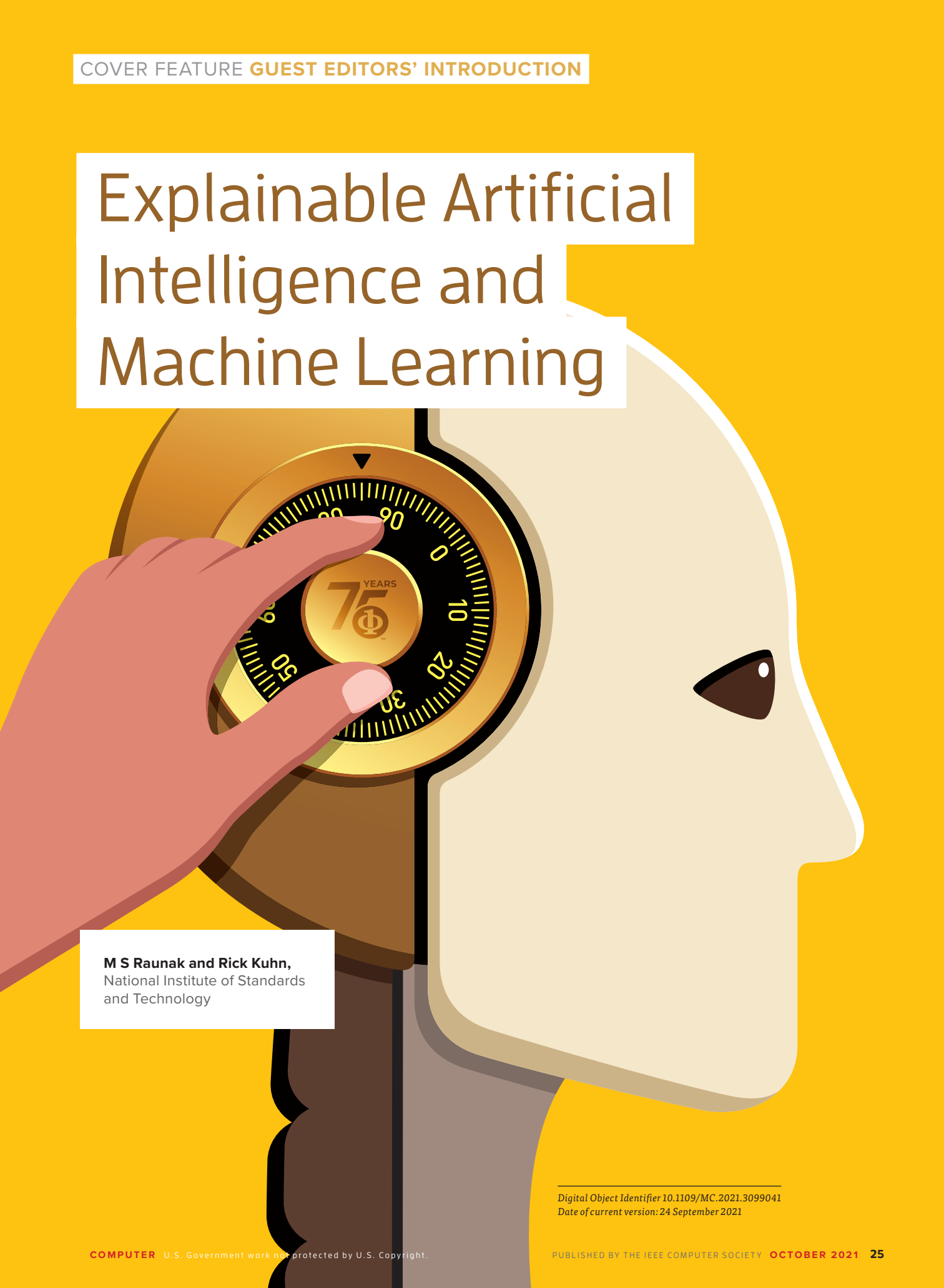


# Explainable Artificial Intelligence and Machine Learning



**M S Raunak and Rick Kuhn,**  
National Institute of Standards  
and Technology

Digital Object Identifier 10.1109/MC.2021.3099041  
Date of current version: 24 September 2021

*Explainable artificial intelligence (XAI) is a crucially important aspect of improving trust and reliability of artificial intelligent and machine learning (AI/ML) systems. These articles shed light on some of the important and useful aspects of XAI and its application.*

**F**rom medical diagnostics to financial fraud detection to Mars mission, applications of artificial intelligence and machine learning (AI/ML) systems are at an inflection point, where their mass adoption appears to be around the corner. Can stakeholders really trust these systems and rely on the decisions they make? One important ingredient needed to increase the trust and reliability of these opaque systems is to shed light on the process behind the decisions they produce, that is, to explain them. Explaining an AI/ML system's decision is a multifaceted problem. Is the explanation relevant, is it accurate, is it useful for the particular stakeholder? These are all crucial questions that need to be answered for achieving

articles here have looked into AI explanation from varying perspectives and for multiple groups of audience.

