SSGE Research Series · Frontier Note 004

Why Explainability Is Different from Evidence

Version 1.0

June 2026

Approx. 7 min read

A research reflection on explainability, evidence, and trustworthy AI.

An explanation may persuade. Evidence can be verified.

Abstract

Explainability and evidence are often discussed as if they describe the same idea. This note explores why they may represent fundamentally different functions in trustworthy AI.

Core Question

As artificial intelligence becomes increasingly capable, a growing number of methods attempt to explain how models reach their conclusions. Feature attribution, Attention visualization, Counterfactual analysis, Natural-language explanations. Together, these approaches seek to answer an important question:

Why did the model produce this decision?

But another question is becoming equally important.

What actually happened during execution?

Are these two questions asking for the same thing?

Or are explanation and evidence fundamentally different?

Why This Matters

The terms *explainability* and *evidence* are frequently used as if they were interchangeable. In practice, they serve different purposes. An explanation helps people understand. Evidence helps people verify. An explanation may describe a possible reasoning process. Evidence preserves an observable execution process.

As AI systems increasingly participate in decisions with scientific, economic, legal, and societal consequences, this distinction becomes increasingly important. Understanding a decision is valuable. Verifying a decision may become indispensable.

Current Perspective

Explainability has become one of the central research topics in trustworthy AI.

Researchers continue to develop methods that reveal which inputs influenced predictions, visualize internal representations, or generate human-readable descriptions of model behavior. These advances provide important insights into increasingly complex systems.

Yet explanations remain, by their nature, interpretative. They reconstruct understanding after execution has already occurred. Different explainability methods may produce different interpretations of the same decision. This does not diminish their value. It simply reflects their purpose. Explanation seeks understanding. It does not necessarily preserve execution.

SSGE Perspective

Perhaps evidence should no longer be regarded simply as another form of explanation. Rather than describing why a decision might have occurred, evidence attempts to preserve what actually occurred during execution.

Execution traces, Runtime state transitions, Operational artifacts, Decision records. These are not explanations of execution. They are components of execution itself. An explanation may help us understand a completed decision. Evidence may allow us to reconstruct the operational process that produced it.

One is interpretative. The other is executable. One may persuade. The other can be independently examined. These two ideas are complementary rather than competing. Future trustworthy AI systems may ultimately require both.

Open Questions

Should AI governance prioritize explanations or evidence?

Can evidence remain meaningful even when no human-readable explanation exists?

Can execution evidence become independently verifiable across different AI architectures?

Might future trustworthy systems distinguish more clearly between understanding decisions and reconstructing them?

These questions remain open.

Takeaway

Explainability helps us understand intelligence. Evidence helps us trust intelligence. Perhaps trustworthy AI will require both—

but they should no longer be treated as the same concept.

This note reflects an ongoing exploration rather than a finalized theory. Future observations may refine, extend, or challenge the perspectives presented here.