06.25

Computer

Data Storage, Knowledge Graphs, Software Testing, and Action Engines

IEEE
IEEE COMPUTER SOCIETY

vol. 58 no. 6

www.computer.org/computer

Get Published in the *IEEE Transactions on Privacy*

This fully open access journal is soliciting papers for review.

IEEE Transactions on Privacy serves as a rapid publication forum for groundbreaking articles in the realm of privacy and data protection. Submit a paper and benefit from publishing with the IEEE Computer Society! With over 5 million unique monthly visitors to the IEEE Xplore® and Computer Society digital libraries, your research can benefit from broad distribution to readers in your field.

Submit a Paper Today!

Visit computer.org/tp to learn more.

IEEE COMPUTER SOCIETY | IEEE

Digital Object Identifier 10.1109/MC.2025.3562047

Computer

JUNE 2025

FEATURES

Digital Object Identifier 10.1109/MC.2025.3557139

JUNE 2025

CONTENTS

ABOUT THIS ISSUE

DATA STORAGE, KNOWLEDGE GRAPHS, SOFTWARE TESTING, AND ACTION ENGINES

Articles explore various areas of technology.

FEATURES CONTINUED

Department

Membership News

COLUMNS

Circulation: *Computer* (ISSN 0018-9162) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036. IEEE Computer Society membership includes a subscription to *Computer* magazine.

Postmaster: Send undelivered copies and address changes to *Computer*, IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Canadian GST #125634188. Canada Post Corporation (Canadian distribution) publications mail agreement number 40013885. Return undeliverable Canadian addresses to PO Box 122, Niagara Falls, ON L2E 6S8 Canada. Printed in USA.

Computer

IEEE COMPUTER SOCIETY http://computer.org // +1 714 821 8380
COMPUTER http://computer.org/computer // computer@computer.org

EDITOR IN CHIEF

Jeffrey Voas
NIST, USA
j.voas@ieee.org

ASSOCIATE EDITOR IN CHIEF, INTERNET OF THINGS
Joanna F. DeFranco
The Pennsylvania State University, USA
jfd104@psu.edu

ASSOCIATE EDITOR IN CHIEF, COMPUTING PRACTICES
Vladimir Getov
University of Westminster, U.K.
v.s.getov@westminster.ac.uk

ASSOCIATE EDITOR IN CHIEF, COMPUTING'S ECONOMICS
Nir Kshetri
The University of North Carolina at Greensboro, USA
nbkshetr@uncg.edu

ASSOCIATE EDITOR IN CHIEF, SOFTWARE ENGINEERING
Phil Laplante
The Pennsylvania State University, USA
plaplante@psu.edu

ASSOCIATE EDITOR IN CHIEF, CYBERTRUST
James Bret Michael
Naval Postgraduate School, USA
bmichael@nps.edu

2025 IEEE COMPUTER SOCIETY PRESIDENT
Hironori Washizaki
Waseda University, Tokyo, Japan
washizaki@waseda.jp

AREA EDITORS

BIG DATA
Domenico Talia
University of Calabria, Italy

CLOUD COMPUTING
Schahram Dustdar
TU Wien, Austria

COMPUTING FUNDAMENTALS
Joanna F. DeFranco
The Pennsylvania State University, USA

CYBER-PHYSICAL SYSTEMS
Oleg Sokolsky
University of Pennsylvania, USA

CYBERSECURITY
Rick Kuhn
NIST, USA
M S Raunak
NIST, USA

EMBEDDED COMPUTING
Marilyn Wolf
University of Nebraska, USA

EMERGING TECHNOLOGIES
Irena Bojanova
NIST, USA
Elena Loukolanova
International Monetary Fund, USA

HIGH-PERFORMANCE COMPUTING
Vladimir Getov
University of Westminster, U.K.

INTERNET OF THINGS
Michael Beigl
Karlsruhe Institute of Technology, Germany

SECURITY AND PRIVACY
James Bret Michael
Naval Postgraduate School, USA

SOCIAL–PHYSICAL–CYBER SYSTEMS
Mike Hinchey
University of Limerick, Ireland

SOFTWARE ENGINEERING
Benoit Baudry
KTH Royal Institute of Technology, Sweden
Christof Ebert
Vector Consulting Services/University of Stuttgart, Germany
Phil Laplante
The Pennsylvania State University, USA

COLUMN AND DEPARTMENT EDITORS

AFTERSHOCK
Solom Heddaya
Heddaya Projects LLC, USA

ALGORITHMS
Doron Drusinsky
Naval Postgraduate School, USA

ARTIFICIAL INTELLIGENCE/MACHINE LEARNING
Hsiao-Ying Lin
Huawei France, France

COMPUTING ARCHITECTURES
Timothy Jones
University of Cambridge, U.K.
Robert Mullins
University of Cambridge, U.K.

COMPUTING'S ECONOMICS
Nir Kshetri
The University of North Carolina at Greensboro, USA

COMPUTING THROUGH TIME
Ergun Akleman
Texas A&M Univ., USA

CYBER-PHYSICAL SYSTEMS
Dimitrios Serpanos
University of Patras, Greece

CYBERTRUST
James Bret Michael
Naval Postgraduate School, USA

DATA
Norita Ahmad
American University of Sharjah, United Arab Emirates
Preeti Chauhan
Google, USA

EDUCATION
George Hurlburt
U.S. Federal Service (Retired), USA
Sorel Reisman
California State University, USA

GAMES
Michael Zyda
University of Southern California, USA

HUMANITY AND COMPUTING
Domenico Talia
University of Calabria, Italy

INDUSTRY INSIGHTS
Christof Ebert
Vector Consulting Services, Germany

INTERNET OF THINGS
Joanna F. DeFranco
The Pennsylvania State University, USA

IT INNOVATION
Mark Campbell
EVOTEK., USA

MEMORY AND STORAGE
Tom Coughlin
Coughlin Associates, USA

MICROELECTRONICS
Conrad James
Sandia National Laboratories, USA

OPEN SOURCE
Dirk Riehle
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

OUT OF BAND
Hal Berghel
University of Nevada, Las Vegas, USA

PREDICTIONS
Dejan Milojicic
Hewlett Packard Labs, USA

SOFTWARE ENGINEERING
Phil Laplante
The Pennsylvania State University, USA

SPOTLIGHT ON TRANSACTIONS
Antonio Mastropaolo
College of William and Mary, USA

STANDARDS
Jyotika Athavale
Synopsys, Inc., USA

50 & 25 YEARS AGO
Erich Neuhold
University of Vienna, Austria

ADVISORY PANEL

Carl K. Chang (EIC Emeritus), Iowa State University, USA
Sumi Helal (EIC Emeritus), University of Bologna, Italy
Keith Miller, retired, USA
Bill Schilit, Google, USA
George K. Thiruvathukal, Loyola University Chicago, USA
Ron Vetter (EIC Emeritus), University of North Carolina Wilmington, USA
Alf Weaver, University of Virginia, USA

IEEE COMPUTER SOCIETY

CS PUBLICATIONS BOARD

Charles (Chuck) Hansen (VP for Publications), Irena Bojanova, Greg Byrd, Sven Dickinson, David Ebert, Minyi Guo, Lizy K. John, Joaquim Jorge, Daniel S. Katz, Klaus Mueller, San Murugesan, Jaideep Vaidya;
Ex officio: Hironori Washizaki, Melissa Russell, Robin Baldwin

CS MAGAZINE OPERATIONS COMMITTEE

Lizy K. John (Chair), Bo An, Troy Kaighin Astarte, Jeffrey Carver, Sigrid Eldh, Fahim Kawsar, Hsien-Hsin Sean Lee, Charalampos (Babis) Z. Patrikakis, Sean Peisert, Balakrishnan Prabhakaran, Weisong Shi, Jeffrey Voas, Pak Chung Wong

COMPUTER STAFF

Journals Production Manager
Joanna Gojlik
j.gojlik@ieee.org
Cover Design
Patrick George
Peer Review Administrator
computer-ma@computer.org
Periodicals Portfolio Specialist
Priscilla An
Periodicals Operations Project Specialist
Christine Shaughnessy
Compliance Manager
Jennifer Carruth

Periodicals Portfolio Senior Manager
Carrie Clark
Senior Advertising Coordinator
Debbie Sims
Director of Periodicals & Special Projects
Robin Baldwin
IEEE Computer Society Membership Director
Erik Berkowitz
IEEE Computer Society Executive Director
Melissa Russell

IEEE PUBLISHING OPERATIONS

Senior Director, Publishing Operations
Dawn M. Melley
Director, Editorial Services
Kevin Lisankie
Director, Production Services
Peter M. Tuohy

Associate Director, Digital Assets & Editorial Support
Neelam Khinvasara
Senior Manager, Journals Production
Patrick Kempf

Digital Object Identifier 10.1109/MC.2025.3557141

Permission to reprint/republish this material for commercial, advertising, or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to the IEEE Intellectual Property Rights Office, 445 Hoes Lane, Piscataway, NJ 08854-4141 or pubs-permissions@ieee.org. Copyright © 2025 IEEE. All rights reserved. IEEE prohibits discrimination, harassment, and bullying. For more information, visit www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

How Artificial Intelligence Is Reshaping Our Lives: A Framework to Unravel Its Complexities

Antonio Mastropaolo, William & Mary

This article from IEEE Transactions on Software Engineering *introduces LUNA, a universal analysis framework designed to enhance the interpretability, reliability, and trustworthiness of large language models by addressing key challenges such as hallucinations, biases, and inconsistencies.*

Artificial intelligence (AI) refers to a collection of methodologies and techniques that allow machines to emulate human intelligence. One of the most transformative advancements in AI in recent years has been the emergence of large language models (LLMs), which are deep learning systems capable of understanding and generating human-like text. The success of these models hinges on two critical factors: the availability of vast amounts of data and the computational power to process it.

Data act as the foundation, embedding knowledge into the model, similar to how a child learns from their parents' experiences. Meanwhile, computational power, particularly through graphics processing units, provides the technical backbone that allows LLMs, the "learning child," to efficiently process and internalize this knowledge.

Digital Object Identifier 10.1109/MC.2025.3549482
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

EDITOR **ANTONIO MASTROPAOLO**
William & Mary, amastropaolo@wm.edu

The rapid growth in data availability and hardware capabilities has driven a revolutionary shift in AI, enabling the creation of increasingly advanced and efficient methods. As of 2 February 2025, these advancements continue to push the boundaries of AI, leading to breakthroughs in areas ranging from language understanding to multimodal learning and beyond.

Truth be told, data and computational power are often seen as the sole main drivers behind today's transformative technologies, while often times we forget about the fundamental scientific breakthroughs that acted as a catalyst for the AI revolution. A pivotal moment, or, more aptly, "the milestone," was the introduction of the transformer model,[a] proposed by Vaswani et al. in 2017.[5] This architecture revolutionized AI by enabling models to process and generate human-like text with unprecedented accuracy. Building on this foundation, a new wave of models has been advancing the automation of diverse practices and fields.[1,2,3] These developments not only enhance existing workflows but also open up new opportunities for innovation, driven by the ability of LLMs to generate accurate, contextually relevant, and human-like responses. Such innovations reflect the continuous evolution of the technology, paving the way for broader and more transformative applications in the future. And yet, despite the tangible and immediately observable improvements, fundamental questions remain unanswered. For instance, have you ever wondered, Would you trust an AI to manage your finances, or to prescribe your medication? These unresolved issues, particularly concerning the interpretability of LLMs' outputs, inevitably limit their applicability and potential benefits. In critical domains, where understanding why event A leads to event B can have consequences as significant as human lives, these limitations become even more pressing.

It is now clear that without addressing these challenges, the full value and impact of these advancements cannot be realized. Despite their remarkable capabilities, LLMs still operate as opaque systems, making their decision-making processes opaque and difficult to trust.

> The success of these models hinges on two critical factors: the availability of vast amounts of data and the computational power to process it.

This month's article tackles the critical yet underexplored aspects of LLMs' output consistency, through LUNA,[4] a universal analysis framework designed to evaluate LLMs in a flexible and human-interpretable way. At its core, LUNA enables the creation of an abstract model, a simplified representation of the original LLM. This abstraction serves as a powerful tool to break down and understand the complex inner workings of LLMs, which encode information in billions of parameters in the form of patterns. By addressing challenges such as hallucinations[b] and biases,[c] LUNA represents a significant step forward in improving the reliability and trustworthiness of LLMs, ensuring they can be safely and effectively applied across diverse real-world contexts.

By employing modeling techniques, the authors of this article aim to create an abstract representation of LLMs that simplifies and supports their analysis, enabling the assessment and mitigation of critical issues such as hallucinations, biases, and inconsistencies. This marks a major milestone as it lays the groundwork for future research and facilitates the creation of more dependable and trustworthy AI systems. With frameworks like LUNA and similar future approaches, the safe and effective application of LLMs across diverse real-world contexts becomes increasingly attainable, ensuring these groundbreaking technologies can be deployed with greater confidence and impact.

[a] A deep learning architecture defines how the elements of a network are structured and interact.

[b] Hallucinations refer to instances where LLMs generate confident but incorrect, nonsensical, or fabricated information, often due to gaps in training data or overgeneralization.

[c] Biases in LLMs arise when the models produce outputs that reflect unfair, skewed, or prejudiced perspectives, often inherited from the training data or societal patterns embedded in the data.

REFERENCES

1. D. Guo et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," 2025, *arXiv:2501.12948*.
2. "Github copilot: Your AI pair programmer." GitHub. Accessed: Feb. 15, 2025. [Online]. Available: https://copilot.github.com/
3. J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
4. D. Song et al., "LUNA: A universal analysis framework for large language models," *IEEE Trans. Softw. Eng.*, vol. 50, no. 7, pp. 1921–1948, Jul. 2024, doi: 10.1109/TSE.2024.3411928.
5. A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2027, pp 6000–6010.

ANTONIO MASTROPAOLO is an assistant professor of computer science at the Department of Computer Science, William & Mary, Williamsburg, VA 23185 USA. Contact them at amastropaolo@wm.edu.

ELSEWHERE IN THE CS

Computer Highlights Society Magazines

The IEEE Computer Society's lineup of 11 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

Computing in Science & Engineering

Building and Sustaining a Community Resource for Best Practices in Scientific Software: The Story of BSSw.io

The authors of this October–December 2024 *Computing in Science & Engineering* article introduce the Better Scientific Software site (https://bssw.io), a platform that hosts a community of researchers, developers, and practitioners who share their experiences and insights on scientific software development. Since 2017, this collaborative hub has gained traction within the scientific computing community, attracting a growing number of readers and contributors eager to share ideas and elevate their software development practices.

IEEE Annals of the History of Computing

A Compelling Image: The Tower of Babel and the Proliferation of Programming Languages During the 1960s

In 1961 and in 1969, two towers of Babel of programming languages were published. The first one appeared on the cover of *Communications of the ACM*. The second one appeared on Jean E. Sammet's *Programming Languages: History and Fundamentals*. These two towers have come to symbolize the fragmentation that plagued the development of programming as a result of the multiplicity of notations. The author of this article, featured in the January–March 2025 issue of *IEEE Annals of the History of Computing*, argues that contrary to this common view, the tower on the cover of *Communications* should be understood as a proud display of the research that profoundly transformed computer programming during the late 1950s.

IEEE Computer Graphics and Applications

Understanding Collaborative Learning of Molecular Structures in AR With Eye Tracking

In this November/December 2024 *IEEE Computer Graphics and Applications* article, the authors present an approach for onsite instruction of multiple students accompanied by gaze-based monitoring to observe patterns of visual attention during task solving. They focus on collaborative processes in augmented reality (AR) that play an essential role in onsite and remote teaching alike. From a teaching perspective, it is important in such scenarios to communicate content and tasks effectively, observe whether students understand the task, and help appropriately.

IEEE Intelligent Systems

Large-Scale Package Deliveries With Unmanned Aerial Vehicles Using Collective Learning

Unmanned aerial vehicles (UAVs) have significant practical advantages for delivering packages, and many logistics companies have begun deploying UAVs for commercial package deliveries. To deliver packages quickly and cost-effectively, the routes taken by UAVs from depots to customers must be optimized. The authors of this January/February 2025 *IEEE Intelligent Systems* article present an innovative,

Digital Object Identifier 10.1109/MC.2025.3558105
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

practical package delivery model wherein multiple UAVs deliver multiple packages to customers, who are compensated for late deliveries.

Internet Computing

A Generative Modeling Method for Digital Twin Shop Floor

Digital twin as a key enabling technology for achieving digitization, flexibility, and customization in shop floors has attracted significant attention. However, the shop floor involves diverse assets across multiple dimensions, scales, and interdisciplinary fields, making the modeling process complex. To address this issue, this article from the January/February 2025 issue of *IEEE Internet Computing* analyzes the construction process of ontology-based information models and proposes a generative modeling method for digital twin shop floors driven by large language models.

micro

UCIe: Standard for an Open Chiplet Ecosystem

Universal chiplet interconnect express (UCIe) is an open industry standard die-to-die physical layer, link layer, and protocol layer for chiplets. It has industry-leading key performance indicators and has successfully coalesced the industry around a common die-to-die specification. This article featured in the January/February 2025 issue of *IEEE Micro* provides an overview of existing UCIe technology and highlights the work being done to create the next layer of standards required for an open chiplet ecosystem.

MultiMedia

Cryptanalyzing an Image Encryption Algorithm Underpinned by a 3-D Boolean Convolution Neural Network

This October–December 2024 *IEEE MultiMedia* article analyzes the security performance of an image encryption algorithm based on a 3D Boolean convolutional neural network (CNN). The algorithm utilizes the convolutional layers of a CNN as the encryption component, thereby achieving low-precision computations. However, due to its low-precision computation, this encryption algorithm employs one-to-one XOR and modulo operations, altering individual pixel values exclusively during encryption without diffusing changes to neighboring pixels. Capitalizing on this vulnerability, the authors propose chosen plaintext attacks on the one-round and multiple-round versions of this encryption algorithm.

pervasive COMPUTING

Multilabel Classification Model for Infant Activity Recognition Using Single Inertial Sensor

Recording and sharing childcare information is crucial for accurately assessing a child's health status and taking appropriate action in case of illness or other emergencies. In this article, featured in the October–December 2024 issue of *IEEE Pervasive Computing*, the authors implement a machine learning model to recognize multilabeled infant activities using a chest-mounted low-sampling-rate accelerometer. The performance evaluation considering multilabel classification shows that their proposed model reaches over 88% in the F1 score in the best case.

