AI Governance Roundtable #2: Transparency and Explainability

How can and should organizations operationalize transparency/explainability? What needs to be understood by whom, when, and why?

*This is the **second** of a series of roundtables convened by AI Singapore for representatives from industry, government, and academia to discuss responsible AI. Such discussions are typically too narrow and too broad. Too narrow in that a few voices dominate the discussion – notably those in the United States and Europe, with China sometimes included. Too broad in that discussion is often limited to generalities and principles. This project aims to address both aspects of this problem, involving a wider set of stakeholders — in particular those from Southeast Asia — in more focused discussions of specific challenges in the application of Responsible AI to particular questions.*

Rapporteur: Chen Dawei, AI Singapore

Location: Google SG
Date: 23 February 2024

1 Introduction

The widespread and escalating integration of AI systems in decision-making processes across diverse sectors underscores the need for a good understanding of these systems. However, the complexity and black-box nature of certain modern AI systems makes them difficult to comprehend. Therefore, it is important to provide necessary and meaningful information about these complex AI systems to build user trust, ensure accountability, facilitate audit and oversight for regulatory compliance, and enhance further model improvement for developers.

Despite these needs, we remain in the early stages of determining **what** kind of information should be provided, to **whom** (i.e., users, deployers, regulators), **when** (i.e., ex-ante or ex-post), and **why**. To address this issue and derive policy and regulatory implications, AI Singapore convened this roundtable bringing together regulators, practitioners, and academics to delve into the subject—*how organizations can and should operationalize transparency/explainability*.

Before delving into specific problems and potential solutions, it is important to clarify the stakeholders discussed in this report. Our previous report addresses the responsibility for AI between developers (i.e., creators of AI models) and deployers (i.e., innovators utilizing developer-created AI models and interfacing directly with users of AI products). Semantically, developers and deployers could be the same or different organizations. Considering the entire development and deployment lifecycle, transparency and explainability obligations that developers owe to the deployers and regulators and those that deployers owe to users and regulators can differ. Therefore, we use the terms developers and deployers rather than organizations to avoid misunderstandings and confusion. Users refer to individuals interacting with deployer-designed AI products, while governments act as regulators, drafting legislation and policies impacting AI.

After being clear about the stakeholders, we firstly define transparency, explainability and their differences, followed by a discussion on the importance of embracing meaningful transparency and explainability for modern AI systems in this section.

1.1 Definitions

Since both transparency and explainability aim to provide essential and meaningful information about AI systems to stakeholders, some people might group these two **interconnected** terms together. For instance, the AI transparency and explainability principle from OECD advocates that AI actors should disclose meaningful information to (1) "foster a general understanding of AI systems, including their capabilities and limitations," (2) "make stakeholders aware of their interactions with AI systems," (3) "...enable those affected by AI systems to understand the outcome," (4) "...enable those adversely affected by an AI system to challenge its output". However, the roundtable discussion offers more nuanced definitions, differences and relationship between transparency and explainability.

AI transparency strives to ensure that AI systems are designed, developed and deployed in a manner that facilitates human **oversight** through **openness** to external scrutiny[1]. It emphasizes the accessibility of information regarding **essential aspects**, or even the entire lifecycle, of AI systems. This information includes notices of AI intervention (e.g., make users aware that they

[1] Chesterman, S., Gao, Y., Hahn, J. and Valerie, S., 2023. The Evolution of AI Governance. *Authorea Preprints*.

are interacting with AI, content is AI-generated, decisions are AI-assisted), proper details about data (e.g., sources, manipulation), implemented algorithms (e.g., models selection, training infrastructure), potential risks and limitations, mitigation and safety measures, and more. The **broad** focus of transparency aims to ensure a general understanding of AI systems for users and overall accountability and ethical compliance for regulators.

AI explainability aims to ensure that AI systems are designed, developed, and deployed in a manner that facilitates **interpretability by laypersons**. It centers on making users **understand the outcomes** (e.g., predictions, recommendations, decisions) generated by AI systems, emphasizing the "why" (e.g., logic, reasoning) behind the outcomes. For instance, providing insights into the factors contributing to an AI outcome could make users better understand the results and assist them in making informed decisions. The **specific** focus of explainability is beneficial for diagnosing errors and biases, as well as improving model performance.

