Abstract:

The abstract underscores the imperative for Explainable AI (XAI) owing to the pervasive impact of AI systems in critical domains, especially in decision-making and individual rights. It advocates for a multidisciplinary strategy to confront these challenges, incorporating technical, philosophical, social, and psychological facets. The intricacy of conveying AI reasoning to non-experts is emphasized, necessitating foundational models and a comprehensive theoretical framework. The translation of intricate AI-generated explanations into understandable language is crucial for bridging the divide between AI reasoning and human comprehension. Surrogate models, such as expert systems, have the potential to enhance the understanding of complex AI reasoning among broader audiences. The abstract urges the establishment of a robust theoretical foundation spanning diverse disciplines to effectively address the intricacies of Explainable AI in contemporary socio-technical landscapes.

Introduction:

The introduction explores the role of Artificial Intelligence (AI) in decision-making processes, with a focus on its impact on individual rights and societal well-being. It distinguishes between automated decisions and authentic AI systems, drawing attention to the misapplication of the term "AI" and associated misconceptions. The complexity of Explainable AI (XAI) is discussed, underscoring the challenges of rendering complex AI reasoning accessible to those lacking technical expertise. The introduction categorizes complex AI mechanisms into varying levels of explainability, highlighting the complexities in articulating AI decision-making in societal domains and public opinion. Emphasis is placed on transparency as a means to uphold public confidence in the performance and validity of AI systems. The introduction advocates for a balanced approach that integrates technical understanding with human comprehension in intricate socio-technical landscapes.