SECURITY & PRIVACY

Security Policy as Code

Engineering software systems to fulfill security requirements remains challenging. In this article, featured in the March/April 2025 issue of *IEEE Security & Privacy*, the authors advocate for designing and implementing software systems around integrated advanced security policies, capturing security requirements. They report on experience gathered with this approach in confidentiality-preserving data analytics.

Software

How Pandemics Changed a Public Software Ecosystem: Omaolo Case

When faced with a common threat, the ecosystem of companies needs to unite to face the challenge. The authors of this article from the March/April 2025 issue of *IEEE Software* describe how Omaolo, a public digital platform for welfare and healthcare, and its supporting software ecosystem evolved during the 2020–2022 COVID-19 pandemic in Finland.

IT Professional

Security-by-Design Issues in Autonomous Vehicles

As autonomous vehicle (AV) technology advances toward maturity, it becomes imperative to examine the security vulnerabilities within these cyberphysical systems. In this article, featured in the January/February 2025 issue of *IT Professional*, the authors spotlight imminent challenges faced by AV operators and explore emerging technologies for comprehensive solutions. They outline the diverse security layers, spanning physical, cyber, coding, and communication aspects, in the context of AVs.

Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *Computer* does not necessarily constitute endorsement by the IEEE or the IEEE Computer Society. All submissions are subject to editing for style, clarity, and space.

Reuse Rights and Reprint Permissions: Educational or personal use of this material is permitted without fee, provided such use: 1) is not made for profit, 2) includes this notice and a full citation to the original work on the first page of the copy, and 3) does not imply IEEE endorsement of any third-party products or services. Authors and their companies are permitted to post the accepted version of IEEE-copyrighted material on their own webservers without permission, provided that the IEEE copyright notice and a full citation to the original work appear on the first screen of the posted copy. An accepted manuscript is a version which has been revised by the author to incorporate review suggestions, but not the published version with copyediting, proofreading, and formatting added by IEEE. For more information, please go to: http://www.ieee.org/publications_standards/publications/rights/paperversionpolicy.html. Permission to reprint/republish this material for commercial, advertising, or promotional purposes or for creating new collective works for resale or redistribution must be obtained from IEEE by writing to the IEEE Intellectual Property Rights Office, 445 Hoes Lane, Piscataway, NJ 08854-4141 or pubs-permissions@ieee.org. Copyright © 2025 IEEE. All rights reserved.

Abstracting and Library Use: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Unlock Your Potential

WORLD-CLASS CONFERENCES — Over 195 globally recognized conferences.

DIGITAL LIBRARY — Over 900k articles covering world-class peer-reviewed content.

CALLS FOR PAPERS — Write and present your ground-breaking accomplishments.

EDUCATION — Strengthen your resume with the IEEE Computer Society Course Catalog.

ADVANCE YOUR CAREER — Search new positions in the IEEE Computer Society Jobs Board.

NETWORK — Make connections in local Region, Section, and Chapter activities.

Explore membership today at the IEEE Computer Society
www.computer.org

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3562049

50 & 25 YEARS AGO

EDITOR **ERICH NEUHOLD**
University of Vienna
erich.neuhold@univie.ac.at

JUNE 1975

https://www.computer.org/csdl/magazine/co/1975/06

UPDATE (p. 9ff): "1975 National Computer Conference-Report on Selected Sessions; The first National Computer Conference on the west coast, NCC '75 was held in Anaheim from 19 May to 22. The conference's technical program was divided into three categories: Science and Technology, Methods and Applications, and Interaction with Society." *[Editor's note: Just a little history (see also http://bitsavers.trailing-edge.com/pdf/afips/AFIPS_Conference_Dates.txt). The Joint Computer Conferences started in 1951 and ended in 1987. At the beginning they happened twice a year (East and West). Then, organization was taken over by American Federation of Information Processing Societies (AFIPS) in 1962. They were combined and called The National Computer Conference 1973 and ended in 1987 with the demise of AFIPS. Especially in the early years, they were quite formative for the American computer research and application communities.]*

Guest Editor: Structured Programming: Highlights of the 1974 Lake Arrowhead Workshop; William F. Ross (p. 21): "But important efforts to address software technology have been underway during the same period—notably in the 60's by Bohm, Jacopini, Dijkstra, Parnas, and others in the university environment, and in the 70's by such people as Mills and Baker, who defined specific elements of a methodology directed toward reducing software costs and improving software quality. ... Subsequently—and inevitably—all this attention prompted questions as to whether SP *[Editor's note: structured programming]* is in fact feasible and economically practical on an industrywide basis, and if so to what extent." (p. 22) "A recurring question at the workshop was, Which of the SP techniques provided the greater benefits? Although many speakers presented extensive experience and attested to significant benefits in employing SP, none was able to quantify each specific technique's incremental contribution toward the totality of benefit." *[Editor's note: A very interesting workshop that raised, 50 years ago, many programming questions, not only concerning "structured programming," the in-word of the time. Of course, many other software development ideas came about later, but even today many of the questions raised are still not answered. This introduction is followed by many short, interesting articles. For lack of space, I will only cite the titles and leave it to you, my readers, to delve into them.]*

Session I: Structured Programming: Concepts and Definitions: Overview; John Naughton (p.23) ... **Structured Programming: Review of Some Practical Concepts; Clairmont McGowan** (p.24) ... **Structured Programming: Fortran: Can it be Structured – Should it be?; Ellis Horowitz** (p. 30).

Session II: Structured Programming: A Quantitative Assessment: Overview; Barry W. Boehm (p. 38) ... **Structured Programming at AUTO; Charles E. Holmes** (p. 41) ... **Applying Structured Programming to Command, Control, and Communication Software Development; Gene R. Katkus** (p. 43) ... **An Implementation of Structured Code Techniques on a Real-Time System; James P. Romanos** (p. 48) ... **Measuring Programming Improvement at IBM-FSD; Robert Mc Henry** (p. 49) ... **Experience and Accomplishments with Structured Programming; E. Kent Gordon** (p. 50) ... **Concluding Remarks; Barry W. Boehm** (p. 53).

Session III: Structured Programming: Problems, Approaches, and Techniques: Overview; Robert R. Brown (p.55) ... **Structured Programming: Agony and/or Ecstasy; John R. Brown** (p. 56) ... **Experience with Module-Level Specification Methods; John W. Brackett** (p. 58).

Session IV: Impact of Structured Programming on Evolving Technologies and Related Programming Technologies: Overview; Guy de Balbine et al. (p. 62) ... **The Need for Improved Programming Language; Charles T. Zahn,**

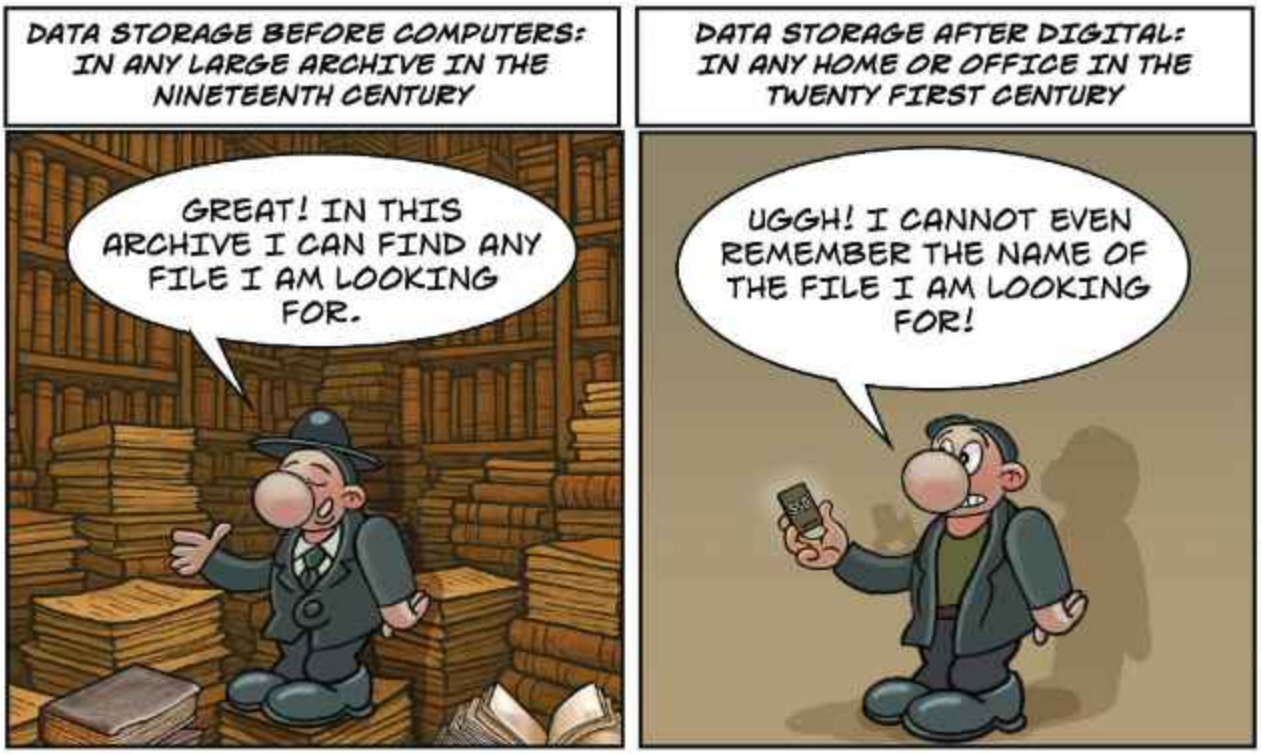
Digital Object Identifier 10.1109/MC.2025.3551701
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

Jr. (p.63) ... **The Structuring Engine: A Transitional Tool; Guy de Balbine et al.** (p. 64) ... **A Formal Design Medium for Software; Larry Robinson** (p. 66) ... **Structured Code via Stack Machine; Richard Bigelow** (p. 67) ... **Structured Programming Applied to Equipment Testing; Judy Townsley** (p. 68) ... **Software Development for Distributed Systems; David J, Faber** (p. 68) ... **The Need for Language Advances; Alfred S. Liddle** (p. 69) ... **Preparing for Future Needs; Noah S. Prywes** (p. 70) ... **Conclusions; Guy de Balbine** (p. 72).

COMPUTING THROUGH TIME
DATA STORAGE

BY ERGUN AKLEMAN
ERGUN.AKLEMAN@GMAIL.COM



BEFORE THE DIGITAL ERA, DATA STORAGE RELIED ON PHYSICAL MATERIALS AND MANUAL ORGANIZATION. STONE TABLETS AND CLAY TABLETS WERE AMONG THE EARLIEST FORMS, INSCRIBED WITH IMPORTANT RECORDS IN ANCIENT CIVILIZATIONS. THESE GAVE WAY TO SCROLLS, BOUND LEDGERS, AND MANUSCRIPTS, CAREFULLY PRESERVED IN LIBRARIES AND ARCHIVES. OVER TIME, INSTITUTIONS USED FILING CABINETS AND PAPER ARCHIVES TO MANAGE GROWING VOLUMES OF INFORMATION—EACH FOLDER AND SHEET METICULOUSLY LABELED AND CROSS-REFERENCED. THOUGH ROBUST AND TANGIBLE, THESE SYSTEMS WERE SPACE-INTENSIVE, PRONE TO DEGRADATION, AND REQUIRED SIGNIFICANT EFFORT TO MAINTAIN AND RETRIEVE DATA.

Digital Object Identifier 10.1109/MC.2025.3558106
Date of current version: 29 May 2025

Special Tutorial; Introduction to the Role of Redundancy in Computer Arithmetic; D. E. Atkins (p. 74): "Rather, the focus will be on the other two potential benefits: more specifically, on the judicious use of number systems employing redundancy in representation. ... Redundancy in the Partial Product: The use of carry-save adders to accelerate the iterative portion of digital multiplication is well-known. Its basis is the realization that carries need not be propagated during each addition of a long series of additions, provided that carries are explicitly stored." (p. 75) "Redundancy in the Multiplier: Another acceleration technique, typically used in conjunction with introducing redundancy into the partial product, is that of recoding the multiplier from a nonredundant to a redundant digit set. ... Redundancy in the Quotient: To accelerate the performance of division, redundancy may be introduced into the representation of the quotient. ... Redundancy to Provide Structural Flexibility: The emphasis has been on the use of a so-called signed-digit number system." *[Editor's note: An interesting, but rather specialized, tutorial about increasing the efficiency of numeric calculations. Using references like Wikipedia, you will find a difference between signed-number representations (mostly talked about here) and signed-digit representations, a much wider field.]*

JUNE 2000

https://www.computer.org/csdl/magazine/co/2000/06

Are Too Many Programmers Too Narrowly Trained?; David Clark (p. 12): "The use of programmers with narrow training can lead to buggy or broken applications, as well as expensive delays of product releases. ... A broad, engineering-based education is important because programming, as well as other elements of computer technology, has become very complex. ... Some critics say that software development companies, in their rush to profitability, are interested only in hiring people with the specific programming skills." (p. 14) "To improve programmer quality, educators and consultants often suggest that the software industry adopt a model of mandatory licensing and certification, as is used in the medical and legal professions." *[Editor's note: An interesting article that remains valid even today. Higher educational institutions offer broad computer education, not just skills, and many certification programs are around. Unfortunately, there was never an agreement as to what actually would be the best education and certification to prepare students for their software development future.]*

Distributed Net Applications Create Virtual Supercomputers?; George Lawton (p. 16): "The most computationally intensive long-term distributed Internet application on the planet, SETI@home, which analyzes signals in space looking for signs of intelligent life, runs on more than 2 million computers and processes an average aggregate of 12 teraflops, according to the project's chief scientist." (p. 17) "In coarse-grained applications, clients communicate with master and proxy servers but not with each other. ... In fine-grained applications, participating clients must communicate with each other." (p. 20) "CONCERNS: ...SECURITY" (p. 21) "Bandwidth consumption ... Internet congestion ... for the first time we can truly aggregate this unused resource for valuable applications." *[Editor's note: An interesting article that cites, besides SETI ("search for extraterrestrial intelligence" that ended 2020) a number of other applications. Of course, distributed Internet computing stayed with us using quite a number of different names, e.g., network, grid, cloud, fog, edge, etc. but in all cases based on the same basic principles.]*

News Briefs; Ed: Anne C. Lear (p. 22ff): "Love Hurts: New E-Mail Worm Afflicts Millions: A treacherous successor to last year's Melissa virus charged across the Internet in early May. ... New Chip Helps with Network Security: A new network processor could ward off denial-of-service and other network attacks by greatly accelerating the rate at which networks filter data packets. ... In addition to filtering, Juniper's processor can sample, count, or log packets." *[Editor's note: Despite promises made then, malware news, as well as techniques to solve those problems, still abound. We know, however, that those problems have not disappeared but became even worse. True solutions, I believe, are not around as governments, their agencies, and corporations are quite involved in producing all types of malware.]*

Components: What If They Gave a Revolution and Nobody Came?; Peter M. Maurer (p. 28): "Unlike earlier so-called revolutions—such as structured or object-oriented programming—component-level programming is a true revolution on a par with stored-program computers and high-level languages. ... Despite the lack of press coverage, there is no question that a revolution has taken place. Visual Basic, the first language that supported component-level programming, is now the preferred language for new development." (p. 31) "Today OLE *[Editor's note: object linking and embedding]* controls are called ActiveX controls ... based on a technology known as the Common Object Model (COM), which in turn is based on the concept of published, immutable interfaces." *[Editor's note: This very interesting article analyzes in detail concepts of component-based programming and predicts many aspects that made component-based programming an essential technique of today's system development. Complex systems as well as APPs are using the technique for the save reuse of components.]*

Winning Teams: Performance Engineering during Development; Robert S. Oshana. (p. 36): "Loosely defined, software performance engineering *[Editor's note: SPE]* is a set of techniques designed to gather data, construct a system performance model, evaluate that model, manage the risk of uncertainty, evaluate alternatives, and verify the models and results. ... Integrating SPE techniques across the various functional organizations proved instrumental in mitigating these risks." (p. 37) "As mentioned previously, SPE is a set of techniques for constructing and evaluating system performance models." *[Editor's note: A very interesting article as it uses*

a digital signal processor project to exemplify the various stages where performance engineering has played an important role.]

Guest Editor: Real-Time Distributed Object Computing: An Emerging Field; Eltefaat Shokri et al. (p. 45): "While the field of object-oriented real time computing (ORC) is young, it is growing quickly because it offers such a wide range of applicability, from complex real time systems to the next generation of computing and communication devices. ... With this collection of articles, we've tried to emphasize innovative and practical solutions for integrating OO *[Editor's note: object oriented]* computing technologies into RT *[Editor's note: real time]* systems engineering methods. ... We are pleased to be able to present four valuable articles that cover a wide range of issues that center on developing object-oriented, real time distributed computing systems." *[Editor's note: The four articles that I present below, only with their titles, authors, and abstracts, are very interesting to read as they explain pros and cons quite in detail on what approaches were taken to arrive at the proposed extensions, respectively, adjustments of the discussed methodologies.]*

The Real-Time Specification for Java; Greg Bollella et al. (p. 47): "The RTSJ *[Editor's note: real-time specification for Java]* provides a platform that will let programmers correctly reason about the temporal behavior of executing software. Two members of the Real-Time for Java Experts Group explain the RTSJ's features and the thinking behind the specification's design."

An Overview of the Real-Time CORBA Specification; Douglas C. Schmidt et al. (p. 56): "OMG's *[Editor's note: object management group's]* Real-Time CORBA *[Editor's note: common object request broker architecture]* provides standard policies and mechanisms that support quality-of-service requirements end to end. Such support enhances the effectiveness of distributed object computing middleware as a platform for performance-sensitive real time systems."

A Generic Framework for Modeling Resources with UML; Bran Selic (p. 64): "For real time systems, designers must consider physical and logical resources. Developers can use the OMG's Unified Modeling Language to model resources and thus predict crucial system properties before fully implementing a system."

APIs for Real-Time Distributed Object Programming; K.H. (Kane) Kim (p. 72): "This article focuses on application programming interfaces (APIs) that take the form of C++ and Java class libraries and support high-level, high-precision, real time object programming without requiring new language translators."