Table 1. Differences Between Transparency and Explainability

	Focus	Information	Benefit
Transparency	A general understanding and proper openness of the entire AI system	Notice of AI intervention, data, algorithms, risks and limitations, etc.	Foster trustworthiness, overall accountability, ethical compliance
Explainability	The reasoning behind the outcomes generated by AI system	Logic explanation, factors' contribution, model visualizations, etc.	Foster trustworthiness, error diagnosis, model improvement

Table 1 summarizes the differences between AI transparency and explainability. Since transparency centers on a general and broad understanding of the entire lifecycle of AI systems, some people might also consider explainability as **an integral part** of transparency[2]. However, given their distinct focuses and purposes, we treat them as **two separated concepts** in this report. This differentiation helps us better identify the challenges and solutions needed to achieve meaningful transparency and explainability.

1.2 Why

Given the definitions and differences between transparency and explainability, we will separately discuss the importance of embracing each concept in the following subsections.

[2] Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K. and Kujala, S., 2023. Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. *Information and Software Technology*, *159*, p.107197.

1.2.1 Why AI Transparency

In the development and deployment lifecycle of AI systems, various stakeholders require different information for different purposes. This section discusses why AI transparency is important for each of three key groups: users, deployers, and regulators.

For **users**, transparency can foster trust and facilitate adoption. Firstly, a clear understanding of AI systems through transparent information can make users more comfortable adopting them. Secondly, it allows users to comprehend the potential risks associated with AI systems. Thirdly, it is important for attributing responsibility, understanding who is accountable for "machine-assisted" decisions, and providing avenues (e.g., communication channels, feedback mechanisms) for seeking remedies when adverse outcomes occur.

For **deployers**, first, transparency enables them to understand the entire development process, making it easier to detect and address potential problems before deployment. Second, it ensures accountability and responsibility discussed in our previous report for their actions, decisions, and any issues that arise, fostering responsible AI deployment. Third, it promotes open communication within the development and deployment team, facilitating more effective knowledge sharing.

For **regulators**, transparency is important in ensuring that AI systems comply with legal and regulatory standards. It allows regulators to evaluate whether AI systems are designed and implemented aligned with governance principles[3]. Additionally, it helps identify who is responsible for decisions made by AI systems, ensuring accountability and compliance.

1.2.2 Why AI Explainability

Compared to transparency, explainability focuses on making users understand the outcomes generated by AI systems. It serves two main purposes. First, it helps users **make informed decisions** by fully understanding AI-generated outcomes before making their own decisions (ex ante). Second, it allows users to understand and challenge the AI generated outcomes when **adversely affected** by them, thereby improving their human-AI collaboration strategy in the future (ex post).

Furthermore, explainability facilitates the **identification** and **correction of errors or biases** in AI systems. If an AI system makes a wrong or biased decision, explainability enables AI developers

[3] Singapore identifies 11 AI governance principles: transparency, explainability, repeatability/reproducibility, safety, security, robustness, fairness, data governance, accountability, human agency and oversight, inclusive growth, societal and environmental well-being.

and deployers to understand the reasoning behind it, which is important for addressing and fixing the issue.

For instance, imagine that Alice provides all necessary information and applies for a loan from a bank that uses an AI system to evaluate and make decisions on applications. Unexpectedly, Alice receives a notification of rejection with no further explanation. In this scenario, Alice does not know why her loan application was rejected (i.e., lack of explanation) and cannot take steps to improve her chances in the future (i.e., inability to improve). In the worst case, if the rejection of Alice's application is due to inherent biases in the AI system based on its training data, these biases may remain unaddressed and continue to affect decisions unfairly for other applicants. However, providing proper explanation to clarify why Alice's application was rejected (e.g., which factors contributed most to the decision) could mitigate these problems.

2 Understanding the Problem

Building on the previously outlined definitions, significance, and necessity of transparency and explainability, we identify the challenges associated with operating them in the practices to better understand why they could be issues before we propose possible solutions in the subsequent section.

2.1 Transparency

2.1.1 AI Complexity and Black Box

The increasing complexity of certain modern AI systems make it difficult to achieve meaningful transparency through providing proper and sufficient information. In addition, the black box problem makes it difficult or even impossible to make certain information transparent. They challenge what should be transparent to whom (i.e., diverse stakeholders), and across different use cases (i.e, various risk levels). Therefore, we will first discuss what should be transparent in the following subsection, followed by an examination of to whom different information should be directed. Next, we will discuss use cases across various risk levels. Finally, we will discuss the costs of transparency.

2.1.2 What Should Be Transparent?

At the outset, it's important to recognize that complete understanding is not always necessary. For example, doctors have correctly prescribed aspirin for centuries despite not fully understanding its mechanisms. However, it is important for patients to understand the guidelines, dosage, and precautions—akin to a medical prescription. Similarly, while full comprehension of AI systems is not always necessary, certain key aspects should be clear. As the

complexity and opaque nature of AI systems increases, it is difficult, or even impossible, for users, developers and deployers to fully understand the entire AI systems.

