem, et al. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns* 4, 10 (2023).

[9] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 427–439.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[11] Michael Guihot, Anne F Matthew, and Nicolas P Suzor. 2017. Nudging robots: Innovative solutions to regulate artificial intelligence. *Vand. J. Ent. & Tech. L.* 20 (2017), 385.

[12] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001* (2023).

[13] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[14] Michael Katell, Meg Young, Bernease Herman, Dharma Dailey, Aaron Tam, Vivian Guetler, Corinne Binz, Daniella Raz, and PM Krafft. 2019. An algorithmic equity toolkit for technology audits by community advocates and activists. *arXiv preprint arXiv:1912.02943* (2019).

[15] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[16] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.

[17] Stefan Larsson. 2020. On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society* 7, 3 (2020), 437–451.

[18] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch und Computer 2023*. 81–90.

[19] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

[20] Q Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-driven design process for explainable AI user experiences. *arXiv preprint arXiv:2104.03483* (2021).

[21] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* Special Issue 5 (may 31 2024). https://hdsr.mitpress.mit.edu/pub/aelql9qy.

[22] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2024. Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *Comput. Surveys* 56, 7 (2024), 1–35.

[23] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*. Springer, 231–248.

[24] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[25] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[26] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[27] Brent D Mittelstadt. 2019. AI ethics-too principled to fail? *CoRR* (2019).

[28] OECD. [n. d.]. The RAISE Corporate AI Policy Benchmarks. ([n. d.]). https://oecd.ai/en/catalogue/tools/raise-benchmarks Uploaded on Dec 14, 2023.

[29] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.

[30] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic impact assessments: a practical Framework for Public Agency. *AI Now* 9 (2018).

[31] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[32] Catharina Doria Susannah Shattuck, Ian Eisenberg. 2022. What Is AI Governance and Why Should You Care? Available at https://www.credo.ai/blog/cutting-through-the-noise-what-is-ai-governance (last accessed: 2025/01/22).

[33] Elham Tabassi. 2023. Artificial intelligence risk management framework (AI RMF 1.0). (2023).

[34] Araz Taeihagh. 2021. Governance of artificial intelligence. *Policy and society* 40, 2 (2021), 137–157.

[35] Margareth Theresia. 2024. Newly enacted law sets basis for nat'l development of AI. Available at https://www.korea.net/NewsFocus/policies/view?articleId=264071 (last accessed: 2025/01/23).

[36] Unknown. 2025. Top 8 AI Governance Platforms for 2025. Available at https://www.domo.com/learn/article/ai-governance-tools (last accessed: 2025/01/23).

[37] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–40.

[38] IBM Research AI FactSheets 360 Website. [n. d.]. AI Lifecycle Governance. https://aifs360.res.ibm.com/governance

[39] Bernd W Wirtz, Jan C Weyerer, and Benjamin J Sturm. 2020. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration* 43, 9 (2020), 818–829.

[40] Hazal Şimşek. 2025. Compare Top 25 AI Governance Tools: A Vendor Benchmark [2025]. Available at https://research.aimultiple.com/ai-governance-tools/ (last accessed: 2025/01/22).

A Screening Questions

We leveraged the following questions to screen out ineligible participants. Free-form responses were used to contextualize their multiple choice responses. The first and second authors reviewed and discussed free-form response to determine eligibility.

(1) What is your job title? (free-form)
(2) How much experience do you have working with generative AI governance tools and technology in your role?
   - No experience *(disqualify)*
   - Low experience *(disqualify)*
   - Some experience
   - Experienced
   - Very experienced
(3) How much experience do you have working with LLMs in your role?
   - No experience *(disqualify)*
   - Low experience *(disqualify)*
   - Some experience
   - Experienced
   - Very experienced
(4) Describe the general usage of Generative AI (or LLMs) in your company. (free-form)
(5) Select the statement that best represents your organizations' AI governance initiatives.
   - No AI lifecycle governance *(disqualify)*
   - Some AI policies available to guide AI lifecycle
   - Common set of metrics to govern AI lifecycle
   - Automated Foundation Models/LLM validation and monitoring
   - Fully automated generative AI lifecycle governance
(6) Which, if any, of these tasks are you involved in? Please select all that apply. [Abbreviated]
   - Runs evaluations following the deployment to monitor how the solution performs against the metric threshold set by the risk and compliance owner.
   - Tests the solution to determine whether it meets the goals that are stated in the AI use case.
   - Evaluate the performance (e.g., accuracy) of AI/ML model outputs.
   - None of the above *(disqualify)*
(7) Can you please explain your recent experience validating or monitoring large language models for AI governance? (free-form)
(8) Do you use any solution/tool for AI governance? If yes, what's the name of the tool? Briefly explain what your typical task is using the tool. (free-form)

B Semi-Structured Interview Protocol

Phase 1

(1) Goal: What are your goals/tasks for AI governance?
(2) Challenges: What challenges do you encounter in achieving your goal/doing your tasks?
(3) Needs:
   - What information do you need to explain/understand the LLM outcomes?
   - How do you address your needs currently?
(4) Tools:
   - What tool/solution do you use for AI governance, if any?
   - What are primary features in the tool? Are there any explainability features?
   - How well does the tool support your needs and what can be improved?

Phase 2

(1) Impression: What are your thoughts when you are looking at this screen?
(2) Needs: What kind of questions do you have about the model?
(3) Feedback:
   - What concerns do you have when you are looking at the screen?
   - If there is anything you would like to change about this screen?

Wrap-up

(1) General
   - How would you describe your overall experience?
   - What do you like or dislike about the prototype?
   - Is there anything you would like to change about the prototype?
(2) Value
   - How well does the prototype help you to conduct your AI governance tasks?
   - How do you see this impacting your governance work on LLMs?
(3) Other: Any additional comment or questions?