Communications: Recent Advances in Wireless Networking; Upkar Varshney (p. 100): "After discussing advances in wired networking in a previous column (Recent Advances in Wired Networking, Computer, April 2000, pp. 107–109), I now turn to advances in wireless networking. ... These factors—along with demand for higher bandwidth and global roaming—will continue to push the standardization and the near-future deployment of third-generation wireless networks using terrestrial and satellite components." (p. 101) "These specifications offer the flexibility needed by both the satellite/terrestrial providers to design new third-generation systems." *[Editor's note: The article mostly discusses third-generation systems. As we know, fourth- and fifth-generation systems are now here and the sixth generation is already intensely discussed.]*

The Empire Strikes Back ... with the X-Box; Michael Macedonia (p. 104): "Microsoft's fortunes have certainly gone sour recently. ... Overshadowing these developments, however, is Microsoft's announcement that it plans to build the X-Box." (p. 105) "The console market will prove a tougher challenge, though, because here Microsoft faces dominant, entrenched competition: Sony." *[Editor's note: Despite the title, the article does not predict the success of the X-Box but rather analyzes the expected properties of it. The X-Box was and still is successful today but is positioned in third place after PlayStation and Nintendo.]*

Alive and Well: Jini Technology Today; Jim Waldo (p. 107): "The Jini community is an ongoing experiment in trying to mix open-source development techniques with industrial engineering development. ... In a Jini network, services advertise themselves by saying what Java language interfaces they implement. ... Advertisement and matching of client and service occurs in a Jini lookup service, a place where providers of a service can register what they provide, and clients of services can look for what they need." *[Editor's note: Yes, Jini was alive in 2000 but was later moved to Apache as Apache River, and was finally retired in 2022 due to lack of further activity.]*

The End of Research as We Know It?; Ted Lewis (p. 112): "Venture capitalists stalk the halls of Stanford University's Gates Building, seeking out high-profile research projects and the bright graduate students drawn to them. ... Thus has venture capital rapidly displaced government grants as the preferred source of research funding. ... Once again, rich private citizens will fund most research." (p. 110) "A new urgency is needed in the lab as well, to compress the product-development model as follows: 1. Academic research still provides new-product ideas, but the focus shifts to student dissertations. 2. Venture capital funds a startup company built around a new-product idea." ... (p. 111) "VCs routinely purchase innovation for a few shares of stock, usually at the expense of a student's educational future, financial health, or both." *[Editor's note: This rather pessimistic prediction of our research future has proven wrong, as the government funded research, unfortunately much via the military, has grown significantly over the years. Startup-funded research has also shown its fallacy, as the dot-com crises and oversold claims have shown.]*

EDITOR IN CHIEF **JEFFREY VOAS**
IEEE Fellow; j.voas@ieee.org

Artificial Intelligence, Concrete, and Wood

Michael Chapiro, SolutionsDiscovery
Jeffrey Voas, IEEE Fellow

This message explores the relationship between CO_2 emissions, data centers, and AI.

In the article "The AI Boom Rests on Billions of Tonnes of Concrete" (https://spectrum.ieee.org/green-concrete), we were struck by this comment:

> "Concrete is not just a major ingredient in data centers and the power plants being built to energize them. As the world's most widely manufactured material, concrete—and especially the cement within it—is also a major contributor to climate change, accounting for around 6% of global greenhouse gas emissions. Data centers use so much concrete that the construction boom is wrecking tech giants' commitments to eliminate their carbon emissions. Even though Google, Meta, and Microsoft have touted goals to be carbon neutral or negative by 2030, and Amazon by 2040, the industry is now moving in the wrong direction."

That article inspired us to look at how data centers are impacting CO_2 emissions beyond the impact from concrete. Here are a few facts.

> "According to IEA, estimated global data center electricity consumption, in 2022, was 460 terawatt-hours (TWh) or around 1–1.3% of global final electricity demand, excluding energy used for cryptocurrency mining (estimated to be 110 TWh in 2022) or for data transmission network energy use (estimated to range from 260 to 360 TWh in 2022). Using the value of 350 TWh (40 GW) for data centers, then in 2024, data centers consumed more energy than all but a few countries."[1]

"In 2022, U.S. total primary energy consumption was about 95 quadrillion British thermal units (Btu), which was equal to about 16% of total world primary energy consumption of about 600 quadrillion Btu."[2] Assuming

DISCLAIMER
The authors are completely responsible for the content in this message. The opinions expressed here are their own.

Digital Object Identifier 10.1109/MC.2025.3548825
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

IN THIS ISSUE

In this issue, we're publishing four articles on diverse topics.

In the first article,[A1] the authors argue that digital data are necessary for science research. The authors explain that they already published a Research Data Reference Architecture (RDRA) that is a high-level feature set useful for data storage implementations. However, they acknowledge that their RDRA did not mandate technology choices and did not specify service levels. This article proposes a new Capability Maturity Model that allows organizations to assess their specific requirements, including projected growth, risk profiles, and budgets.

In the second article,[A2] the authors explore the impact of artificial intelligence (AI) on humans. The article discusses Responsible AI as it relates to designing, developing, and deploying more ethical AI systems. The article proposes knowledge graphs for Responsible AI and suggests how these graphs process unstructured information.

In the third article,[A3] the author explores the shift from manual to automated testing, emphasizing the role of AI and machine learning in enhancing efficiency and quality assurance in the software development lifecycle. The article evaluates current practices and discusses advancements, including shift-left testing, continuous testing, and AI-driven metrics. An experimental evaluation is provided that compares traditional testing with AI-enhanced methods.

In the fourth article,[A4] the authors envision a future where information systems actively assist in completing tasks and reducing workloads. This article discusses research on AI agents and artificial capable intelligence that aims to reach the next frontier in information access: *task completion*. This includes task automation and action engines that work with humans and require reliability, safety, and security.

—Jeffrey Voas, Editor in Chief

Digital Object Identifier 10.1109/MC.2025.3557071
Date of current version: 29 May 2025

APPENDIX: RELATED ARTICLES

A1. D. Abramson and J. Carroll, "A capability maturity model for research data storage systems," *Computer*, vol. 58, no. 6, pp. 30–39, Jun. 2025, doi: 10.1109/MC.2025.3540093.

A2. X. Li, Q. Liu, Q. Bai, and X. Xu, "The role of knowledge graph on responsible artificial intelligence realization: Research opportunities and challenges," *Computer*, vol. 58, no. 6, pp. 40–48, Jun. 2025, doi: 10.1109/MC.2025.3545036.

A3. A. Saxena, "Rethinking software testing for modern development," *Computer*, vol. 58, no. 6, pp. 49–58, Jun. 2025, doi: 10.1109/MC.2025.3554094.

A4. S. Nath et al., "From search engines to action engines," *Computer*, vol. 58, no. 6, pp. 59–68, Jun. 2025, doi: 10.1109/MC.2025.3556643.

all U.S. data centers consume 50% of the globally consumed electricity for the data centers, data centers in the United States consume about 3.9 quadrillion Btu or about 4.1% of the total electric energy consumption in the United States.

If we assume that the data transmission network energy use and the energy used for cryptocurrency mining in the United States constitute 50% of the global use for those activities, then we can assume that the United States spent another 4 quadrillion Btu, or about 4.1% of the total electric energy consumption in the United States, for data transmission and cryptocurrency mining.

"By 2026, the energy spent in the United States on data centers is estimated to increase by 1.5 times."[1] This effectively makes energy consumption by data centers in the United States about 6 quadrillion Btu. If electricity consumption for cryptocurrency mining and data transmission network energy usage also increases by 1.5 times, we can add another 6 quadrillion Btu.

If electricity consumption in the United States stays at around 95 quadrillion Btu in 2026, then energy consumption by the data centers, data transmission, and cryptocurrency mining will grow to about 12.1% of the total electric energy consumption in the United States.

According to the Energy Information Administration, "the annual total energy-related carbon dioxide emissions" in 2023 was approximately 4,807 million metric tons (MMT). "The term energy-related CO_2 emissions, as used in these tables, refers to emissions released at the location where fossil fuels are consumed."[3] The electrical power sector was responsible for 1,427 MMT in 2023.[3] Assuming in 2026 that the electrical power sector will be responsible for about 1,500 MMT, we can conclude that the estimated carbon dioxide emissions related to electricity consumption for data centers, data transmission, and cryptocurrency mining (technology) will approach

WRITING FOR *COMPUTER*

Magazines are great places to park stories. *Computer* continually seeks and encourages authors who have good stories. Storytelling can be technical or nontechnical, for example, opinion pieces. If you've ever thought about writing for *Computer*, let me explain your options.

Most readers are aware of *Computer*'s feature articles. There are three types: research feature, perspective, and computing practice. Feature articles undergo the normal IEEE peer review and should be at least 4,000 words in length, including figures and tables, which account for 300 words each. (See the guidelines for submitting a feature article at https://www.computer.org/csdl/magazine/co/write-for-us/15913?title=Author%20Information&periodical=Computer.) You can simply submit a paper on a topic of your choosing, or you can submit a paper to an existing call for papers for a special issue. (See the guidelines for submitting a feature article at https://www.computer.org/publications/author-resources/calls-for-papers?type=mags&publication=co.)

The other side of *Computer* is the column articles. Our current columns are "Cybertrust," "Microelectronics," "Cyber-Physical Systems," "Internet of Things," "IT Innovation," "Open Source," "Data," "Standards," "Computing's Economics," "Software Engineering," "Artificial Intelligence/Machine Learning," "Education," "Algorithms," "Memory and Storage," "Computing Architectures," "Humanity and Computing," "Games," "Notes From the Field," and "Industry Insights." Most of these columns are open for unsolicited articles.

Column articles are usually either opinion pieces or short technical articles. Column articles should be fewer than 2,500 words in length, including figures and tables, which account for 300 words each. Column articles are approved for publication by the column's editor(s). Column articles do not go through the peer-review process that feature articles do. If you wish to write a column article, you should pitch your idea for your article to the column's editor(s) first.

And we have one more article category, which we refer to as *virtual roundtable*. These are created from virtual panel discussions where a handful of experts are given a set of questions concerning a timely technical topic. The question responses are written into the article. These articles are quite popular with the organizers, panelists, and readers. If you care to organize one of these, please contact the editor in chief (EIC) first.

If you're more interested in guest editing a special issue of *Computer*, please first read the guidelines for special issue proposals at https://www.computer.org/csdl/magazine/co/write-for-us/15911?title=Special%20Issue%20Proposals&periodical=Computer. If you care to organize one of these special issues, please contact the EIC first. Note that one guest editor for a special issue must be a current member of the Editorial Board of *Computer*.

Writing for *Computer* is rewarding! Think about it.

–*Jeffrey Voas*, *Editor in Chief*

Digital Object Identifier 10.1109/MC.2025.3559636
Date of current version: 29 May 2025

about 180 MMT in 2026. This amount constitutes about 50% of the CO_2 emissions from the entire U.S. residential sector and more than 50% of the entire U.S. commercial sector in 2023.

And finally, Amazon, Microsoft, Google, and Oracle are the leaders in cloud computing. According to *The Washington Post*, Amazon has doubled down on nuclear energy with deals for small reactors.[4]

> "Amazon is leading a US$500 million funding round for X-Energy Reactor, a company that develops small modular nuclear reactors and fuel. It's also working with utilities in Washington state and Virginia on potential SMR projects. Google said Monday it will purchase energy from small modular nuclear reactors built by Kairos Power. The first Kairos Power SMR is intended to come online by 2030. Amazon and X-Energy want to bring more than 5 gigawatts of power projects online by 2039."[5]

And although nuclear power reactors do not directly emit CO_2, mining and refining uranium ore for reactor fuel require energy. "Nuclear power plants require metal and concrete which requires energy. If fossil fuels are used for mining and refining uranium ore, or if fossil fuels are used when constructing nuclear power plants, the emissions from burning those fuels should be balanced against the electricity that nuclear powers generate."[6]

And finally, according to computer scientists at the University of Waterloo in Canada, changing 30 lines of code in Linux could cut energy use at some data centers by up to 30%.[7] And, interestingly, Microsoft is looking to build

data centers using wood to decrease emissions.[8]

The bottom line is that data centers increase CO_2 emissions. One solution to reduce these emissions may be for data centers to run more efficient computing algorithms (for example, proof of work in crypto is an energy "guzzler"). Or maybe, the answer is as old as time—wood.

REFERENCES

1. S. Shankar, "Challenging trends in energy of computing for data centers," *Computer*, vol. 57, no. 12, pp. 134–142, Dec. 2024, doi: 10.1109/MC.2024.3468568.
2. "Frequently asked questions (FAQs) — What is the United States' share of world energy consumption?" U.S. Energy Information Administration (EIA) (.gov). Accessed: Mar. 11, 2025. [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=87&t=1
3. "U.S. energy-related carbon dioxide emissions, 2023," Energy Information Administration, Washington, DC, USA, Apr. 2024. [Online]. Available: https://www.eia.gov/environment/emissions/carbon/
4. C. Mitchell, "Gartner magic quadrant for strategic cloud platform services 2023," *CX Today*, Jan. 15, 2024. [Online]. Available: https://www.cxtoday.com/customer-data-platform/gartner-magic-quadrant-for-strategic-cloud-platform-services-2023/
5. S. Najmabadi and E. Halper, "Amazon doubles down on nuclear energy with deal for small reactors," *The Spokesman-Rev.*, Oct. 16, 2024. [Online]. Available: https://www.spokesman.com/stories/2024/oct/16/amazon-doubles-down-on-nuclear-energy-with-deal-fo/
6. "Nuclear explained: Nuclear power and the environment." U.S. Energy Information Administration (EIA) (.gov). Accessed: Mar. 11, 2025. [Online]. Available: https://www.eia.gov/energyexplained/nuclear/nuclear-power-and-the-environment.php#:~:text=Unlike%20fossil%20fuel%2Dfired%20power,or%20carbon%20dioxide%20while%20operating
7. M. Gooding, "Changing Linux code could cut data center energy use by 30%, researchers claim," *Data Center Dyn.*, Jan. 22, 2025. [Online]. Available: https://www.datacenterdynamics.com/en/news/changing-linux-code-could-cut-data-center-energy-use-by-30-researchers-claim/
8. O. Eldert, "Microsoft is making data centers out of wood — And it's a trend that could redefine the industry," *Yahoo News*, Jan. 2, 2025. [Online]. Available: https://www.yahoo.com/news/microsoft-making-data-centers-wood-110013459.html?guccounter=1

MICHAEL CHAPIRO is the lead principal solutions architect at SolutionsDiscovery, Gaithersburg, MD 20878 USA. Contact him at mchapiro@solutionsdiscovery.com.

JEFFREY VOAS, Gaithersburg, MD 20899 USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.

Assured Autonomy, Artificial Intelligence, and Machine Learning

Phil Laplante, IEEE Fellow

The most recent IEEE Workshop on Assured Autonomy, Artificial Intelligence and Machine Learning provided a forum for experts to reflect on the state of AI and security. This virtual roundtable summarizes their observations and recommendations.

Artificial intelligence (AI) and machine learning (ML) systems are increasingly seen in areas such as self-driving land vehicles, autonomous aircraft, and medical systems. AI systems should equal or surpass human performance, but given the consequences of failure in these systems, how do we determine that the data gathered to train an AI system are suitably representative of the real world? How do we assure the public that these systems work as intended and will not cause harm?

SECURITY AND AI/ML SYSTEMS

This virtual roundtable is based on panels focused on security and AI/ML systems at the Third IEEE International Workshop on Assured Autonomy, Artificial Intelligence and Machine Learning (WAAM 2024) on 30 October 2024. We explored various issues of AI/ML systems and security, including but not limited to the secure design of AI/ML systems, adversarial AI, using AI/ML to secure other systems, and using AI/ML for post-incident analysis. Participants listed next provided a wide range of views on research, experiences, and best practices.

The hope of this virtual roundtable is to provide a better understanding of the state of trust and assurance for autonomous systems and ML. Please note that the views presented here are those of the roundtable participants

Digital Object Identifier 10.1109/MC.2025.3548739
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

ROUNDTABLE PANELISTS

Tracy (Trac) Bannon is a software architect and researcher, MITRE Corporation, Bedford, MA 01730 USA. Contact her at tbannon@mitre.org.

Payel Das is a principal research staff member and a manager, IBM Research Artificial Intelligence (AI), IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10597 USA. Contact her at daspa@us.ibm.com.

Joanna F. DeFranco is an associate professor of software engineering and Associate Director, Doctor of Engineering in Engineering Program, The Pennsylvania State University, University Park, PA 16802 USA. Contact her at jfd104@psu.edu.

Alwyn Goodloe is a computer scientist, Formal Methods Group, Safety Critical Avionics Systems Branch, NASA Langley Research Center, Hampton, VA 23681 USA. Contact him at a.goodloe@nasa.gov.

Rick Kuhn is a computer scientist, Computer Security Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. Contact him at kuhn@nist.gov.

Erin Lanus is a research assistant professor at the National Security Institute and Affiliate Faculty Computer Science, Virginia Tech, Arlington, VA 22203 USA. Contact her at lanus@vt.edu.

Sandeep Neema is a professor of computer science, Vanderbilt University, Nashville, TN 37235 USA. Contact him at Sandeep.Neema@vanderbilt.edu.

Sara Rampazzi is an assistant professor in the Department of Computer and Information Science and Engineering (CISE), University of Florida, Gainesville, FL 32611 USA. Contact her at srampazzi@ufl.edu.

M S Raunak is a computer scientist, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. Contact him at ms.raunak@nist.gov.

Anoop Singhal is a senior computer scientist in the Computer Security Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. Contact him at anoop.singhal@nist.gov.

David Stracuzzi is manager, Machine Intelligence Department, the Sandia National Laboratories, Albuquerque, NM 87123 USA. Contact him at djstrac@sandia.gov.

Girija Subramaniam is the founder of Forcing Function LLC, Montgomery Village, MD 20886 USA. Contact her at girija@forcingfunction.org.

and authors only. We think this virtual roundtable will give a nice baseline on this topic for 2025 in this very rapidly changing field.

***COMPUTER*:** What are the current and potential future security threats to AI and ML systems?

ERIN LANUS:

Key issues. As with many other systems, ML models have primarily been evaluated in terms of their performance, and as with many other systems, security and privacy may have a slight cost in terms of performance. We need to ensure that evaluations of AI-enabled systems include security and quantify the cost of losses due to security violations against the cumulative cost of performance dips due to the employment of defense mechanisms.