Given the inherent complexity and the black-box nature of certain AI systems, determining what should be transparent to stakeholders is a significant challenge. Historically, lengthy documentations (e.g., license agreements, privacy policies) have been used as tools for transparency because these systems were manageable and comprehensible. However, certain modern AI systems present a different challenge. They are so complex that full disclosure, such as complete source code disclosure, is **neither feasible** due to trade secrets **nor meaningful** since users are often **unwilling** to sift through extensive documentation or **unable** to grasp highly technical materials. Therefore, rather than simply making the entire code of the AI system public—offering transparency in a literal sense—more practical insights might include notices of AI intervention (e.g., make users aware that they are interacting with AI, content is AI-generated, decisions are AI-assisted), proper details about data (e.g., sources, manipulation), implemented algorithms (e.g., models selection, training infrastructure), potential risks and limitations, mitigation and safety measures, and more.

2.1.3 To Whom?

As mentioned in the introduction, different stakeholders–including AI users, deployers, and regulators–require varying levels of AI transparency for distinct purposes. Generally, developers are responsible for providing appropriate transparency to both deployers and regulators, while deployers should ensure proper transparency for users and regulators. In the following, we will discuss what aspects of AI systems should be transparent to users, deployers and regulators[4].

For users of AI, the goal is to have a **general understanding** of the entire process of the AI system to trust it, feel comfortable interacting with it, and know how to find potential communication channels and supporting documentation if something unexpected occurs. Therefore, deployers should provide users with clear and concise information about notices of AI intervention, data sources, model basics (e.g., algorithms, capabilities), usage policy, risks and limitations, and feedback mechanisms, all in plain and easily understandable language, as many users of AI are not AI experts.

Deployers aim to have a **deeper understanding** of AI models than general users before deploying them. Additionally, deployers may need to fine-tune or retrain AI models with domain-specific proprietary data. Therefore, developers should provide deployers with detailed information about data (e.g., sources, access, manipulation), model basics (e.g., algorithms, size, components, structure, capabilities, updates), risks and limitations, usage policy, and feedback mechanisms.

[4] A useful resource is the foundation model transparency index created by Stanford researchers, which identifies and classifies 100 transparency indexes into 13 major dimensions.

Regulators focus on **ethical and regulatory compliance**. Developers and deployers should provide regulators with information about data compliance (e.g., privacy, copyright), model performance compliance (e.g., fairness, non-discrimination, safety, security), risks and limitations.

Table 2 summarizes the goals and transparency requirements for the three stakeholder groups discussed earlier.

Table 2. Transparency Requirements from Three Stakeholder Groups

Stakeholders	Goals	What Should Be Transparent
Users	A general understanding of AI systems before adoption	Notices of AI intervention, data basic, model basic, risks and limitations, usage policy, feedback mechanisms
Deployers	A deep understanding of AI systems before deployment	Data details, model details, usage policy, data, risks and limitations, feedback mechanisms
Regulators	Audit AI systems for legal or ethical compliance	Data compliance, model performance compliance, risks and limitations

2.1.4 Risk Level

Apart from considering what information should be provided to whom, the type and extent of transparency also vary in terms of risk level. AI systems that present low or negligible risks should prioritize stability, and a minimal level of transparency may be adequate. In contrast, AI systems that significantly affect human rights, safety, or security must adhere to strict controls, ensuring a high degree of transparency.

2.1.5 Costs

Transparency comes at costs. There exist trade-offs between transparency and other AI ethical principles. For instance, disclosing too much information might compromise **security** and **privacy** because hackers could potentially reverse-engineer the data used in training the model. Additionally, such disclosures risk exposing **intellectual property** or trade secrets, as adept hackers might reconstruct the model itself. Furthermore, the **administrative** and compliance costs associated with implementing transparency can be significant for developers and deployers. These costs could add another layer of complexity to the ethical development and deployment of AI technologies.

2.2 Explainability

2.2.1 AI Complexity and Black Box

Apart from **not being transparent**, the increasing complexity and black-box nature of certain AI systems can lead to an **inability to explain**. While some underlying algorithms, such as those used in Alice's loan applications might be explainable, certain modern AI algorithms, especially those based on deep neural networks, are inherently unexplainable. This makes it difficult—even impossible—for developers, deployers and users to understand how the AI system functions and why it may fail in some cases. This complexity complicates efforts to achieve meaningful explainability.

