Transparency and Explainability

Munongedzi Mabhoko, Clarkson University, mabhokm@clarkson.edu

Introduction

Transparency and explainability have become central themes in debates about the governance of modern artificial intelligence systems. Public expectations of algorithmic accountability have grown in response to high profile failures, rising regulatory scrutiny, and widening awareness of the covert ways algorithms shape everyday life. Academic literature increasingly recognizes that the challenges of explainability are not limited to model architecture or interpretability techniques. They reach further into the social, political, and infrastructural assumptions embedded in data pipelines, model training procedures, and institutional decision making practices. The "Open Data Trojan Horse" framing argues that the rhetoric of openness and democratized data access often obscures hidden mechanisms through which opacity, surveillance, and extractive data practices become normalized under the banner of transparency.

Meanwhile, technical literature on explainability, interpretability, and model verification shows gaps in our ability to audit modern black box systems, especially under adversarial or high stakes conditions. Research such as TAD, Trigger Approximation based Black box Trojan Detection for AI reveals how models can contain dormant malicious behaviors that remain undetected even when explainability methods appear to provide insight. This paper reflects on these dual strands of scholarship and examines how they complicate contemporary assumptions about AI transparency. I argue that meaningful explainability requires moving beyond visible model behavior and toward structural transparency about dataset lineage, model provenance, institutional incentives, and the broader governance ecosystems that enable hidden risks to emerge.

Transparency as a Governance Ideal and its Structural Contradictions

Transparency is often treated as a technical quality rather than a social and political condition. Many AI explainability frameworks equate transparency with opening model weights, showing attention maps, or exposing gradient based explanations. Yet the "Open Data Trojan Horse" critique shows that transparency initiatives that focus on surface visibility can distract from deeper asymmetries of power embedded in data infrastructures.

Open data programs frequently promise democratized access to information. They promote the idea that visibility encourages accountability. But practical deployments often reinforce the opposite. Open data can enable extraction without meaningful reciprocity. It can allow

corporations, researchers, and governments to build systems that individuals have little power to contest. Even when datasets are openly available, the assumptions baked into data collection, labeling, preprocessing, and curation remain obscure. These upstream decisions shape model behavior far more than downstream explainability techniques can reveal.

In this context, transparency becomes a symbolic gesture rather than a substantive safeguard. It creates the impression of openness while concealing the very mechanisms through which harm emerges. As a result, transparency without structural governance reforms can function as a Trojan Horse that legitimizes opaque algorithmic systems while masking their risks.

Explainability and the Myth of Interpretive Sufficiency

Explainability research attempts to provide accessible rationales for model predictions. Techniques such as SHAP, LIME, GradCAM, feature attribution, and counterfactual explanations promise to translate opaque decision processes into human understandable terms. Yet scholars increasingly argue that explainability is limited by what a system is designed to show. The interpretive surface often conceals deeper vulnerabilities.

Several explainability papers highlight concerns such as:

- Explanations can be misleading or selectively faithful.
- Feature attribution methods reveal correlations but not causal logic.
- Counterfactuals depend on model internal structure that remains opaque to auditors.
- Explanations may provide a false sense of comprehension that undermines safety and accountability.

This literature reveals a fundamental tension. Explainability aims to bridge the gap between algorithmic reasoning and human understanding. But the gap is not purely computational; it is socio-technical. Explanations are shaped by training data, distribution shifts, institutional deployment conditions, and adversarial incentives. A model's surface level explanation may be logical while its internal decision boundaries reflect structural bias or hidden triggers that explanations fail to expose.
This limitation is especially clear when considering Trojaned or backdoored models.

Trojaned Models and the Failure of Conventional Transparency

The paper TAD: Trigger Approximation based Black box Trojan Detection for AI illustrates how easily harmful or malicious behaviors can be embedded in state of the art models even when their outputs appear explainable. Trojaned models behave normally in standard conditions but

activate harmful outputs when exposed to a specific trigger pattern. These triggers can be imperceptible, semantically meaningless, or visually subtle.