Another issue is that organizations often structure AI security within AI and not within cybersecurity. This separation can lead to the loss of security expertise gained through decades of experience. Many AI security issues are mitigated or the attack surface is narrowed through traditional cybersecurity practices. For example, the correct use of access control, encryption, homomorphic encryption, and integrity checking can all reduce the size of the attack surface for data poisoning by protecting data at rest and in computation, such as when using third-party services for storage or model training. Of course, securing against data poisoning may also require additional new approaches due to the statistical nature of attacks, such as the number of samples needed to be manipulated to impact the model or even physics-informed detection when the physical relationships present in the uncorrupted data are known and can be modeled. Still, cybersecurity is already a varied field (ranging, for example, through databases, networks, software, sensors, and social engineering), and securing each type of system requires domain subject matter expert knowledge.

AI security could benefit from identifying other commonalities among the varied systems for which we already apply cybersecurity. For example, security differs from safety in that safety violations often occur in rare events that happen with some measurable probability while security violations often occur in unexpected events that an intelligent adversary intentionally causes by hunting for a weakness; adversarial thinking is an important and teachable skill. Systems that do or do not include AI may all be susceptible to spoofing attacks, making an input look like a different input; examples from traditional cybersecurity include phone phreaking and spam detection evasion, while patch and perturbation attacks are used to evade ML perception. Differential privacy, a mechanism for ensuring individual privacy by guaranteeing that the result of a query is not substantially dependent on any individual's data, has been employed for protecting privacy in ML systems.

State of security. We have known for about a decade that ML models are vulnerable to novel attacks. While important and alarming, the practicality of the attacks is not always clear. Perturbation-based attacks that add noise to pixels are easily exploitable if there is a vulnerability in the other system components that allows an attacker to inject noise, either by supplying their own noised sample, possibly bypassing the system input/sensor, or by having noise added to the sample somewhere on the communication channel between the input and model. Of course, there are many other things an attacker could do with this level of access that would not require much of the knowledge needed to construct an adversarial sample, such as directly changing the output of the model to the desired class. That said, these attacks may be stealthy. Work that demonstrated the applicability of physical attacks—placing stickers on street signs, 3D printing turtle shells with rifle patterns, or using light patterns from lasers to spoof against a lidar sensor—makes it possible to attack the model with no system access so long as the adversary can manipulate the environment. We are also seeing ML not as the target of attack but used to deliver a malicious payload. For example, malware can be stored in the least significant bits of model weights and bypass security scanning without impacting model performance.

Some defense mechanisms for perturbation attacks, like retraining on a dataset with adversarial samples, put the defender in the "whack-a-mole" position of responding to attacks after they are found. Others, like smoothing mechanisms that involve aggregating the predictions for samples taken around the potentially compromised point, can be more proactive. Ideally, we would like defenses with a guarantee of security like what is done in cryptography, either in terms of computational difficulty or provable unconditional security. We also want to think holistically about the entire system. A lot of attention is currently being paid to tweaking attacks against an ML model because it's exciting, but there may be easier attack vectors elsewhere in the system. In cybersecurity, we know that the attacker is not going to try to break the cryptographic algorithm if they can find a weakness in the protocol or if they can just steal the cryptographic key. We need security employed across the entire system. Additionally, as with all systems, AI-enabled system components are not siloed, and weaknesses may occur where the components interact. ML developers should be given some basic training in cybersecurity and deeper knowledge of AI security to help secure the final composed system.

> AI security could benefit from identifying other commonalities among the varied systems for which we already apply cybersecurity.

Main threats. One of the biggest threats against AI/ML systems is data poisoning due to the data-centric nature of these systems. Data poisoning is the act of tampering with the training data to cause a model to learn an incorrect function. It may be targeted to a specific class, possibly with a backdoor trigger that creates an exploitable vulnerability known only to the attacker, or untargeted, where the goal is degraded model performance more generally. Data poisoning is challenging because it can occur in numerous places throughout the AI/ML lifecycle and is often difficult to detect. As models have difficulty generalizing beyond the training data, the best data to use for training are operationally relevant data. For systems that operate in contested environments, the adversary may poison the original data at the time of collection. In this case, there may be no clean data to compare against. Open source data without authentication and integrity checking are particularly vulnerable to data poisoning. Data poisoning may also occur at rest while the data are stored when there are weaknesses in the system's access control or during computation such as by training on a cloud platform or using open source packages with malicious code. Models may be updated with new samples to avoid data drift over time via retraining or online learning, but this creates another attack vector for data poisoning.

Like data poisoning, the assumptions we make about the data may not be correct. For example, anomaly detection is an unsupervised learning technique where the "normal" data are modeled, and inference classifies incoming samples that don't match the original distribution as "abnormal." Anomaly detection used to identify attacks in network traffic requires that attack traffic is not in the collected data. However, zero-trust principles for security tell us to assume that an adversary is already present in the network.

Mis/disinformation. Generative AI (GAI)[a] has increased the scale at which mis/disinformation campaigns can be waged and has lowered the bar to entry. Previously, special skills in specific domains were needed to create forgeries that would pass scrutiny, but generative technologies can automate the process of creating deepfakes in modalities like image, video, voice, and text. The good news is that GAI often leaves fingerprints that are discernable to a human on inspection. On the other hand, GAI bots can create so much fake content that they drown out the truth,

[a]The findings of another panel on GAI and DevSecOps will appear in the July edition of my "Software Engineering" column.

which has the similar effect of exhausting a human operator receiving too many false positives and wears down the ability to perform inspection.

Measures. The AI security literature quantifies the effectiveness of attacks in terms of how many attacks achieved their intended goal: the attack success rate (ASR). The ASR is often parameterized by aspects of the attack, such as the strength of perturbation applied, as high-strength perturbations can often completely disable ML perception but at the cost of being highly detectable to a human operator (not stealthy). Defenders may measure their models in terms of the number of attacks against which they are robust or the size of the robustness ball that serves as a distance around a sample. As in the cybersecurity community, security evaluations should first be made by defining the adversary model as well as the assets to protect by any defense mechanisms in place. Coverage criteria have been used extensively in software testing to quantify the coverage of the system by the test, and combinatorial coverage has been adapted for characterizing the ML input space. There is likely a role for coverage criteria over the adversary characteristics to put into context where and when models are vulnerable.

Requirements. Security is only proven against a specific adversary model. Yet, for ML systems, requirements are often implicitly specified by the data. Both the intended use requirements and the security requirements—what security properties (for example, confidentiality, integrity, and availability) the system should have, under what conditions, and against which adversary—should be specified like other information systems.

Quality. In cybersecurity, and especially cryptography, we are taught Kerckhoff's principle, summarized as "security should not rely on obscurity." While a lack of knowledge about the system may slow down the adversary, it rarely stops them. In the case of ML, even if the adversary does not have access to the model directly, they may know quite a bit. If they know the task being performed, they may be able to guess the model architecture (or similar) and even possibly the dataset. With query access, they can build a surrogate model, and adversarial samples often transfer from one model to another, even when they have different architectures.

ALWYN GOODLOE: Most organizations have siloed AI into separate AI-focused groups isolated from traditional software engineering. We often hear how "AI is different," meaning that traditional cyber and software engineering techniques do not apply to the brave new world of AI/ML. This is, of course, nonsense! All the old security threats still apply, albeit sometimes in a slightly different guise. Hence the importance of having cybersecurity practitioners embedded in any AI group.

The biggest nontraditional security threat comes from the fact that in ML, data are algorithmic, meaning that the utmost care must be taken to secure the data. In particular, the provenance and integrity of data are critical to prevent adversaries from attacking a system by maliciously creating/altering data. While practices are improving, a great many ML models are built from open source data, which is not secure. In addition, ML systems, especially those for perception, are subject to adversarial attacks where a scene is slightly altered (for example, putting tape on highway signs), causing brittle systems to return answers that are wildly off. While a great deal of research has focused on how to ensure that ML-enabled systems are robust against such attacks, there remains a large gap between the research and practice.

ANOOP SINGHAL: The last several years have witnessed the rapidly increasing use of ML systems in multiple industry sectors. Auto-driving cars use ML to process the images/videos from cameras to understand the traffic signals and real-time traffic around them. ML has been used to translate text from one language to another in several systems. Deep learning has been used in products such as Google and Mozilla to understand speech.

It is widely recognized that the existing security analysis frameworks and techniques, which were developed to analyze enterprise (software) systems and networks, are not very suitable for analyzing ML systems. ML systems have new kinds of causality relationships that cannot be handled by current approaches for security analysis. For example, attack graphs are fundamental tools for enterprise security analysis but mainly focus on relationships between security vulnerabilities [such as Common Vulnerabilities and Exposures (CVEs)] and exploits (which mainly focus on newly gained permissions/accesses). In contrast, a good foundation for analyzing security issues in ML systems must also capture the causality relationships involved in data poisoning and evasion attacks using adversarial examples. It is clear that such causality relationships are not really relevant to traditional attacks that involve the exploitation of common vulnerabilities (CVEs).

Evasion attacks and data poisoning attacks can make ML systems misbehave. *Evasion attacks* refer to crafting adversarial examples after the training phase so that models produce incorrect outputs. *Data poisoning attacks* refer to modifying the training data so that the trained model will be maliciously altered.

There is a need to develop new techniques for modeling attacks on ML systems using causality graphs. These graphs can be used to capture the data, model, and library dependencies in a specific ML system:

SARA RAMPAZZI:

Main threats. From a cybersecurity perspective, there are two key

challenges when we rely on AI algorithms to make automatic critical decisions involving direct interactions with humans, such as in autonomous driving or health-care diagnostics. The first challenge is to overcome the assumption that AI models are trusted based only on the quality of their input data. This arises because we trust AI based on how well formed, complete, and unbiased the data used to train these models are. However, since we don't fully understand how these algorithms extract information and features from the training data, we cannot fully grasp the extent of the relationships and correlations they identify, especially in complex systems. This makes these algorithms prone to be manipulated by attackers as they exploit relationships hidden in the data.

The second challenge is creating standardized and widely accepted measures of the reliability and security of these models. Currently, the performance and accuracy of AI models are often mistakenly equated with their security and safety. An accurate model does not necessarily mean that it is secure and reliable for human use. From the cyberphysical system security perspective, we saw the emergence of two main threats as follows.

Overlooked interactions between AI, software, and hardware: AI algorithms used in autonomous or semiautonomous cyberphysical systems make decisions and provide predictions based on data captured and elaborated by hardware and low-level software. However, electronic components, sensor data, and software can be manipulated by attackers, leading to incorrect AI decisions. For instance, physical attacks use external signals to change sensor measurements and trick AI models into believing false information, even if the model is designed correctly. Security measures that do not consider the overall interactions among software, hardware, and AI model input/output correlations can be easily circumvented and exploited by attackers.

Lack of external feedback to validate the automatic decisions. Often, AI algorithms are only validated by themselves or evaluated by other AI models. This has been shown to generate loopholes and backdoors, especially in complex and sophisticated cyberphysical systems. Attackers can exploit the lack of proper external feedback as a point of access.

Measuring and testing. To quote Stuart Russell, one of the major contributors to AI, we should consider incorporating the concept of uncertainty where humans or the system itself is asking for external feedback (for example, by humans, other system components, or external systems) when the model is dealing with uncertain situations and scenarios instead of leaving the algorithm, taking immediate action, and then evaluating the final outcome. Efforts to regulate and standardize data collection procedures, along with the creation of universal safeguards and automated guidelines for designing and testing models (similar to the MISRA rules in the automotive domain), can help assess the security and safety of AI/ML systems without disclosing direct information about the model architecture and training data.

Another crucial point is to inform users about conformity and compliance with regulations and tests. This can incentivize manufacturers to address potential issues and help users build trust in this new technology. Similar approaches have been taken with safety-critical systems such as avionics and medical devices, and we should envision something similar for AI.

COMPUTER: What are the biggest challenges in managing risks in public systems?

PAYEL DAS:

Model landscape. Some of the concerns with AI systems and autonomous agents are the availability of broad attack surfaces (for example, a model interacting with models, code/tools, software, and data centers), which escalates their safety risk and trust considerations. Another related issue is the infusion/influence of a model on another model through the interaction during training (synthetic data), evaluation [large language model (LLM) as a judge], etc. This puts the autonomous agents at a very high risk of data contamination and biased testing. It is important to consider model testing in a dynamic humanized manner that considers the tradeoff between safety and utility.

Data landscape. The publicly available data can have strong bias and fairness issues. While current data curation techniques for generating training data for AI models do consider the use of bias/fairness/misinformation/hate/toxicity detectors, their definitions of those dimensions are static and therefore may not suit when there is a data drift. There are also risks in missing unexplored dimensions and/or defining those dimensions as per Western world standards.

Regulatory frameworks. I advocated for a U.S. Food and Drug Administration (FDA)-like government body for AI regulation, which will define safety and regulation dimensions and implement a plan around those. The panel discussed the European Union (EU) AI Act, which is ahead in AI regulation. I shared some first-hand feedback from business owners from the EU who were concerned about the strict rules and regulations described in the EU AI Act and the lack of clarity and transparency around them. The concern is also that as of now, the plan of action around the EU AI Act is to be defined, which puts businesses, particularly the low-margin ones, in a state of uncertainty.

Data privacy. We discussed the importance of model intention characterization before the model gets

access to private data. The goal would be to allow access only to the models that meet the safety standards. This again puts us back to the necessity of a regulatory body like the FDA for AI.

DAVID STRACUZZI:

Model landscape. One major risk in any AI application relates to trust. Ideally, we'd like our end users, particularly those who use AI-based systems to perform their daily jobs, to trust and rely on AI systems with the goal of accelerating and improving the quality of their work. The risk arises when we try to develop AI-based systems to engender that trust. Trust can be manipulated, often unintentionally, causing people to over-rely on AI and other data analysis systems, using the AI's judgment in place of their own even when unwarranted.

In practice, the goal of developing an AI-based system is to maximize the decision-making performance of the end user, with trust between the user and the AI arising as a side effect of good task performance and communication. AI systems must provide sufficient information to support the end user's decision-making process without manipulating their natural risk profile or bogging the user down in excess detail. Developing such AI-based systems will not be simple as recent research indicates that there are many mechanisms by which information systems can influence human users. Moreover, research also shows that maximizing the performance of the combined human-AI system can often involve tuning the AI to perform well against task-specific metrics as opposed to broad-based metrics such as accuracy or F1 traditionally used in the AI research literature. The AI community needs to work with cognitive scientists and experimental psychologists to develop improved human-AI interaction methods, along with methods for assessing human-AI system performance.

LANUS:

Privacy and security. Differential privacy, a mechanism for ensuring individual privacy by guaranteeing that the result of a query is not substantially dependent on any individual's data, has been employed for protecting privacy in ML systems. Often, privacy and utility are expected to be at odds as the idea is that sufficiently obscuring individual details leads to modeling an approximation of the function. However, some research suggests that, in fact, differential privacy applied to protect ML training data has the benefit of preventing overfitting. That is, the ML model should be learning a more general function to generalize to unseen data and avoid memorizing individual samples.

Addressing bias and fairness. What best practices should be adopted to mitigate bias in training datasets, particularly in health care and transportation? This depends on the type of bias involved. When the bias is a statistical skew associated with the function to be learned, mitigating bias may have the unwanted effect of destroying the utility of the model. In some cases, bias exists in the population to be modeled as the result of institutional bias, but it is noise and not the true function. In these scenarios, there are a variety of fairness metrics that can be utilized, such as disparate impact; however, this requires some domain knowledge, such as identifying groups that may be (un)favored or the identification of sensitive attributes or those that may be proxies for sensitive attributes. Doing this well requires an interdisciplinary team probably composed of an ML developer, a statistician or data scientist, and a social scientist who understands the population being modeled and harmful outcomes.

TRACY BANNON:

The use of AI-generated synthetic data in GAI-augmented systems introduces significant risks. When models rely on other models for training or validation, it creates a self-referential loop that can amplify inaccuracies and biases—a phenomenon I sometimes described as "incestuous AI." This practice risks degrading trust and reliability over time, especially in high-stakes applications. Understanding the specific context of the system—its purpose, users, and impacted populations—is paramount in mitigating the risks. AI systems are not one size fits all. Solutions, inclusive of embedded AI models and training data, must be designed to reflect the unique challenges and nuances of their intended environment.

Geocentric biases inherent in region-specific data can undermine fairness and equity in public-facing systems. Mitigating these risks requires an understanding of the representation gaps in datasets. It also requires that data scientists and engineers are committed to evaluating protected attributes and coverage. Human involvement in the design and evaluation process is critical to identifying biases and then addressing them in ways that align with the system's goals. What is the purpose of the system, who are the ultimate decision makers, and who are the impacted populations? For example, fairness-focused systems designed to improve equitable access to resources must define explicit measures to counter bias, such as adjusting data inputs or weighting outcomes to balance inequities. Conversely, statistical or exploratory models may require less correction but still need transparency about what is missing from the training dataset(s).

Balancing focused models and mega-models is another challenge. While larger models provide broad applicability, they often introduce systemic biases that are difficult to detect and correct. On the other hand, focused models tailored to specific applications may lack robustness in diverse scenarios. A contextualized approach—tailoring models to specific applications with a focus on high-quality comprehensive

training data—could mitigate these risks. Involving humans to validate outputs and alignments with the intended purpose may help navigate the tradeoffs between broad applicability and focused accuracy.

A final, though somewhat tangential, challenge lies in the tendency to anthropomorphize AI systems, treating them as human-like decision makers rather than probabilistic tools. This misplaced perception can lead to over-reliance on AI-generated outputs and the underestimation of their limitations. Clear communication about what AI systems can and cannot do—as well as educating stakeholders on the importance of human oversight in AI-driven decisions—is crucial to reducing misunderstandings. GAI tools must be seen as augmentations to human judgment, not replacements, especially in contexts where fairness, safety, or trust are critical.

COMPUTER: What are the biggest challenges in managing risks in government systems

GOODLOE:

Tradeoffs. Focusing on the realm of safety-critical systems, reinforcement learning is simply not an acceptable technology in this domain due to the difficulty in analyzing the behavior of such systems. The unpredictable nature of such systems poses a security risk, especially if the attacker can discern the objective function.