2.2.2 What Should Be Explained?

A general and broad understanding provided by transparency is sometimes insufficient for naturally complex and opaque AI systems. Users are entitled to explanations of adverse decisions, where explainability becomes important. For instance, reconsider Alice's case: the AI actor (i.e., the bank) could provide a detailed report explaining which factors influenced its decision, allowing Alice to understand why she was rejected and giving her the option to request a human review of the AI's decision, thus improving her profile for future applications.

Providing a breakdown of contributing factors is not the only way to achieve explainability for complex AI models. The fundamental goal is to **help users understand the outcomes generated by AI systems**. Apparently, simply sending all model parameters or model details does not help. Deployers should clearly explain how the outcomes are reached, what datasets are being used, the importance of each attribute, the potential risks and limitations.

Explanations can be either global or local. A **global explanation** aims to clarify the AI model's behavior for the entire dataset, providing insights into the model's logic and structure, evaluating its generalization, and allowing comparison with other models. On the other hand, a **local explanation** focuses on the AI model's behavior for a specific instance, identifying potential errors or biases and offering feedback for improvement.

2.2.3 Tradeoff Between Explainability and Accuracy

The tradeoff between explainability and model performance (i.e., **accuracy**) is well debated and documented. While models like decision trees and regression models offer significant explainability, they often tend to be less accurate. On the contrary, some models are hard to be explained, or even not explainable, but they are much more accurate than decision trees and regression models.

2.2.4 Risk Level

One the one hand, the risk-based approach to regulation demands greater explainability for high-risk AI applications. On the other hand, there exists a tradeoff between explainability and accuracy. It appears that we are constrained to deploying only those models that are less accurate but more explainable in high-risk use cases. In essence, a strict requirement for explainability could narrow our choices to simpler models like decision trees, potentially excluding more powerful AI models in the high-risk use cases. This tradeoff underscores the importance of a "**Goldilocks**" regulation—neither too restrictive nor too lenient–to effectively balance these considerations.

3 Possible Solutions

Considering the challenges outlined, the roundtable discussion highlights that **no single approach** can universally address the need for transparency and explainability. All stakeholders, including regulators, developers, deployers, and academics, play important roles in establishing meaningful transparency and explainability tailored to their target audiences across various use cases. In this section, we explore what these three groups of stakeholders can act to achieve meaningful transparency and explainability for modern AI systems.

3.1 Developers and Deployers

3.1.1 Motivations

The concept of the “invisible hand” suggests that self-interested individuals may inadvertently promote the general welfare of society in a free market. Advocates of this market-based approach argue that market forces alone can lead developers, deployers and users to achieve efficient and meaningful AI transparency and explainability without the need for regulatory intervention. This approach might be effective for several reasons. Firstly, developers and deployers, driven by competition and reputational concerns, might self-regulate and voluntarily adopt appropriate levels of AI transparency and explainability. Secondly, deployers might also be proactive in meeting their users’ needs. Thirdly, deployers typically understand their target users’ needs better than others, sometimes even including users themselves. Fourthly, users’ requirements for transparency and explainability vary widely. Similar to attitudes toward privacy, some users prioritize convenience over transparency and may not value detailed explanations of AI processes.

3.1.2 Audiences-centric

The first question developers and deployers need to answer when deciding how to implement transparency and explainability is: *what should be transparent and what should be explained to*

their target audiences. Apart from what we have discussed in Table 2, they should also consider: (1) who are your target audiences? (2) what do your audiences **need to know** according to the regulatory requirements? (3) what do your audience **want to know** to gain trust, promote adoption, and make informed decisions?

3.1.3 Interface Design

Based on the answers to the first question, developers and deployers need to consider the second question: *how to **display** the information to their audiences*. We can learn from the evolution of privacy policies. Instead of presenting length documents directly, it might be more reasonable and effective to highlight and visualize key information to their audiences, similar to privacy labels and cookie consents. Nonetheless, access to the full documentation and the availability of communication channels and feedback mechanisms remain essential.

3.1.4 Algorithmic Explainers

The discussion highlights various tools that generate post-hoc explanations on certain complex AI models (e.g., deep neural networks), including AI Verify[5], SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), What-if Tool, AI Explainability 360, among others. These tools enable users to understand the outcomes generated by complex AI models. However, explaining those emerging and complex generative AI models proves to be more challenging compared to traditional AI models. As mentioned earlier, providing factor contributions is not the only way to achieve explainability. For simpler AI models, such as regressions and decision trees, visualization can be more effective.