Progress in the field of XAI requires a thorough understanding of the kinds of errors that users make for different types of explanation systems. In "Explainable Recommendations and Calibrated Trust: Two Systematic User Errors," Naiseh et al. describe a methodology to understand "trust calibration" errors, in which users either overtrust, following incorrect recommendations, or undertrust, rejecting correct conclusions of AI systems. Evaluating how much trust to place in a decision by an expert, either human or machine, is always part of processes involving decision making. But the factors users can rely on are different

overview from a global perspective of the model and were comfortable with complex explanations when properly structured. Explanations can be presented from multiple viewpoints, and this structured approach has a positive correlation with trust in the system. Since trust is a basic requirement for users, the multiperspective framework can be of significant value in increasing usage in the field of AI-supported investment advising.

In "Explainable Machine Learning for Fraud Detection," Psychoula et al. illustrate the importance of weighing tradeoffs in explainability when used in real-time systems. In particular, the "best" explanation may not always be most effective for a task when speed is important, and the authors show that it may be better to use a combination of XAI methods in some applications. An analysis of tradeoffs is, of course, fundamental in engineering, and this article demonstrates that engineering AI systems will involve this type of analysis as well. It is important for designers to understand the characteristics of different approaches to XAI to evaluate how each may be incorporated into practical applications. Like many other domains, one explanation approach is unlikely to fit all. Instead of seeking one "best" approach to XAI, designers must determine what technique is best for their particular applications and for their specific audience.

Srinivasan et al., in "Explainable AI for Chiller Fault-Detection Systems: Gaining Human Trust," consider an XAI application that is somewhat rare in consumer applications but common in industrial usage—the scenario of

**USERS EITHER OVERTRUST, FOLLOWING INCORRECT RECOMMENDATIONS, OR UNDERTRUST, REJECTING CORRECT CONCLUSIONS OF AI SYSTEMS.**

explainability. The explainable AI (XAI) research area is new and rapidly growing. Different approaches, techniques, and tools are being developed and explored toward achieving meaningful, accurate, and useful explanations.

In this special issue, we have included a diverse group of articles addressing different aspects of XAI in a variety of domains. From real-time systems to human-in-the-loop fault detection, the

for human versus AI experts, so progress in XAI will require understanding how user trust is calibrated.

In "Usability, User Comprehension, and Perceptions of Explanations for Complex Decision Support Systems in Finance: A Robo-Advisory Use Case," Deo and Sontakke consider how explanations can be organized to improve user understanding. For the application studied, users preferred a partial

applications that involve both human and machine experts interacting to solve a fault diagnosis problem. The authors show how XAI can be used to reduce fault detection time in building air-handling systems, when AI systems are used to assist human users who must physically inspect equipment, and who must also make decisions regarding the probability of a failure in the near future. The resulting integration of XAI improves the accuracy of fault impact assessment, reducing faults and improving planning for maintenance operations.

Another aspect of human integration with AI systems is provided by Da Poian et al. in “Science Autonomy and the ExoMars Mission: Machine Learning to Help Find Life on Mars.” Space research missions involve open-ended questions rather than binary failed/nonfailed determinations, and there may be no clear-cut decisions. In this application, the AI systems must recommend courses of action likely to yield useful data. The authors consider this problem of science autonomy, how to combine human and machine insights to not only answer questions but also to pose new ones.

In “Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI,” Holzinger and Müller present some new ideas related to measuring the quality of an AI’s explanation at real time, collecting human-generated feedback through sensors and utilizing them for updating and improving the explanation. The authors discuss the information flow between the AI decisions and the mental model of the human utilizing the decision. With a motivating

## ABOUT THE AUTHORS

**M S RAUNAK** is a computer scientist at the National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. His research interests include verification and validation of “difficult-to-test” systems such as complex simulation models, cryptographic implementations, and machine learning algorithms. Raunak received a Ph.D. in computer science from the University of Massachusetts Amherst. He is a Member of IEEE. He can be reached at raunak@nist.gov.

**RICK KUHN** is a computer scientist in the computer security division at the National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. His research focuses on combinatorial methods in software verification and testing and extending these methods for assurance and explainability in AI and machine learning. Kuhn received an M.S. in computer science from the University of Maryland, College Park. He is a Fellow of IEEE. He can be reached at kuhn@nist.gov.

**ONE IMPORTANT CONSIDERATION IS THE ROLE OF HUMANS IN THE EXPLANATION PROCESS AND HOW BEST TO INTEGRATE IT.**

example of histopathology, that is, the changes in tissues due to disease, the authors illustrate the use of human-AI interface in a medical diagnostic for better explainability.

**E**veryone agrees on the importance of explanation in AI and ML algorithms. However, achieving explainability is not

simple. One important consideration is the role of humans in the explanation process and how best to integrate it. There is unlikely to be one overall best approach for explaining different AI and ML systems. With the diverse group of articles in this special issue, we believe the readers will be treated with a range of perspectives emanating from different aspects of XAI. 