Hallucinations. For the case of AI-generated code containing hallucinated calls to functions and packages that do not exist, approaches such as static analysis tools may be helpful. Possibly harder to detect are cases where the wrong function is called. I once had to deal with the consequences of an application developer using the same name as a low-level system call. The linker would link in the system call, resulting in erratic behavior at execution. We had to educate our developers to be careful when choosing function names. I could easily see this happening in LLM-generated code as well.

Will we ever get there? Well, yes and no! AI experts rightfully advertise ML as the solution for problems where you can't precisely know what you want to build but can collect a lot of examples, and you are willing to accept periodic failure. Neither of `these are acceptable when we are talking about safety-critical systems. Yet there are many applications in these domains where we know how to build them using conventional means, but applying ML could yield significantly improved performance. In these cases, techniques like runtime assurance and novel approaches to verifying ML-enabled systems are robust and may be able to provide significant assurance. On the other hand, areas such as perception, where we have no idea how to characterize what the system is to do and what it isn't supposed to do, may remain indefinitely remain an open research question.

M S RAUNAK:

Tradeoffs. The risks associated with online or evolving models are likely to be higher compared to statically trained models. The standards, tests, and validation processes are easier to apply on a statically developed model. The tradeoff, of course, is that a static model can be more predictable and easier to verify and validate. However, it may lose its efficacy over time. On the other hand, a dynamic or online learning model will be changing continuously, making it very difficult to verify and validate. I expect the development and use of dynamic or online learning models to become more prevalent and dominant over time. Therefore, specific standards—as well as testing, validations, and monitoring strategies—need to be developed for these evolving or online learning models.

Hallucinations. Hallucinations, especially in AI-generated code, are one of the biggest sources of vulnerability. Code injection, taking advantage of hallucinated methods and packages, is an easy way to design an attack. Some of software engineering's traditional approaches toward code quality, such as code reviews, will remain relevant and may even become more important to guard against such attacks. Moreover, systematic thorough testing with special emphasis on boundary analysis and input space coverage will be useful mitigation methods against these vulnerabilities.

Red teaming. Red teaming is a well-accepted approach to address the risks of AI/ML systems. The October 2023 Presidential executive order on the safe, secure, and trustworthy development and use of AI specifically mentions the importance of using red teaming. It is an adversarial method where a dedicated team looks for weaknesses that the developers might have missed. While Red teaming is useful and should be used for discovering vulnerabilities, one should also be aware of its limitations. This approach is not exhaustive. Thus, it may provide a false sense of security. Systematic exploration of vulnerability through proper testing should not be avoided. One must also be conscious of the highly resource-intensive and costly nature of the approach. It takes a lot of time and effort to properly apply red teaming for AI/ML vulnerability identification. If one does not allow the required time and properly trained individuals with the right tools, red teaming will not serve its intended purpose.

Will we ever get there? We surely hope so. With enough resources and rigorous verification and validation approaches, we have been successful in producing ultra-reliable software in the traditional aviation industry. We need to develop similar metrics, measurement processes, and their

systematic and rigorous application in AI. However, we are far from that place now. I see two additional challenges while building reliable AI/ML systems, namely the inherent uncertainty revolving around their behavior and the potentially evolving nature of the systems. We need metrics and measurement processes that can address these characteristics. Novel approaches will be required.

BANNON:

Tradeoffs: One of the most pressing concerns for government systems is the risk of model training data vulnerabilities. Many of the training data for GAI are derived from open or public sources, which inherently carry risks of bias, inaccuracy, and intentional poisoning. The concept of model poisoning, recognized by research groups, including MITRE, and supported by the ATLAS framework, represents a serious threat vector. Poisoning can lead to models embedding both unintentional vulnerabilities and maliciously engineered behaviors. This issue is particularly critical for government systems, which range from human resources applications to scientific research to cyberphysical systems like weapons platforms and autonomous vehicles. Ensuring the integrity of model training data and implementing robust validation processes must be prioritized to mitigate these risks.

Hallucinations: The concept of hallucinations in GAI also demands a more nuanced perspective. Hallucinations are often mislabeled as "bugs," but they are a natural consequence of how GAI models generate output. These models create outputs by sampling from learned patterns and probabilities in their training data, which can result in novel combinations of information. While these outputs may seem nonsensical to humans in the loop, they are consistent with the model's probabilistic architecture.

Rather than dismissing hallucinations as errors, the focus should shift to better understanding when and why GAI generates outputs outside expected norms. This ties back to the discussions in Panel 2 about context and purpose; human involvement is critical not only for validating GAI outputs but also for interpreting and integrating them within the unique requirements of the system. Without this contextual lens, the risks of overreliance or misuse of GAI outputs—be it for content generation or augmenting human reasoning—will only grow.

> One immediate way to address some of the risks is to apply today's modern software practices, that is, DevSecOps principles.

Will we ever get there? The overemphasis on code generation in discussions about GAI in software engineering risks overshadowing broader, equally critical usage patterns. My research has identified two primary applications of GAI: content generation (including but not limited to code) and augmenting human reasoning. While code generation draws attention due to its immediate impact, there is insufficient focus on the necessary measures and metrics to evaluate GAI's effectiveness and risks across these broader use cases. For instance, while multiple studies from institutions like MIT and Stanford highlight security vulnerabilities in generated code, we lack comprehensive frameworks to assess GAI's contributions to augmenting human decision making—a crucial application in systems where high assurance and contextual reasoning are required. One immediate way to address some of the risks is to apply today's modern software practices, that is, DevSecOps principles. Embedding security and validation earlier in the software development lifecycle (SDLC) may reduce vulnerabilities introduced by GAI-generated outputs. Automated dependency validation and static analysis tools could identify potential security flaws in generated code before deployment.

STRACUZZI: Will we ever get there? My comments about understanding the goals of AI systems in terms of improving work and decision performance in humans apply equally to government systems. The main challenge is that the AI research and applications community does not yet have a set of general theories around system safety, reliability, security, or user interaction. As a result, we have to play "whack-a-mole" on all of these fronts, as Erin noted earlier. Part of the solution will come from continued research by academia, industry, and government labs to develop those theories.

Another part will come from the composition of the development teams that construct these public and government AI systems. The explosive growth in AI software packages and downloadable models across programming languages and platforms makes AI system development accessible to almost anyone. Everyone can experiment, which is great! However, developing an AI system for others to use, particularly when mistakes come with a cost, requires expertise along at least four dimensions.

AI. The software packages used to create AI-based systems often hide complex assumptions about training data, the deployment environment, and a host of other factors. While there are no guarantees, formal training and experience with AI algorithms and techniques increase the chances of success.

The domain. AI-based systems are often purpose built, incorporating deep knowledge and assumptions about the application that are not necessarily embedded in the training data. Managing this background information, identifying the needed performance metrics, and determining whether the resulting systems meet the needs of the application environment require input from a domain expert.

The human-AI interface. Joint decision making, as any married couple can attest, requires precise and careful communication, as discussed previously in greater detail. AI systems can produce more sophisticated behavior—and more complex justifications for that behavior—than many past software and automation systems. Special care and study are needed to ensure that AI systems have the intended impact on user task performance.

Security. AI-based systems, including the human-AI interface, create a novel set of attack surfaces that include and extend beyond traditional cybersecurity concerns. Anticipating and controlling these attack vectors are not a part of traditional training for most experts in AI, application domains, or human-AI interaction. Likewise, traditional cybersecurity concerns are also relevant to AI systems, as others have noted in greater detail.

By requiring that AI application development teams have adequate representation from these four distinct areas of expertise, we can at least ensure that the issues associated with each area have received consideration. Importantly, the intersections of these four areas each generate novel challenges and research needs. This is not to say that we should halt AI application development until these challenges can be fully resolved, but they do need to be addressed on a per-application basis (also known as "whack-a-mole") until the research community can develop the fundamental theories needed to anticipate and address the technical challenges robustly.

COMPUTER: What should be done for awareness, education, training, and certification?

GIRIJA SUBRAMANIAM: Data on the Internet that are being used to train many AI models are increasingly composed of data generated by AI models themselves. This has the effect of perpetuating and even amplifying the initial bias that existed in the human-generated data. This phenomenon is called a *data feedback loop* or *synthetic data feedback* and needs to be addressed urgently. The fact that 40% of the software code on GitHub is already AI generated shows the rapid rate at which AI is being adapted in the world of software. The exponential nature of AI development and its use implies that this problem will only accelerate in the days to come. We need policies and procedures to watermark not just software code generated using AI models but also data generated by AI. This, in turn, will aid in the selective use of data used to train AI models.

While AI might not replace all jobs, it will increasingly replace entry-level positions—with the need for human intelligence being reserved for complex tasks and experienced positions. It is now a well-known fact that the availability of GPS navigation on smartphones has impaired the ability of humans to navigate without digital aids and has reduced our spatial memory, visual orientation, and mental mapping skills. We can easily extrapolate these developments to other aspects of problem solving like writing code. The increasing abstraction of programming languages along with the advent of AI implies that the coding language of tomorrow is "plain English." Does this mean that the average person whose skills can be replicated by AI loses cognitive ability? How will this affect the ability to produce experienced software developers or architects if the entry-level positions and the opportunity to gain experience are lacking?

How will we determine if an AI-generated software/system is safe enough to be deployed in real-world systems? Will these be based on statistical benchmarks, and how will these benchmark values be identified? Is it sufficient if they are statistically safer than human-generated code? Even if that is the case, they might not be acceptable to the public at large. A classic example is the presence of train operators in high-speed trains even though the response time of a human is not sufficient to take timely action in case of an alarm or visual event.

JOANNA DEFRANCO: The highlights of this panel were about education and training. AI/ML education/training could fall into two buckets: 1) developers and 2) users. We could differentiate the objectives of each where developers need to gain knowledge and experience to design and develop safe AI-enabled systems, and users and society need to gain knowledge to understand the implications of the use of AI-enabled systems.

Training for users. The example I gave was the use of a life-saving medical device that a patient needs to configure for use. In addition to improving the management of a chronic illness, the "smart" devices could also assist in preventing an emergency, but the patient needs to be aware of how the algorithms work to manage the device settings. The patient "training" for these devices focuses on superficial use—without the consideration of edge cases, which can be dangerous for users. For example, an insulin pump's AI algorithm doesn't automatically adjust settings if the patient is about to exercise (which may cause a blood sugar low or an adrenaline rush, causing a blood sugar high) for a competitive or intense event. In other words, the device does not "know" how fast a person will run, how intense the game

will be, or what they ate before the game to manage those situations.

Required training for regulatory authorities and policymakers. Regulatory authorities and policymakers should be required to follow a set of training standards set by standards-setting organizations such as the National Institute of Standards and Technology (NIST), IEEE, the EU, etc. They need to understand the safety and privacy risks while writing policy. Policies should also enforce disclosure of specific AI risks on all products.

Education for developers. An additional consideration for all is requiring an ethics course as part of all general education curricula. The course could emphasize the need for a focus on risk, safety, security, and privacy in products.

RICK KUHN:

Awareness. One of the top concerns will be failure modes that are new or significantly different from these systems. AI/ML has different failure modes than conventional software, and attackers will find new ways to exploit them. These new failure modes are a security concern because fewer defenses are available.

Adversarial imaging is an example, where a change in an image that is not noticeable to a human causes the AI to recognize an object as something completely different (such as changing a stop sign to a speed limit sign). AI hallucinations are a related problem. For example, a study of software produced through GAI found that about 20% of the code contained calls to functions and open source packages that do not exist. This is a security problem if an attacker could access the generated code or could induce the GAI to produce calls to specific nonexistent functions. It would then be possible to compromise the system by adding a malicious function with the specified name to the relevant open source repository. There are undoubtedly many novel attacks related to the different characteristics of AI/ML. Identifying and defending against them will be a long-term challenge.

Education. As with all areas of technology, universities can cover the basics well, but keeping pace with a rapidly evolving field will be a challenge. Educators face these problems in many fields of technology and science and deal with them by keeping their research current and relevant and involving students with these advances as well.

Training. One of the most important points to get across to potential users is to be cautious and skeptical of the promises of GAI. There is a problem of "garbage in, gospel out" as people often place far too much confidence in systems that appear to be producing intelligent responses. Here again, the hallucination problem could lead to disaster in the fields cited, where a mistake—however rare—may be fatal. It may be best to encourage users to treat AI/ML as a tool that may save time but can easily lead to error and require constant checking of results.

Certification. The answer would depend primarily on whether they are offering their services to the public, where some certification is needed to demonstrate competence, in the same way as doctors and nurses have certifications required by law. AI/ML will primarily involve engineering, and about 20% of engineers have a professional engineer license. Civil engineers are often licensed as they offer their services to cities, businesses, and others who need some assurance of their competence. Conversely, aerospace engineers are rarely licensed because they are employed by governments or large aerospace firms that can judge their qualifications. It may be helpful to develop a licensing program for AI/ML engineers who offer their services to the public. Core competencies would presumably include the topics covered in an undergraduate program in AI/ML or data science.

SANDEEP NEEMA: The emergence of GAI provokes a fundamental challenge in education, particularly in disciplines like computer science where an important component of education and training is "learning by doing." In most computer science curricula, courses are accompanied by homework assignments, which require a student to solve a computing problem by developing an algorithm, programming the algorithm, deploying, testing, and validating in a context. There is a large range of such assignments specific to different subjects, different focus areas, and different universities, and while a systematic study has not been conducted yet, it is anecdotally evident that most of these assignments are solvable by GAI implementations such as ChatGPT and its variants. Different universities are coping with this rapid emergence in different ways—ranging from limiting the use of GAI to mutating the assignments toward low-resourced ecosystems (for example, using xv6 as a pedantic OS instead of Linux). However, most of these are stopgap and short-term solutions. A long-term viable solution would require a fundamental rethinking of the role of human (software or otherwise) engineers and a corresponding shift in curriculum and training for a future where GAI is integral to all engineering enterprises and human engineers are increasingly "on the loop," as opposed to "in the loop."

SUMMARY

AI is a transformative technology that will impact all aspects of human life. It is crucial for both technologists and end users to understand AI's core concepts, challenges, and opportunities. Developing standardized metrics and processes for AI safety and security, updating curricula, and fostering collaboration between academia, industry, and government are necessary for

the responsible development and use of AI/ML systems. This virtual roundtable provided a small step in that direction and I want to thank the panelists for their insightful comments.

SUBRAMANIAM: AI is a fundamental technology that will affect all aspects of human life and will eventually be used to generate code that is deployed in every industry known to man. It is critical that the technologist as well as the end user is educated about the core concepts, challenges, and opportunities that AI brings. A blind reliance on AI's output without understanding the how and why of the modeling process opens a pandora's box of errors and misunderstanding. Early education regarding the core concepts of AI along with real-world examples that undertake a comparative analysis of outputs from multiple AI models and tools will help educate the public about critical aspects of AI like bias, hallucinations, data privacy, and transparency.

BANNON: Context and principled practices are foundational to the adoption of GAI augmentation. Across these panels, the importance of aligning AI design, implementation, and validation with the system's intended purpose and operational environment was a consistent theme. The risks introduced by GAI—such as hallucinations, systemic bias, and misplaced anthropomorphization—underscore the need for transparency, rigorous validation, and continuous oversight. Clear and repeatable measurements are essential to assess GAI's performance, risks, and benefits, providing a foundation for trust and improvement.

Applying today's leading DevSecOps principles provides a pathway to address some of the quality and security challenges introduced by GAI augmentation. Practices such as early security checks, automated validation pipelines, and continuous monitoring embed resilience into GAI workflows, reducing vulnerabilities. These principles are as relevant to public and government systems as they are to transforming the SDLC.

Collaboration between government, industry, and academia remains critical. By leveraging robust validation methods, iterative improvements, and domain-specific frameworks, we may be better situated to deliver systems that are secure, reliable, and aligned with societal and operational needs. With thoughtful context-aware approaches and shared responsibility, the transformative potential of GAI augmentation may be better realized while effectively managing risk. GAI in the SDLC has groundbreaking potential with limitations and challenges that need to be understood and planned for.

STRACUZZI: Maybe the most important thing to remember about AI is that, until quite recently, it was a niche research community. Specific uses of AI have been filtering into industry and society for decades, but the last five to 10 years represent the first time that demand for cutting-edge AI has filtered into business, industry, and society at large. There's a notion in the AI research community that has been around since at least the early 1970s with uncertain attribution that AI is the collective name for problems that we do not know how to solve yet. Once we attain success in any domain, it ceases to be AI. The last decade or so has finally broken from that limiting perspective.

However, a side effect of that historical view is that the AI R&D community has limited experience in hardening and deploying cutting-edge technology. The current explosion of AI, deployed across society in its raw and relatively untested forms, represents a new paradigm. As others have argued here, managing this new paradigm of AI research, development, and deployment so that society can benefit while avoiding major risks and pitfalls will require immediate, extensive, and aggressive collaboration between academia, industry, and government. There does not appear to be an analogy from AI to any other new technology in terms of the combined rate of growth and breadth of impact. Our collective response requires equal doses of caution and practicality—for example, recognizing that AI applications are here to stay, that continued development is in U.S. national security interests, and that it may take decades for theory to catch up with practice. We need to do the best we can with existing development and security tools while pushing hard for continued improvement.

DeFRANCO: For experienced developers, AI should be used with caution as a tool or assistant. For new developers—learn the basics of software development; otherwise, you won't know what you don't know. In general, society should be cautious with AI output. We are in the MapQuest stage of navigating AI tools—not at the GPS stage where updated traffic patterns are provided while you drive.

RAMPAZZI: Like many other technological breakthroughs in the past, it is important to recognize that GAI/ML might not be the perfect solution to apply to everything at all costs. For those applications where AI can be beneficial, adopting a zero-trust security model ("never trust, always verify") can prevent dramatic setbacks when damage has already occurred. This is a fundamental step for adoption in safety-critical cyberphysical systems and infrastructures where a single mistake can endanger society and humans on a large scale. Collective effort in addressing our current limitations in understanding how AI models make decisions—and how software, hardware, and humans interact holistically—can help prevent AI-driven issues that might emerge years or even decades in the future.

