International Journal of Natural Computing Research
Volume 11 • Issue 1

Insights Into Incorporating Trustworthiness and Ethics in AI Systems With Explainable AI

Meghana Kshirsagar, University of Limerick, Ireland*
https://orcid.org/0000-0002-8182-2465

Krishn Kumar Gupt, Technological University of the Shannon, Athlone, Ireland
https://orcid.org/0000-0002-1612-5102

Gauri Vaidya, University of Limerick, Ireland

Conor Ryan, University of Limerick, Ireland

Joseph P. Sullivan, Technological University of the Shannon, Athlone, Ireland
https://orcid.org/0000-0003-0010-3715

Vivek Kshirsagar, Government Engineering College, Aurangabad, India

ABSTRACT

Over the past seven decades since the advent of artificial intelligence (AI) technology, researchers have demonstrated and deployed systems incorporating AI in various domains. The absence of model explainability in critical systems such as medical AI and credit risk assessment among others has led to neglect of key ethical and professional principles which can cause considerable harm. With explainability methods, developers can check their models beyond mere performance and identify errors. This leads to increased efficiency in time and reduces development costs. The article summarizes that steering the traditional AI systems toward responsible AI engineering can address concerns raised in the deployment of AI systems and mitigate them by incorporating explainable AI methods. Finally, the article concludes with the societal benefits of the futuristic AI systems and the market shares for revenue generation possible through the deployment of trustworthy and ethical AI systems.

KEYWORDS

Agriculture, Black-Box Approach, Decision Support Systems, Deep Learning, Healthcare, Lime, Machine Learning, PDP, Responsible Artificial Intelligence, Security, Smart City

INTRODUCTION

Artificial Intelligence (AI) has extended into a significant technological shift in the recent decades such that each industry has been empowered in increased productivity, intelligent solutions, automation,

DOI: 10.4018/IJNCR.310006

*Corresponding Author

optimization, etc. to name a few. With the introduction of the determinant of trust, AI can no longer be treated as a black-box model without a clear understanding of what is going on inside. As AI ethics pose the single largest challenge towards widespread deployment, a trustworthy AI framework can help companies to design, develop, and deploy AI systems that they can trust. Better policies to manage ownership of personal data through adopting regulatory like the General Data Protection Regulation (GDPR), one can easily overcome the inappropriate usage of data, possible from uncovering behavioural patterns through data mining. AI systems can indeed be made more trustworthy and responsible by making them more traceable and explainable for the prediction of outcomes and decisions. In this research work, we present an overview of the determinants of AI – trust, the recent market trends and factors that can lead to responsible AI and software engineering.

LITERATURE METHODOLOGY

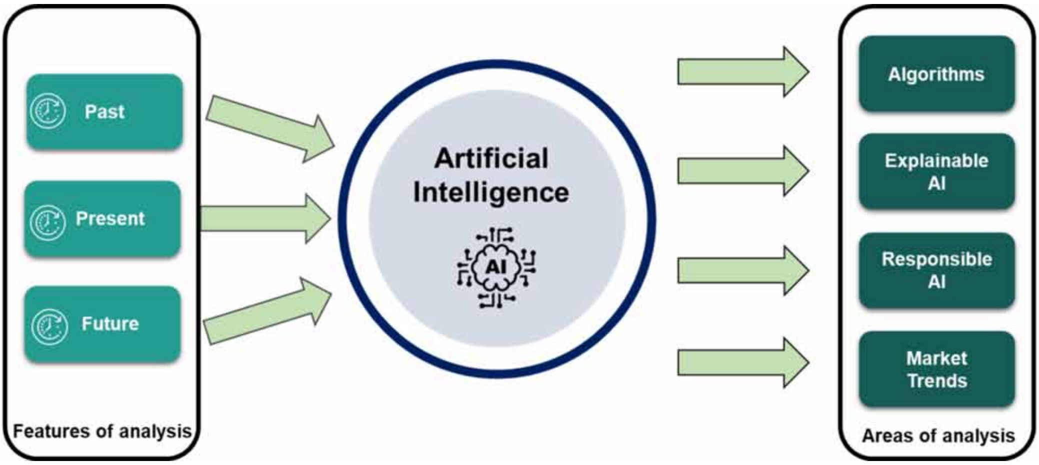
Our literature review is built by analyzing information across a range of sources. Figure 1 illustrates the pipeline of the proposed study where we integrate data based on information from the past, the current, and the predicted future for AI-powered applications. The objective of the study is to discuss the impact of AI-powered products on the wider community.

We present the evolution of AI, the Machine Learning (ML) algorithms in practice, applications in use, and the future of businesses and technologies with the integration of AI and its determinants like trust, big data and ubiquitous computing. We conducted a detailed study of all the ML and Deep Learning (DL) algorithms along with their use cases. We have an in-depth discussion on how incorporating explainability and interpretability into AI applications can lead to robust, trustworthy, fair and transparent AI systems. Finally, we bring to attention the importance of responsible AI engineering leading to regulated and accountable AI systems of the future.

The unique contributions of our proposed study are:

1. The evolution of AI and deep learning technology over the past seven decades;
2. Popular ML algorithms along with use cases drawn from diverse application domains;
3. Intelligent business models and market trends for industrial AI-powered products;
4. Incorporating Responsible AI for Trustworthy AI systems.

Figure 1. Pipeline of the proposed study


Past
Present
Future
Features of analysis
Artificial Intelligence
AI
Algorithms
Explainable AI
Responsible AI
Market Trends
Areas of analysis