TAD presents a black box Trojan detection method that reconstructs potential triggers by analyzing the sensitivity of model outputs to synthetic perturbations. The approach does not rely on model internals or parameter access. It infers the existence of hidden behaviors by searching for minimal patterns that disproportionately influence predictions.

This research exposes an important flaw in dominant transparency assumptions. A model can present highly interpretable explanations, consistent attention maps, and stable feature attributions while simultaneously containing hidden behaviors that explanations cannot reveal. Explanations faithfully represent the benign operational mode, not the malicious or adversarial one. Thus transparency that relies on model behavior under normal conditions is structurally insufficient for safety critical systems.

Trojan detection research highlights that transparency must include adversarial awareness. It must treat models as potentially compromised artifacts, not static mathematical functions. Without this shift, explainability tools risk reinforcing trust rather than uncovering vulnerabilities.

The Open Data Trojan Horse and Hidden Risks in Data Pipelines

The metaphor of the “Trojan Horse” is equally applicable to the datasets used to train modern AI systems. The belief that publicly available datasets promote fairness and reproducibility overlooks the structural inequities built into data generation processes. Opening data does not erase the historical or political conditions under which the data was created.

Combining this with Trojan detection literature reveals an even deeper challenge. Openness can accelerate the propagation of poisoned datasets, subtle triggers, or latent vulnerabilities across research pipelines. Models trained on publicly available datasets are often treated as trustworthy by default. But as recent backdoor attacks demonstrate, poisoning can occur at any stage, including during data scraping, annotation, or preprocessing.

The transparency discourse assumes that more visibility leads to more accountability. But visibility is often partial and strategically curated. True transparency requires the ability to interrogate provenance, to audit data lineage, and to question the power relations that determine what becomes visible and what remains hidden.

Toward a Holistic Framework of Explainability and Structural Transparency

To address these limitations, explainability must move beyond interpretability of model outputs. A meaningful framework should integrate:

1. Provenance based transparency
   Understanding the origins, authorship, and assumptions of datasets and models.
2. Adversarial transparency
   Recognizing that models can be manipulated or compromised, and incorporating Trojan detection and red teaming into standard governance practices.
3. Infrastructural transparency
   Examining institutional incentives, version control processes, access controls, and auditability mechanisms across the full ML pipeline.
4. Socio technical explainability
   Providing explanations not only for predictions but also for how algorithmic decisions distribute benefits and burdens across social groups.
5. Policy aligned documentation
   Using datasheets, model cards, transparency reports, and regulatory disclosures as operational artifacts, not symbolic gestures.

These principles shift transparency from a visual metaphor to a structural one. Instead of asking what a model shows, the question becomes what the system reveals about its origins, risks, and governance environment.

Discussion

The combination of open data critiques and Trojan detection research reveals a fundamental truth about contemporary AI governance. Transparency is necessary but insufficient. Explainability provides intuition but not assurance. Opening data provides access but not accountability. Black box auditing provides warning signals but not root cause explanations.

The future of responsible AI depends on integrating these fragmented approaches into a cohesive governance ecosystem. Systems must be transparent not only in their outputs but in their construction. They must be explainable not only in model logic but in institutional logic. And they must be auditable under both normal and adversarial conditions.

The Open Data Trojan Horse shows that transparency initiatives can unintentionally mask harm. TAD shows that explainability tools can fail to expose dangerous backdoors. Together, they reveal that AI safety requires more than technical interpretability. It requires structural reforms, governance design, and an expanded conception of what transparency demands.

Transparency and explainability in AI must evolve beyond simplistic assumptions about visibility and interpretability. The Open Data Trojan Horse highlights the political and structural dimensions of transparency, while TAD demonstrates the technical fragility of explainability in black box systems. Together, they show that contemporary AI systems operate within layers of hidden risk that explanations alone cannot reveal. A meaningful transparency agenda must treat AI systems as socio-technical infrastructures shaped by data provenance, institutional incentives, adversarial threats, and governance mechanisms. Only then can explainability function as a tool for accountability rather than a veneer of trust.