LANUS: There is a common idea that AI/ML systems fail on their own without an adversary due to the lack of ability to generalize to new or changing environments, and so security is

an issue to tackle later. This attitude directly contradicts security by design principles; it can be costly to redesign a system to add security later after losses are incurred and an adversary may have gained persistent knowledge or access. While AI does present some new attack vectors, most of what we already know about cybersecurity is applicable to AI-enabled systems. Siloing AI systems as being different from other information systems and AI security as different from cybersecurity leads to the loss of gained expertise in securing a wide variety of systems, especially given that AI is almost always integrated into a larger information or cyberphysical system. We need to bridge the gap between cybersecurity and AI experts; both should receive foundational training in the other disciplines. We also need to think holistically about securing the entire system, including the entire ML development pipeline, with an essential focus on data security and carefully weigh risk from vulnerabilities against the performance cost of employing defense mechanisms.

RAUNAK: AI will inevitably encompass many aspects of our personal, professional, and societal lives. We are yet to know all the different ways it will impact us. There will need to be a concerted effort to ensure the security and safety of the AI/ML systems. One area that requires careful attention is the development of standardized metrics and processes to measure how safe and secure an AI/ML system is. New and rigorous approaches toward verification, validation, and assurance of these systems will be necessary. Developers and users of these new systems will need to be aware of the different ways these systems can fail or become vulnerable. Updates in the curricula will be required in higher education as well as training of the existing workforce. A pervasive discussion of ethical implications is also necessary throughout the Society. A close continuous collaboration and a concerted effort from academia, industry, and government will be needed to ensure the safe and secure development and use of these AI/ML systems.

PHIL LAPLANTE, State College, PA 16801 USA, is a computer scientist and software engineer, a Fellow of IEEE, and an associate editor in chief of *Computer*. Contact him at plaplante@psu.edu.

Digital Object Identifier 10.1109/MC.2025.3562087

IEEE COMPUTER SOCIETY

www.computer.org

PURPOSE: Engaging professionals from all areas of computing, the IEEE Computer Society sets the standard for education and engagement that fuels global technological advancement. Through conferences, publications, and programs, IEEE CS empowers, guides, and shapes the future of its members, and the greater industry, enabling new opportunities to better serve our world.

OMBUDSMAN: Contact ombudsman@computer.org.

CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

PUBLICATIONS AND ACTIVITIES

***Computer*:** The flagship publication of the IEEE Computer Society, *Computer*, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

Periodicals: The IEEE CS publishes 12 magazines, 18 journals

Conference Proceedings & Books: Conference Publishing Services publishes more than 275 titles every year.

Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.

Technical Communities: TCs provide professional interaction in more than 30 technical areas and directly influence computer engineering conferences and publications.

Conferences/Education: The IEEE CS holds more than 215 conferences each year and sponsors many educational activities, including computing science accreditation.

Certifications: The IEEE CS offers three software developer credentials.

AVAILABLE INFORMATION

To check membership status, report an address change, or obtain information, contact help@computer.org.

IEEE COMPUTER SOCIETY OFFICES

WASHINGTON, D.C.:
2001 L St., Ste. 700,
Washington, D.C. 20036-4928
Phone: +1 202 371 0101
Fax: +1 202 728 9614
Email: help@computer.org

LOS ALAMITOS:
10662 Los Vaqueros Cir.,
Los Alamitos, CA 90720
Phone: +1 714 821 8380
Email: help@computer.org

IEEE CS EXECUTIVE STAFF

Executive Director: Melissa Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership Development: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

Digital Object Identifier 10.1109/MC.2025.3557143

IEEE

revised 14 January 2025

IEEE CS EXECUTIVE COMMITTEE

President: Hironori Washizaki
President-Elect: Grace A. Lewis
Past President: Jyotika Athavale
Vice President: Nils Aschenbruck
Secretary: Yoshiko Yasuda
Treasurer: Darren Galpin
VP, Member & Geographic Activities: Andrew Seely
VP, Professional & Educational Activities: Cyril Onwubiko
VP, Publications: Charles (Chuck) Hansen
VP, Standards Activities: Edward Au
VP, Technical & Conference Activities: Terry Benzel
2025–2026 IEEE Division VIII Director: Cecilia Metra
2024–2025 IEEE Division V Director: Christina M. Schober
2025 IEEE Division V Director-Elect: Leila De Floriani

IEEE CS BOARD OF GOVERNORS

Term Expiring 2025:
İlkay Altintaş, Mike Hinchey, Joaquim Jorge, Rick Kazman, Carolyn McGregor, Andrew Seely

Term Expiring 2026:
Megha Ben, Terry Benzel, Mrinal Karvir, Andreas Reinhardt, Deborah Silver, Yoshiko Yasuda

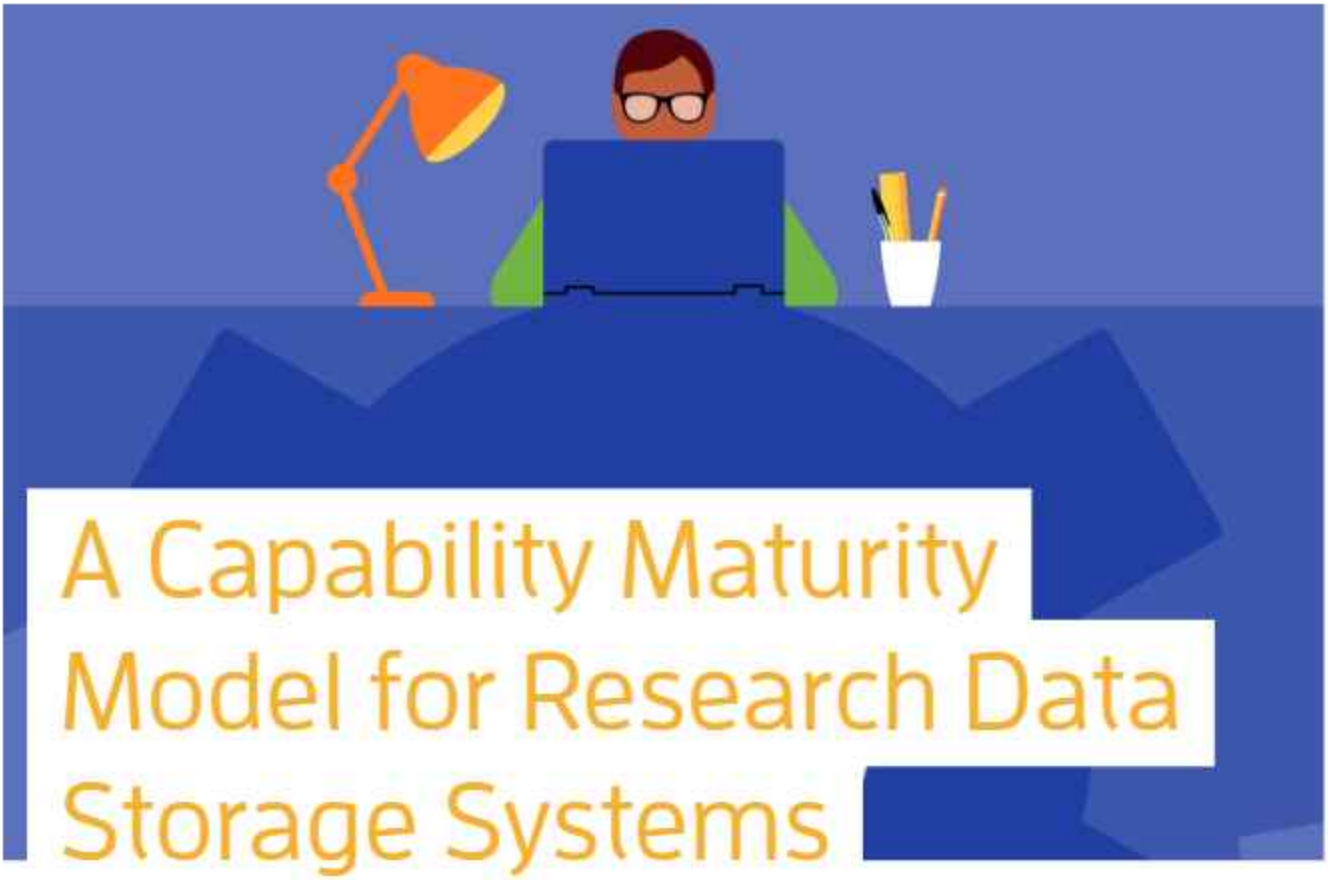
Term Expiring 2027:
Sven Dickinson, Alfredo Goldman, Daniel S. Katz, Yuhong Liu, Ladan Tahvildari, Damla Turgut

IEEE EXECUTIVE STAFF

Executive Director and COO: Sophia Muirhead
General Counsel and Chief Compliance Officer: Ahsaki Benion
Chief Human Resources Officer: Cheri N. Collins Wideman
Managing Director, IEEE-USA: Russell Harrison
Chief Marketing Officer: Karen L. Hawkins
Managing Director, Publications: Steven Heffner
Staff Executive, Corporate Activities: Donna Hourican
Managing Director, Member and Geographic Activities: Cecelia Jankowski
Chief of Staff to the Executive Director: Kelly Lorne
Managing Director, Educational Activities: Jamie Moesch
IEEE Standards Association Managing Director: Alpesh Shah
Chief Financial Officer: Thomas Siegert
Chief Information Digital Officer: Jeff Strohschein
Managing Director, Conferences, Events, and Experiences: Marie Hunter
Managing Director, Technical Activities: Mojdeh Bahar

IEEE OFFICERS

President & CEO: Kathleen A. Kramer
President-Elect: Mary Ellen Randall
Past President: Thomas M. Coughlin
Director & Secretary: Forrest D. Wright
Director & Treasurer: Gerardo Barbosa
Director & VP, Publication Services & Products: W. Clem Karl
Director & VP, Educational Activities: Timothy P. Kurzweg
Director & VP, Membership and Geographic Activities: Antonio Luque
Director & President, Standards Association: Gary R. Hoffman
Director & VP, Technical Activities: Dalma Novak
Director & President, IEEE-USA: Timothy T. Lee



A Capability Maturity Model for Research Data Storage Systems

David Abramson and **Jake Carroll**, The University of Queensland

In this article, we contribute a new Capability Maturity Model that allows organizations to assess their specific requirements, including projected growth, risk profile, and budgets. The intended audience includes senior research/academics, CIOs, CDOs, and those responsible for implementing operational infrastructure in research organizations.

Access to high-quality data is revolutionizing science and research in general, delivering new results from existing data, building evidence to existing theories, and producing new computational methods. For example, recent infrastructures that enabled the measurement of gravity waves not only built cases to support Einstein's theory of relativity but also opened entirely new research methods on gravity wave physics. Somewhat contentiously, previous theoretical models based on partial differential equations are selectively being replaced by models that have learned a response surface from experimental data.[1] In fields as diverse as the humanities, new computational methods based on the growing amount of human-centric data are enabling novel research avenues.[2] These examples require advanced data infrastructures that can respond to increasing demands for high performance and scale, as well as support rich access models to increasingly complex data.

To date, the implementation of research data platforms has largely advanced in an ad hoc way, often driven by the

Digital Object Identifier 10.1109/MC.2025.3540093
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

urgent need to deliver operational infrastructure within a constrained budget and sometimes driven more by what is available in the market than what would provide a powerful, flexible, and extensible system. Without a powerful underpinning model, practitioners often build systems that only meet a subset of the requirements, do not interoperate with each other, and often scale poorly. This leads to diminished organizational outcomes due to missing or unreliable operational capability and a lack of strategic alignment.[3] We acknowledge that organizations have different starting points, and consequentially, we provide a continuum of capability in each underpinning research data reference architecture (RDRA) component. This approach can be thought of as a complementary precursor to increasingly accepted research data governance models.[4]

RDRA FEATURES

In this section we enumerate eight features and two subfeatures that make a data storage and access system suitable for research data. Enumerated in prior work, this minimum viable set of features forms the RDRA[5] and subsequently drives implementation choices.

Resilience

A resilient system provides assurance that data will persist for an agreed retention period according to legislative obligations, academic conventions, or independent evidential validation. This is particularly important for research infrastructures because there is an increasing expectation that some data, either captured from the real world or generated by computations, will be available for reprocessing both by the original research team and, in time, other researchers. These retention periods can vary, often exceeding standard equipment refresh cycles. Without appropriate lifecycle management, the technologies used to read the media on which research data are stored may become unobtainable, thus rendering the content lost.[6] Reproducibility is a foundational tenet of good research practice and therefore requires a resilient data store. Codes of conduct are directing researchers to ensure data are available and open for independent validation. Also, as data citations increase, it is important that a data referent remains available (unmodified) if an external reference has been created. Resilience can also be expressed by the measurement of availability. Availability is often defined by the uptime and quality of service metrics, including the number of days without service outages, recovery time objectives, and recovery point objectives.

Discoverability

A discoverable system is one where data can be found easily, even if they are not openly available. In many cases, data form the evidential base of research and are important assets of an organization. Just as companies maintain asset registers for their capital equipment and intellectual property, catalogues of data need to be created and maintained. A significant number of research organizations have absolutely no idea what data they hold, partly because they are held in multiple distributed storage systems but mainly because there is a lack of catalogue services. Certainly, some of the largest research data storage facilities now see discoverability as paramount to maintaining a sustainable ongoing operation.[7] Maintaining sufficient metadata to attribute and track changes has the additional advantage of making it possible to record the provenance of data. This is becoming an increasingly important topic in research data workflows, repositories, and digital librarianship.

Manageability

Given the increasing volume and scale of research data, they need to be organized to make them manageable. Storage can be allocated in arbitrary units, with varying granularity. Traditionally, file systems allocate storage in single file units, but these can be clustered into directories when multiple files share similar properties. With research data, a common pattern is to store multiple objects or files for a single project. For example, a research project that works on population-level genomes might store its data as a single "collection" for the population. Prior work addresses the challenges in various file system layouts and the efficiency and accessibility of different design choices.[8] Because of this, we argue that data are more manageable if "collections" are the unit of allocation. Collections can contain multiple files (or objects) and have shared metadata at the collection level. We argue that this level of granularity is sufficient and flexible enough to apply to any research domain.

Accessibility

Research is increasingly collaborative with research data having increasingly complex accessibility requirements across multiple individuals, groups, and organizations. Consequently, access control features, applied on at least the collection level, are not trivial and require maintenance. Furthermore, different users might have different access rights. For example, the lead chief investigator of a project might have the ability to read, write, and delete files in a collection, but collaborators may only have the rights to read the files.

Governable

Research organizations need to build and follow procedures and processes to decide whether data allocations can be made, whether access rights can be changed, and ultimately if or when data can be deleted. Roles, such as data custodian, data owner, and the like, have different rights and responsibilities, and these need to be reflected in data governance concepts. For example, a data owner may be allowed to change the access rights to a collection, but a data custodian may be the only person who can authorize deletion. It is important that the governance function is integrated into the research data storage architecture from the beginning to enable traceability, auditability, and reporting, rather than being an afterthought. Legislative and sociocultural requirements may include collective benefit, authority to control, responsibility, and ethics principles for the management of Indigenous data[9] that must also be accommodated from the beginning of a research data storage system (RDSS) initiative.

Scalability

Scalability in research data can often be characterized as small numbers of large-scale collections through to large numbers of small-scale collections. Storage infrastructures can start out small but invariably grow over time. This occurs for many reasons, including technology improvements driving greater ability to gather and generate data. Additionally, institutional growth is generally associated with increasing data generating projects and researchers. Finally, as more traditional research areas become digitized, so too does the number of data collections. Data stores can grow to many petabytes, and at this scale it is usually not cost-effective to store data using a single technology. It may seem obvious that any storage architecture should be efficient, but the increasing scale and complexity of research data make this an imperative. Thus, in many real-world research stores of significant scale, multiple storage tiers are used to keep active data on faster but more expensive media. Less active data can be held in slower, cheaper, and more energy-efficient technologies. Minimizing the capital spent on infrastructure may free up funding to be spent on the actual research. Another aspect of scalability is the ability to grow the infrastructure sustainably as demands increase. For example, increases in the user base might require additional network bandwidth or storage throughput.

Versatility

Research pipelines are heterogenous, and no single processing platform will suffice to support all domains. For example, hypersonic engineers might make extensive use of high-performance supercomputers, but humanities researchers may choose cloud platforms or desktops. These platforms have different access requirements: supercomputers usually need high-performance parallel file systems, whereas cloud platforms either use web or object-based storage systems. Mobile systems may also make use of cloud-based sync-and-share platforms. As a result, an architecture must deliver data using a variety of protocols and techniques that minimize unnecessary data movement and maximize storage efficiency.[10] An RDSS should be flexible enough to offer different classes of performance by using the appropriate protocol depending upon the workflow and research question.

Security

An RDSS needs to be both logically and physically secure to avoid inadvertent leakage of information that could compromise ethical or privacy obligations or primary investigator privileges or impede accepted academic conventions of independent validation. Security levels vary across different organizations, domains, and projects, from completely open to completely closed. Several security levels have been proposed for research data (such as open, protected, sensitive, etc.), and these must be enforced at the most basic levels of an architecture. Moreover, the security classifications must be enforced through protocols and access pathways, often at an individual user level and by the actual storage platforms. Related to this is the principle of least privilege[11] and the minimization of "superusers."

SECONDARY FEATURES

In the "RDRA Features" section, we discussed a minimum set of features that are required to form our RDRA. Here we add two additional secondary features that, while important in their own right, can be derived from the primary eight features.

Citability

An important tenant of good research is to cite sources, and increasingly, data management must include the ability to cite datasets, as espoused by various open science and data initiatives. A data artifact is citable 1) if it can be discovered and if there is 2) a namespace and 3) an access protocol. Each of these properties is covered under discoverability, manageability, and versatility. Accordingly, any system that possesses these three primary features should also support the ability to cite data.

Interoperability

Research is frequently a collaborative process carried out across institutions, states, and among nations. For collaboration to be successful, standards of communication or protocols must exist. For systems to communicate, interoperability must be present. If 1) versatility and 2) discoverability exist in an implementation, systems can interoperate. By combining protocols from versatility and self-descriptive records, structures, and minimum viable metadata present in discoverability, systems can interoperate. Increasingly, scientific communities are using repositories to facilitate interoperability among sites.[12]

A CAPABILITY MATURITY MODEL

Within each of the features presented earlier, there are different service levels, and not all systems must implement the most stringent levels to meet the RDRA. Considering this flexible approach, the five foundational steps in the original IEEE Capability Maturity Model (CMM)[13] are readily adapted to the RDRA. Figure 1 provides a matrix, mapping the IEEE CMM to our RDRA.