3.1.5 Open Source

The discussion also touched on the open-source approach as a method for fostering collaboration and promoting AI transparency. While this approach might appear to be an "ultimate" solution by offering complete access to all information (i.e., the source code) which allows **external experts** to review and verify the workings and claims of transparency, it often only provides the **illusion of transparency**. In reality, it does not necessarily provide meaningful transparency or explainability, especially for **laypersons**. This is because most people are not equipped to understand the intricacies of machine-assisted decisions buried within thousands of lines of source code. Additionally, there is a concern that open source LLMs could be **misused**, as malicious actors may exploit their full access to re-engineer models for nefarious purposes.

[5] Apart from providing explainability, AI Verify is in fact a broader AI governance testing framework and toolkit consisting of 11 AI ethics principles and corresponding testable criteria and testing process.

3.2 Regulators

3.2.1 Motivations

The market-based self-regulated approach has potential shortcomings. Firstly, the **i**mbalance of **market power** between deployers and users might lead deployers to take advantage of users' cognitive limitations. Secondly, the long-term impacts of AI-assisted decisions may be neglected. For example, while content recommendation systems can offer personalized content in the short term, they might also lead to **echo chambers** over time. These long-term effects, potentially detrimental to both individuals and society, may not align with the immediate interests of deployers and might be overlooked by users. In such cases, regulatory intervention becomes essential to address these issues.

3.2.2 "Goldilocks" Regulation

The discussion emphasizes the need for "Goldilocks" regulation—neither too restrictive nor too lenient. The debate begins with the question of whether it is appropriate to regulate AI transparency and explainability at this **early stage** of AI adoption. Additionally, it explores what aspects should be regulated and the extent of regulatory oversight necessary to achieve desired outcomes. The call for a standardized regulatory framework is also a point of contention. Regulation is seen as an additional layer of safety, ensuring that industry members adhere to established standards.

Regulators are encouraged to collaborate closely with the industry to standardize disclosures effectively. However, it is acknowledged that the disclosing parties have the deepest understanding of their target audiences. Therefore, these parties should have considerable discretion in determining the most useful forms of transparency and explainability for their contexts and target audiences. Moving forward, it would be beneficial for regulators to establish standards for transparency and explainability across different AI applications, as a **one-size-fits-all approach is not feasible**. These standards should take into account the trade-offs between transparency, explainability and costs discussed in Section 2.

Many countries have identified transparency and explainability as key principles in their white paper and developing regulations. The European Union's AI Act, the first of its kind, adopts a risk-based approach that mandates higher transparency for high-risk AI systems as outlined in Article 13. Additionally, Article 50 specifies the transparency obligations for providers and users of certain AI systems.

3.2.3 Assistive Tools (Optional)

Beyond high-level regulatory frameworks, regulators could offer specific assistive tools to help developers and deployers achieve meaningful AI transparency and explainability. For instance,

IMDA Singapore collaborates with companies from various sectors and scales to develop the AI Verify toolkit, which helps developers and deployers test their AI systems against 11 AI ethical principles. Moving forward, a **standardized checklist** could be a helpful, convenient and time-saving tool for developers and deployers, especially for small and medium enterprises (SMEs), to ensure compliance.

3.3 Academics

Academics can also contribute to the development of meaningful transparency and explainability. Further research is needed to determine **what types of information** should be provided and how best to **display** it to support human decision-making such that users could have a clear and better understanding of AI systems and make informed decisions. This research should encompass both high-level theoretical perspectives and specific interface designs. Deployers could collaborate with academia to conduct field experiments to understand, design, and validate improved human-AI interactions.

4 Conclusions

The widespread adoption of AI systems to assist human decision-making has increased the need for transparency and explainability. However, the inherent complexity and opacity of AI systems make it challenging for organizations (e.g., developers and deployers) to provide meaningful transparency and explainability. Additionally, the diverse backgrounds of target audiences and various use cases further complicate this task. Organizations also face trade-offs between transparency/explainability, and factors such as accuracy, privacy, trade secrets, and administrative costs.

Given these challenges, this roundtable explores what three groups of stakeholders can act to achieve meaningful transparency and explainability for modern AI systems. Regulators could collaborate with other stakeholders to propose a "Goldilocks" regulatory framework and provide assistive tools for organizations to comply with regulatory requirements. Organizations should adopt an audience-centric approach, developing strategies for transparency and explainability that are customized to their target audiences and specific use cases. Academics could contribute by researching the appropriate types of information and the best methods of presentation to achieve meaningful transparency and explainability.

Funding: This roundtable was founded via a charitable grant from Google.org, as part of Google's Digital Futures Project.


Rachel Teo (Presenting)
AI Singapore
2nd AI Governance Cross-Sector Roundtable
Google Cloud