InterXAI White Paper

Reimagining Explainable AI through Intertextual Critique

Version 1.0 – July 2025

Executive Summary

InterXAI is a human-centered platform that bridges Explainable AI (XAI) and intertextual analysis. It empowers users to compare machine-generated model explanations with human interpretations, fostering critical engagement, accountability, and enriched understanding of AI behavior. Combining computational linguistics, literary methods, and digital humanities, InterXAI offers a hybrid framework for interpretability and critique in machine learning systems.

---

1. Vision & Purpose

Modern AI systems are increasingly opaque, and existing XAI tools often center the machine's perspective. InterXAI flips the lens. We ask not just how a model works—but how its outputs are interpreted, challenged, and reframed by human users.

By drawing from fields like hermeneutics, reader-response theory, and intertextuality, InterXAI offers a method to analyze, visualize, and question AI decisions in light of cultural, ethical, and narrative frames.

---

2. Core Features

XAI Engine

- Integrates SHAP, LIME, and other ML explanation libraries
- Provides model-level explanations of classification, regression, and generative outputs
- API-ready and extendable to vision and multi-modal models

Human Annotation Engine

- Allows annotators to critique, contextualize, or reframe model outputs
- Supports tagging of metaphors, references, omissions, and alternative interpretations
- Tracks consensus or divergence across multiple readers

Intertextual Linker

- Links model outputs to external corpora (e.g., Wikipedia, classic texts, media archives)
- Highlights allusions, reused phrasing, or conceptual drift
- Integrates with existing NLP pipelines for text similarity and citation detection

Comparison Orchestrator

- Merges machine and human insights side-by-side
- Detects gaps between technical explanations and human critiques
- Generates visual dashboards and exportable insights

---

3. Use Cases

- **Education**: Teaching AI literacy and critical thinking via annotated model outputs
- **Digital Humanities**: Analyzing AI-generated interpretations of historical or literary texts
- **Media & Ethics**: Auditing content recommendation or moderation explanations
- **NLP Research**: Comparing generated text explanations with scholarly commentary
- **Policy & Compliance**: Visualizing accountability in high-stakes automated decisions

---

4. System Architecture

- **Input:** Literary, historical, social media, or user-provided texts
- **Processing:** XAI engine + human-centered annotation layer
- **Integration:** LLMs, SHAP/LIME, Intertextual APIs
- **Export:** PDF/HTML/JSON formats for scholarly, public, or audit purposes

Deployment available via GitHub Pages + Netlify frontend. A companion Streamlit dashboard powers real-time analysis and comparison.

---

5. Technical Stack

- Python, Streamlit, SHAP, LIME, Hugging Face Transformers
- Jekyll, GitHub Pages, Netlify (UI & publishing)
- Integration-ready with spaCy, Gensim, Wikidata APIs

---

6. Contribution & Collaboration

InterXAI is an open, evolving project. You can:
- Submit a case study
- Collaborate with us
- Explore existing blog posts and case studies

---

7. Future Roadmap

- Inter-Annotator Analysis Module
- Visualization of Intertextual Threads
- Dataset Builder for XAI + Human Commentary
- Dashboard for Influence Mapping

---

8. Licensing & Ethics

InterXAI embraces the values of transparency, interpretability, and scholarly openness. Code is released under MIT License, with research outputs licensed as Creative Commons (CC-BY-SA). We uphold ethical AI development and inclusive critique.

---

9. Contact

Project Lead: Felix B. Oke
bfiliks4xt@gmail.com
GitHub | LinkedIn

Website: https://interxai.netlify.app