The RDRA CMM proposed here allows an organization to tailor their implementation. Choices may account for tolerance to different levels of risk. In other cases, the CMM allows an organization to tailor its storage infrastructure to match specific technology choices and platforms. Each feature is graded on a five-point scale, and then the overall system maturity is calculated as the arithmetic mean of these scores. This facilitates a solution with varying maturity metrics across the RDRA feature set. For example, an implementation may possess high availability against resilience but have very few access protocols in versatility. Thus, organizations are not compelled to achieve CM = 5 across all features. Figure 2 demonstrates different sites exhibiting unique maturity profiles across the RDRA CMM.

The RDRA CMM extends the software engineering CMM as it uses more than one factor to determine the overall maturity level, while

IEEE Software CMM v1.1	RDRA CMM	Resilience at CMx	Discoverability at CMx	Manageability at CMx	Accessibility at CMx	Governance at CMx	Scalability at CMx	Versatility at CMx	Security at CMx
Optimized (CM = 5)	RDRA CM = 5	• HA • Multisite DR • Data integrity validation • Geographic durability • Version control • Geographic retention	• Repository integration • Web-entered searchability • RBAC search + discovery • Internal + external metadata search • API for search • Discoverability automation • Open access platform integration • Nationally federated	• Automated provisioning • Triggering from other workflows • Automated metadata collection • API driven • Governance level validation • Documented and enforced standard of collection granularity • Policy enforcement	• RBAC • Multi-institution IDM • Single metadata management • Multi-institution namespace mapping and harmonization	• RBAC per component • Comprehensive event audit • Provenance framework • Policy engine environment for metadata reading to record changes • Reporting: data classification, criticality, metadata, user behavior	• Multitier • Data/metadata driven • Software defined storage • Automated data movement connected to use case • Advanced economic, financial and TCO modeling	• Multiprotocol • Single namespace view of all data, irrespective of protocol • No designated ROSS in institution	• Adherence to and certification of standards (ISO, NIST, RAP, ABD BW) • Real-time observability • Event-based alerting • SLP (data loss prevention) • Audit outputs via API to institutional SIEM
Managed (CM = 4)	RDRA CM = 4	• Multisite DR • Data integrity validation • Geographic durability • Version control	• Repository integration • RBAC search + discovery • Web-entered searchability • Internal + external metadata search	• Automated provisioning • Triggering from other workflows • Documented and enforced standard of collection granularity • Policy enforcement	• RBAC • Multi-institution IDM • Single metadata management • Steps towards single namespace	• RBAC per component • Partial event audit • Provenance framework • Policy engine environment for metadata reading to record changes	• Multitier • Data/metadata driven • Software defined storage • Automated data movement connected to use case • Dynamic economic, financial and TCO modeling	• Multiprotocol • Single namespace view of all data, irrespective of protocol • Minimal designated ROSS in institution	• Adherence to a standard (ISO, NIST, RAP, ABD BW) • Event-based alerting • SLP (data loss prevention) • Steps towards institutional BW integration
Defined (CM = 3)	RDRA CM = 3	• Multisite DR • Data integrity validation • Version control	• Basic browse repository integration • Web-entered searchability • Internal metadata search	• Steps towards automated provisioning • Documented standard of collection granularity • Documented manual provisioning	• RBAC • Steps towards multi-institution IDM	• RBAC per component • Basic review event audit	• Basic towards multitier • Predominantly single technology/media type • Basic towards finance and TCO planning	• Web accessible or three protocols • Single namespace view of all data, irrespective of protocol • Minimal designated ROSS in institution	• Steps towards adherence to a standard (ISO, NIST, RAP, ABD BW) • Event-based alerting • Passive monitoring of events to data lake
Repeatable (CM = 2)	RDRA CM = 2	• Multisite DR	• Web-entered searchability • Internal metadata search	• Manual processing	• RBAC • Single institution IDM	• RBAC per component	• Predominantly single technology/media type	• Supports one protocol • Basic connection to single namespace view of all data, irrespective of protocol • No significant number of designated ROSS in institution	• Event-based alerting • Passive monitoring of events to data lake
Initial (CM = 1)	RDRA CM = 1	• Single site	• Manual, internal metadata location per project basis	• Ad hoc to heavy processing	• Single institution IDM	• Steps towards RBAC per component	• Single technology, standalone	• Supports one protocol • Wide siloed environment of designated ROSS in institution	• Passive monitoring of events

FIGURE 1. Mapping IEEE CMM v1.1 to RDRA CMM. API: application programming interface; IDM: identity management platform; DR: disaster recovery; HA: high availability; BRBAC: role-based access control; SIEM: security information and event management; TCO: total cost of ownership.

supporting varying degrees of maturity across individual features.

Here we analyze each feature and discuss the range of implementing options and service levels on that scale.

Resilience

As discussed, RDSSs require a high degree of resilience. Therefore, they need to accommodate different types of failures, and these can be managed by different capabilities, as summarized in Table 1.

The most significant risk involves a catastrophic system failure in which data storage infrastructure is damaged irreparably and to an unrecoverable extent. Such damage can occur through damage to the infrastructure through fire, water, chemical, or data center environmental issues such as humidity or dump-gas suppression events. One mitigating capability involves using only data centers with very high levels of assurance, such that they are largely immune to external damage, flooding, and other hazards. However, even when such hardened infrastructure and associated facilities are available, the most used mitigation strategy involves physical replication of the data, usually in the form of a remote data center some distance from the primary site. High-speed networks support continuous replication of data to multiple sites, such that if one data center becomes inoperable, another can take over operation.

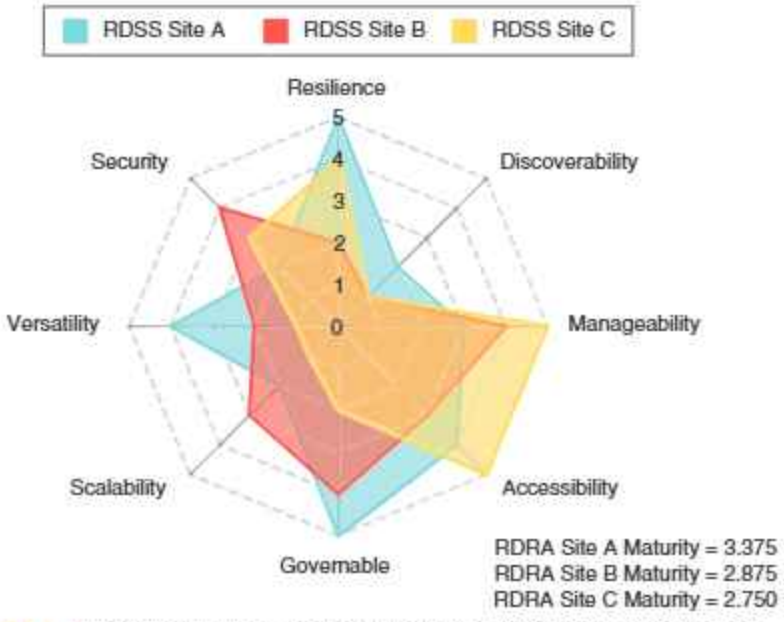


FIGURE 2. The RDRA CMM five-point scale representing the maturity of various sites.

TABLE 1. Resilience functionality and capabilities.

Functionality	Mitigating capability
Catastrophic system failure	Hardened environments, physical replication
Accidental data loss	Backup, versioning
Malicious data loss	Backup, versioning
Device failure	RAID/error correction codes, physical replication
Data corruption	Versioning, backup
Loss of name mapping	Resilient name space mapping

Two types of data loss can occur: either accidental or malicious. The former usually involves users (or scripts) accidentally deleting files, for example, while the latter usually involves an external actor or malware deliberately destroying data. While data can sometimes be recovered using the data replication processes discussed above, risk mitigation usually involves either data backup, in which files are regularly copied to another data device or partition, or file versioning, in which a versioned file system replicates a file every time it is modified or deleted. Both capabilities cannot mitigate catastrophic system failure but are very effective for partial loss. Techniques to mitigate risk at the facility level are well rehearsed and robustly documented.[14]

Complete or partial device failure can cause the loss of a subset of files. This can occur when a whole unit, such as a spinning disk drive, fails or when bits are corrupted through cosmic rays or Redundant Array of Independent Disks (RAID) write holes. Common mitigation capabilities involve error detecting, correcting, and erasure codes, as

routinely used in hardware RAID systems and advanced software defined file systems.

Data can also be corrupted at a finer granularity when individual records or blocks of a file are damaged. This can have several causes, ranging from hardware or software errors through to media or carrier transmission errors, but the risk can again be mitigated with a combination of a versioned file system and data integrity validation techniques.

Finally, there can be a loss of namespace mapping between names and files, which occurs when filename descriptor tables are damaged or lost. Over the years, a variety of capabilities such as those used by the Unix fsck command can be used to rebuild mapping tables, through to complete replication of the metadata catalogues themselves.

Of course, it is impossible to guarantee there will never be a data loss, so all storage providers quote a reliability metric, which is always less than one. For example, Amazon quotes "99.999999999% (11 nines) data durability." As explained in the "RDRA Features" section, a further aspect of resilience is how available data are. Providers can make choices about the level of availability they are willing to offer their organizations, depending upon the criticality and priority of the data. An organization that has sensitivity to latency may choose to put data onto high-speed storage platforms, whereas another organization might choose to put the same data into a slower, more economical storage platform depending on the workload and research data workflow. Another aspect of resilience is availability. Availability complements our earlier discussion on scalability with respect to where data may reside in storage systems for optimal accessibility and usability.

Discoverability

Ever since the first file system, data have been discoverable by associating user-defined file names to the storage blocks. While these metadata are primitive, they allow users to use textual strings that indicate what is in a file, and users can search for strings that match certain characteristics. More advanced commands such as the Unix "find" command extend this to other metadata such as creation time, access time, etc.

Increasingly, researchers want to record additional metadata, such as experimental protocols and laboratory conditions, project identifiers, instrument parameters, etc. They then wish to search using these additional fields, and these capabilities are provided by a range of domain specific repositories such as XNAT,[15] OMERO,[16] etc., and more general repositories such as IRODS.[17]

Both capabilities can operate on the spectrum from individual researchers to the entire organization. When expanded to the organization, a new capability is enabled, namely, other researchers and managers can discover data assets. This can greatly improve overall data management and reuse in the organization because the scale and nature of the data are exposed.

Finally, exposing the metadata outside an organization enables the implementation of a findable, accessible, interoperable, and reusable data management framework because other researchers outside the organization can discover what datasets are available.

Of course, there are also levels of control one may impose on who can discover data. While the examples given in Table 2 suggest wide permissions (such as a whole organization), a finer grain of control can be implemented. For example, a role-based security model may be used to restrict discovery features to those with a need to know and with the appropriate authority.

Manageability

As discussed, the increasing volume and scale of research data call for simplified management processes. One way of managing storage on traditional file systems is to simply create a directory as the unit of allocation. This works for small research data infrastructures and can implement arbitrary organization polices. For example, collaborating researchers can agree on a structure and naming convention for their data and simply create a shared folder. On the other hand, an organizational view in which data are associated with projects can be enforced with a naming convention that indicates which directories are associated with which projects. While this capability is effective, it can be open to misuse and may be subject to

TABLE 2. Discoverability functionality and capabilities.

Functionality	Capability
Basic metadata	Traditional file system
Advanced metadata	Repository
Metadata search within institution	Institutional catalogue, policy constraints
External metadata search	Open repositories

error. It relies on documented protocols that are manually implemented by individual researchers (Table 3).

Clearly, automating the creation of collections is more secure and less likely to fail. A system capability that allows a user to request a collection after which storage is allocated automatically by software would make collection management less error prone and simpler. It also makes it possible to capture metadata pertaining to the collection automatically, providing useful information for the discovery service. Thus, items such as researchers' names, field of research codes, and the like can be gathered at creation time and bound up in an appropriate catalogue.

Finally, collection creation can sometimes be an automated side effect of some other process. For example, if an organization has a centralized project management system, a collection could be created automatically when a project is documented, on the assumption that all research projects will require at least one data collection. This requires the collection management system, described in the last paragraph, to have an application programming interface that can be invoked by other software systems, rather than just through an interaction with users (say, via a Web portal).

Accessibility

While modern research is often collaborative, different collaboration modes require varying infrastructure capabilities. For example, if a project only has a small number of investigators, all of whom work in the same organizational unit (for example, a university department), then a traditional network attached storage system (NAS) can be adequate. These can enforce restrictions that limit access to one or more individuals or members of an arbitrary group. Likewise, access modes such as read, write, delete, etc., can be controlled at a fine grain.

TABLE 3. Manageability functionality and capabilities.

Functionality	Capabilities
Manual creation	Protocol to have shared directories created by users
User-driven creation	System creates collections, can capture metadata as in the discussion on discoverability, ownership, links to projects management system
System driven	System creates collections as side effect of other actions

TABLE 4. Accessibility functionality and capabilities.

Functionality	Capability
Individual or small group collaborations	Network attached storage and shared file systems
Multiple researchers with an organization	Organization-wide authentication
Multiple researchers from independent organizations	Global user namespace and authentication system, identity-based, role-based, token-based

However, projects that have multiple investigators across the organization might require a more powerful capability. In this case, access to data needs to be authenticated, arbitrary usernames within the organization and the file access mechanisms need to allow the data to be exported to a range of distributed platforms (Table 4).

The most substantial form of collaboration concerns multiple investigators from multiple organizations. In this case, authentication needs to be controlled by some sort of access control list involving multiple usernames from a global namespace. This may be controlled via an intermediary or an identity management platform. This capability is found in most cloud-based sharing platforms, although various models still require a homogenous namespace (for example, Google docs can only be shared among Google usernames), although a variety of mechanisms such as unforgeable links can also be used.

Governable

Governance is multifaceted. Some aspects of governance (and good governance in particular) require the capability to trace operations. Others require adherence to corporate or organizational policies and protocols. It is difficult to list all capabilities that make data governable, but Table 5 itemizes some of the more important ones.

Recently, attention has focused on mechanisms that can enhance the repeatability of research, including data and experiments. Studies have demonstrated that many experiments cannot be repeated, calling into question any conclusions drawn from analysis. Whether by poor practice or scientific fraud, the inability to recreate or validate research findings is of great

concern. At the other end of the spectrum, research data often underpin corporate intellectual property, and damage from malicious actors needs to be detected early and quickly addressed. The types of controls that can limit damage from actions like these suggest role-based governance mechanisms are required. Only actors with the correct roles and attributes should be allowed to manipulate research data, and these rights should be constrained to as few individuals as possible.

Likewise, the ability to trace operations at a collection or file level is an important governance feature that enables provenance features for widely used or contributed data collections or forensic audit. In either case, the ability to detect who accessed the data, what they did, and when they did it is essential. This is often enabled with a provenance framework in general, and file-based audit capabilities need to record every operation on every file. In some systems, these mechanisms are built into the file systems themselves.

Different organizations have different processes and protocols in place to protect research data. For example, an organization often writes policies about research data ownership, protection, governance, etc., into a human readable policy library. If such a library could be processed by a policy engine, then some of these protocols could be enforced and checked in real time.

One class of policy concerns data classifications. Some organizations regard all their data as private and protected, but publicly funded universities usually also support other classifications such as openly available, sensitive, official, private, etc. These classifications are best managed by a governance engine that can enforce and check all access requests as they occur.

The need to be able to report on how research data infrastructure is used is also an important governance capability. Currently, very few organizations can report on the types of data they hold, how this links to funding, and which fields of research are supported by the data. If such information is recorded against each collection, in the metadata catalogue mentioned in the discussion on discoverability, then it becomes relatively easy to report against those criteria.

Scalability

While research data stores may start out small, they invariably grow. A variety of technology capabilities can deliver efficient scalability. Three categories of scale have been singled out, and without locking these to a particular number of petabytes, they have been designated as small, medium, and large. Individual organizations and technological choices can determine which capability is the most cost-effective for a given organization.

At the small scale, it may indeed be possible to manage with a single, high-speed tier of storage, as delivered possibly through an NAS solution. However, it is likely that even small organizations will soon need to migrate to a medium scale in which multiple tiers are used, a typically small high-speed tier closely connected to compute capabilities and a slower and larger device for cold storage. The latter is usually much cheaper per byte than the high-speed tier and can also serve as a data archive. At this medium scale, the rate of movement among tiers can often

TABLE 5. Governance functionality and capabilities.

Functionality	Capability
Control operations on research data infrastructure	Role-based governance
Trace operations on collections or individual file basis	Provenance framework, file-based audit
Determine if operations are policy complaint	Policy engine
Enforce collection access controls	Record and enforce data access classification (sensitive, public, etc.)
Report usage and data classifications	Enhanced collection-based metadata and metadata catalogue

TABLE 6. Scalability capabilities.

Scale	Capability
Small scale	Single tier
Medium scale	Multitier with manual data movement among tiers
Large scale	Multitier with automatic movement among tiers

ABOUT THE AUTHORS

DAVID ABRAMSON is a computer scientist at the University of Queensland, Brisbane 4072, Australia. His research interests include computer architecture and high-performance computing research. Abramson received a bachelor of science (honors), a doctor of philosophy, and a doctor of science from Monash University. He is a Fellow of IEEE, a Fellow of the Association for Computing Machinery, a Fellow of the Australian Academy of Technological Sciences and Engineering, and a Fellow of the Australian Computer Society. Contact him at david.abramson@uq.edu.au.

JAKE CARROLL is the director of the Research Computing Centre and a doctoral candidate in the School of Electrical Engineering and Computer Science, University of Queensland, Brisbane 4072, Australia. His research interests include large-scale architecture, systems, and design of digital scientific research infrastructure. Carroll received a bachelor of information technology (honors) from Southern Cross University and an MBA from the University of Queensland. He is a Graduate Student Member of IEEE. Contact him at jake.carroll@uq.edu.au.

be managed manually by users. They explicitly copy data to the high-speed store for use and copy it back to the cold store afterward (Table 6).

However, as the amount of data moves to a larger scale, such manual movement becomes error prone and onerous. Users forget to move data out of the expensive, high-speed store and often replicate the files, and the task of copying data becomes labor intensive. For this reason, many large-scale storage systems have hierarchical storage management features that migrate data automatically.[18] Often, a policy engine controls this data movement, making the system very powerful, lowering the management cost and simplifying data management for researchers.

Versatility

The RDRA argued that a versatile research data infrastructure needs to support multiple protocols depending on how the data are presented to the computational platforms. Depending on the number and type of protocols supported, the system can choose different implementation techniques. For example, if data are only ever accessed on cloud-based virtual machines or containers, then only one protocol might be required. This, in turn, means the data management hardware only needs to implement one protocol.

Some real-world platforms can support a few protocols concurrently on the same hardware, but very few support the complete range in use today. Thus, if an organization requires a wider range of access protocols, it must be prepared to use multiple independent storage systems, tailored to a particular protocol, or find a mechanism for presenting data through multiple protocol engines. That is, if an organization chooses not to use a technology that can support multiple protocols, it may be explicitly creating multiple independent namespaces and unrelated data stores. Of course, the disadvantage of implementing multiple independent silos is that it becomes necessary to copy data among them, increasing the resources used and the complexity in managing data.

Security

The adoption of a security posture for data storage infrastructure and data storage assets is ultimately the responsibility of each institution. However, there are a number of security frameworks that can be adopted, such as the ISO/IEC 27000 family, the NIST cybersecurity framework, and various government-endorsed information security manuals, such as the Australian Information Security Manual. The common aim for each of these is to apply a risk-based approach to establishing, assessing, and continually improving the requisite policies, processes, and controls to govern the security of information assets and the infrastructure used to hold those information assets. Coupled with the security posture for a research data system are legislative compliance obligations, such as the General Data Protection Regulation, the Notifiable Data Breach legislation, and the Security of Critical Infrastructure Act.

The RDRA does not prescribe a particular framework but does recommend that an institution consider what policies, procedures, processes, and controls need to be adopted such that it is appropriate for the classifications of the data assets that it holds and to meet legislative compliance. Further, consideration needs to be given on how the organization's security practices and capabilities mature.

This article serves as a companion to those in which we proposed the rationale for the RDRA and an abstract implementation architecture.[19] Here we have specified a CMM that allows organizations to tune implementations based on budget, appetite for specific risk, and technology choices. The article is intended to guide senior research/academic committees, CIOs, CDOs, and those responsible for implementing operational infrastructure in research organizations.

ACKNOWLEDGMENT

Multiple individuals have contributed to the ideas in this article, specifically, Luc Betbeder-Matibet (UNSW), Stephen Bird (QCIF), Rhys Francis (UoM), Wojtek Goscinski (UQ), Ai-Lin Soo (UNSW), Carmel Walsh, Glenn Wightwick, and J. Max Wilkinson. Further, the feature set was tested in a public forum by the Australasian e-Research Organizations Association (https://www.aero.edu.au), held at Supercomputing Asia 2024 in Sydney, February 2024 (https://sca24.sc-asia.org). We thank the attendees of that forum for their insight and critical feedback.

REFERENCES

1. G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Rev. Phys.*, vol. 3, no. 6, pp. 422–440, 2021, doi: 10.1038/s42254-021-00314-5.
2. "The powers and perils of using digital data to understand human behaviour," *Nature (London)*, vol. 595, no. 7866, pp. 149–150, 2021, doi: 10.1038/d41586-021-01736-y.
3. S. Y. Hua, "Procurement maturity and IT failures in the public sector," *Transforming Government*, vol. 16, no. 4, pp. 554–566, 2022, doi: 10.1108/TG-07-2022-0097.
4. L. Armstrong, B. Flahertyand, and N. Kearns, "Waipapa Taumata Rau research data management capability maturity model v1.2 research data management project," The Univ. of Auckland, Auckland, New Zealand, Nov. 12, 2021. [Online]. Available: https://doi.org/10.17608/k6.auckland.16993660.v1
5. D. Abramson et al., "Why we need a reference architecture for research data," in *Proc. IEEE 19th Int. Conf. e-Sci.*, Limassol, Cyprus, 2023, pp. 1–4, doi: 10.1109/e-Science58273.2023.10254825.
6. R. Layne, A. Capel, N. Cook, and M. Wheatley, "Long term preservation of scientific data: Lessons from jet and other domains," *Fusion Eng. Des.*, vol. 87, no. 12, pp. 2209–2212, 2012, doi: 10.1016/j.fusengdes.2012.07.004.
7. Z. Liu et al., "Improving NASA's Earth satellite and model data discoverability for interdisciplinary research, applications, and education," *Data Sci. J.*, vol. 22, pp. 9–16, Apr. 2023, doi: 10.5334/dsj-2023-009.
8. F. Spreckelsen, B. Rüchardt, J. Lebert, S. Luther, U. Parlitz, and A. Schlemmer, "Guidelines for a standardized filesystem layout for scientific data," *Data (Basel)*, vol. 5, no. 2, 2020, Art. no. 43, doi: 10.3390/data5020043.
9. S. R. Carroll et al., "The CARE principles for indigenous data governance," *Data Sci. J.*, vol. 19, no. 1, 2020, Art. no. 43, doi: 10.5334/dsj-2020-043.
10. K. Julian, and R. P. Luciana, "Potential of I/O aware workflows in climate and weather," *Supercomput. Front. Innov.*, vol. 7, no. 2, pp. 35–53, 2020, doi: 10.14529/jsfi200203.
11. F. B. Schneider, A. Herbert, and K. S. Jones, "Least privilege and more," in *Computer Systems*, A. Herbert and K. S. Jones, Eds., New York, NY, USA: Springer-Verlag, 2004, pp. 253–258.
12. Khvastova, M. Witt, A. Essenwanger, J. Sass, S. Thun, and D. Krefting, "Towards interoperability in clinical research - Enabling FHIR on the open-source research platform XNAT," *J Med. Syst.*, vol. 44, no. 8, pp. 137–137, 2020, doi: 10.1007/s10916-020-01600-y.
13. M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber, "Capability maturity model, version 1.1," *IEEE Softw.*, vol. 10, no. 4, pp. 18–27, Jul. 1993, doi: 10.1109/52.219617.
14. J. C. Bejar et al., "Disaster recovery and data centre operational continuity," *J. Phys.: Conf. Ser.*, vol. 513, no. 6, 2014, Art. no. 062052, doi: 10.1088/1742-6596/513/6/062052.
15. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The extensible neuroimaging archive toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data," *Neuroinformatics*, vol. 5, no. 1, pp. 11–33, 2007, doi: 10.1385/NI:5:1:11.
16. Allan et al., "OMERO: Flexible, model-driven data management for experimental biology," *Nature Methods*, vol. 9, no. 3, pp. 245–253, 2012, doi: 10.1038/nmeth.1896.
17. A. Rajasekar et al., *iRODS Primer: Integrated Rule-Oriented Data System*, Cham, Switzerland: Springer-Verlag, 2010.
18. L. Freeman, "What's old is new again - Storage tiering," SNIA, Santa Clara, CA, USA, 1998. Accessed: Feb. 20, 2024. [Online]. Available: https://www.snia.org/sites/default/education/tutorials/2012/spring/storman/LarryFreeman_What_Old_Is_New_Again.pdf
19. J. Carroll, D. Abramson, and B. R. de Supinski, "An analysis of research data storage systems," in *Proc. IEEE 20th Int. Conf. e-Sci.*, Osaka, Japan, 2024, pp. 1–9, doi: 10.1109/e-Science62913.2024.10678706.

The Role of Knowledge Graphs on Responsible Artificial Intelligence Realization: Research Opportunities and Challenges

Xiang Li, University of Tasmania
Qing Liu, CSIRO Data61
Quan Bai, University of Tasmania
Xiwei Xu, CSIRO Data61, University of New South Wales

The improved performance of artificial intelligence (AI) systems has expanded the scope of AI-powered applications and makes AI play an increasingly important role in human life. Thus, building responsible AI that can be trusted by humans has drawn much research attention in recent years.

Driven by the development of artificial intelligence (AI) algorithms and the availability of big data, the performance of AI systems has been improved significantly in recent years. AI has played an increasingly important role in our daily lives. However, research has indicated that decisions made by some AI systems are discriminating or unreliable,[1] which might be harmful to humans. Building responsible AI (RAI) which can be trusted by humans has drawn much attention from government, industry, and academia recently.

With respect to trust concerns of AI systems, there are similar terms used in literature, such as ethical AI and trustworthy AI. Ethical AI refers to AI systems that align with ethical principles, and trustworthy AI focuses mainly on designing ethical algorithms for AI systems.[1] RAI, on the other hand, has both enforceable laws and ethical

Digital Object Identifier 10.1109/MC.2025.3545036
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

principles requirements.[2] Combining these two perspectives, in this work, we define RAI as legal compliance (enforceable) and ethical alignment (unenforceable) AI systems as shown in Figure 1.

To make the development of AI systems more sustainable and enhance human trust, governments, organizations, and private companies have proposed many regulations and ethical principles.[3] However, to build RAI and enhance human trust in practice, these proposals alone are not enough. It is essential to realize RAI through downstream tasks. For enforceable laws, such as acts and regulations, checking compliance can verify whether the AI systems align with lawful requirements.[4] For nonenforceable ethical principles, operationalizing high-level guidelines in practice can enhance human trust.[2] Knowledge graphs (KGs), as the structured data mode, represent information as directed multirelational graphs, where nodes denote entities and edges represent the semantic relations between them.[5,6] In terms of the ethical principle realization, the most commonly used advantage of KG in RAI is the explainability,[7] realizing the transparency. For compliance checking, the application of KG for RAI is representing regulation requirements to facilitate continuous compliance checking and traceability of the system.[8]

BUILDING RESPONSIBLE AI WHICH CAN BE TRUSTED BY HUMANS HAS DRAWN MUCH ATTENTION FROM GOVERNMENT, INDUSTRY, AND ACADEMIA RECENTLY.

In this work, we focus on how KGs can potentially support building RAI systems. Our contributions include the following:

1. We reveal the capabilities of KGs for both ethical principles and lawful requirements as shown in Figure 1, which brings opportunities for building RAI.
2. We analyze research challenges regarding KGs for RAI, focusing on KG construction and processing, which encompass the challenges of using KGs to enhance large language models (LLMs) for compliance checking. Meanwhile, the limitations of current works that utilize KGs to enhance LLMs in generating reliable outputs have also been analyzed.
3. We analyze a concrete use case which leverages KGs to enhance

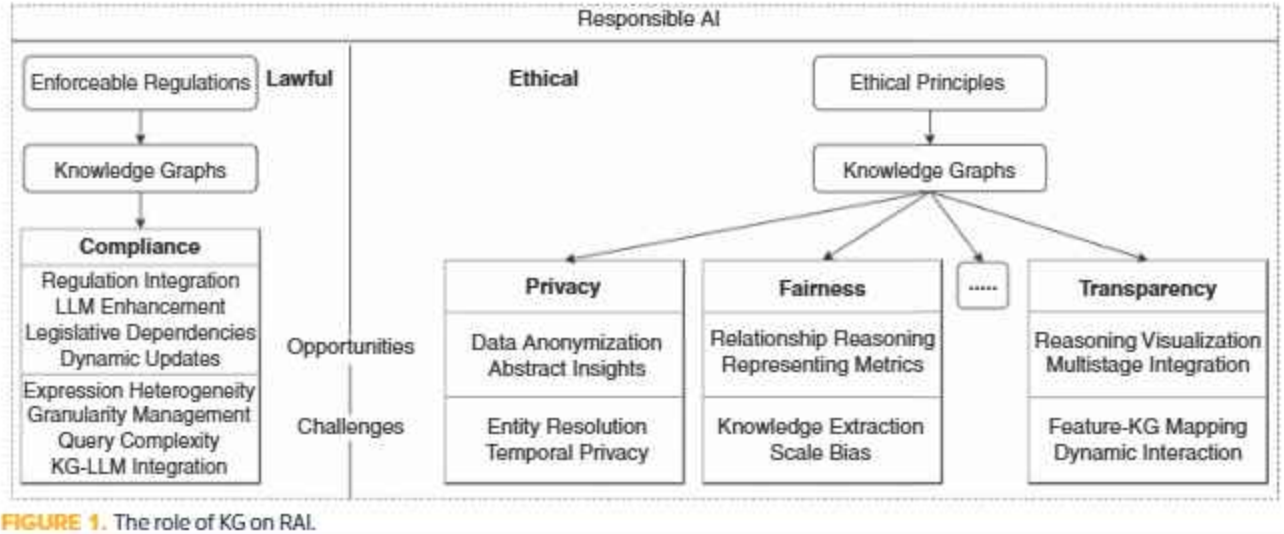


FIGURE 1. The role of KG on RAI.

compliance checking. The implementation demonstrates the potentials and roles of KGs in system improvement.

The rest of this article is organized as follows: the "Related Work" section analyzes and summarizes the related work in RAI and KG literacy, including the evaluation and exemplar applications. The "Research Opportunities" section discusses the opportunities brought by combining representation and KG processing for RAI. The "Research Challenges" section identifies research challenges for KG construction and processing for RAI. The final section provides the article's concluding statements.

RELATED WORK

RAI

Recent literature has established a comprehensive framework for RAI that encompasses both mandatory regulatory compliance and voluntary ethical adherence.

On the regulatory front, significant developments include the introduction of the Artificial Intelligence Act (AI Act)[a] and the General Data Protection Regulation (GDPR).[b] The ethical dimension is guided by various frameworks including Australia's AI Ethics Principles[c] and Ethics Guidelines for Trustworthy AI[d] to name a few.

This work emphasizes privacy, fairness, and transparency among AI ethical principles, rather than focusing on performance-related aspects, such as robustness and accuracy, or broader concepts like human-centered value and well-being. Privacy protection stands as a cornerstone requirement as AI systems must protect both individual privacy and business confidentiality through robust security measures and privacy-preserving techniques throughout the entire data lifecycle. The goal is maintaining a delicate balance between advancing AI capabilities while safeguarding sensitive information through clear protocols and user control over personal data.[1,2]

The principle of fairness in AI systems emphasizes the ethical principle that AI systems should treat all individuals and groups equitably, avoiding discriminatory outcomes based on protected characteristics.[2,9] Transparency represents another critical dimension of RAI. AI systems should clearly disclose their artificial nature and provide explanations for their behaviors and decisions to ensure users understand when and how they are interacting with AI. Transparent AI systems that can generate explanations about their functioning help educate users and enable them to better utilize and explore the technology's capabilities.[2,7]

KGs

Before the AI Act, the GDPR was the widely adopted regulation since most AI systems these days are data-driven AI. They learn patterns from large amounts of input data. GDPR, as data protection law, regulates activities related to personal data collection, processing, and storage. Compliance with GDPR can realize lawful requirements of RAI.

In terms of regulation representation, many works have been proposed to represent requirements and rules in GDPR. Description Logics has been used to chain multiple expressions together to express the rules in GDPR. However, the predefined rules restrict the expressiveness of knowledge representation in the original legal document. Graph-structured ontology has also been designed to represent GDPR. Related work includes GDPRov[e] and GDPRtEXT.[f] GDPRoV captures requirements on personal data processing, storing, and transferring related activities in GDPR. GDPRtEXT, on the other hand, represents the GDPR structure of chapters, sections, articles, points, and key concepts contained in these parts. Another work extracts permissions and obligations in GDPR and forms a KG to represent these requirements. Under the same schema, the GDPR KG is further extended to incorporate other relevant regulations to enrich the lawful requirement expression.[4] This work demonstrates the capability of KGs in representing information from heterogeneous resources and linking multiple regulations.

For ethical AI principles, researchers have demonstrated various KG implementations. Privacy preservation was achieved through data structuring methods that maintain anonymization while representing relations.[5] Fairness applications leveraged graph embeddings with social information and frameworks like the Fairness Metrics Ontology.[9,10] Transparency efforts utilized semantic modeling for feature extraction and relationship mapping in AI pipelines.[7]

With the rapid development of LLMs, they have shown remarkable

[a] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[b] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[c] https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles
[d] https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
[e] https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology
[f] https://openscience.adaptcentre.ie/ontologies/GDPRtEXT/deliverables/docs/ontology

potential in legal applications, demonstrating strong capabilities in understanding and executing legal instructions.[11] However, their application in legal contexts faces several challenges, including limited domain-specific knowledge, the risk of hallucinations, and the need for accurate, up-to-date information processing.[11,12] The hallucination issue of LLMs refers to their tendency to produce artificial responses that deviate from truth or reality by fabricating facts and details not supported by available information or training data.[13] KGs have been proposed as external knowledge bases to reduce hallucinations and improve reliability.[8] Some research has explored combining KGs with LLMs for legal tasks,[13] demonstrating the potential for enhanced accuracy in compliance verification.

Evaluation

KG construction. In general, the KG construction for RAI requires domain knowledge, and the selection and inclusion of RAI-related information have not been standardized. General KG construction evaluations focus on entity alignment, factual accuracy, etc.[6] However, for KGs aimed at providing information for RAI, the quality of construction depends on their usefulness rather than just correctness. It is challenging to set the ground truth and evaluate usefulness instead of correctness for KG construction for RAI.

KG processing. For compliance checking, automated evaluations are constrained to closed-end questions. Simple answers of compliance or noncompliance are not informative to end users without thorough legal analysis. Comprehensive legal analysis, however, heavily relies on domain experts to manually evaluate its quality. Conclusions may not be unified, as different legal experts can reach varying results, making unified evaluation difficult. For realizing RAI values, evaluation faces two main challenges: 1) quantifying ethical values like the system's transparency or robustness; 2) managing conflicting RAI values, for example, transparency and privacy, which complicates assessing trade-offs between them.

Exemplar applications on KG processing for RAI

Combining KG representation capabilities with the natural language understanding and reasoning capabilities of LLMs brings opportunities for enhancing current work on RAI value realization through compliance checking and operationalization. Figure 2 showcases exemplar applications of KGs in RAI realization.

Monitoring and traceability. A KG guided by provenance ontology captures provenance information for legal requirements and AI system information. For example, the use case proposed in Pandit et al.,[14] describes the consent requirements outlined in GDPR. In the provenance KG, the grant of consent, the object of the consent, and the time frame of the consent are defined at the class level. Concrete activities of the system processing personal data, such as collecting users' addresses, are recorded as instances. Using the SPARQL query language, the provenance information about consent can be returned as entities and relations in the KG, indicating that consent is given by the user for

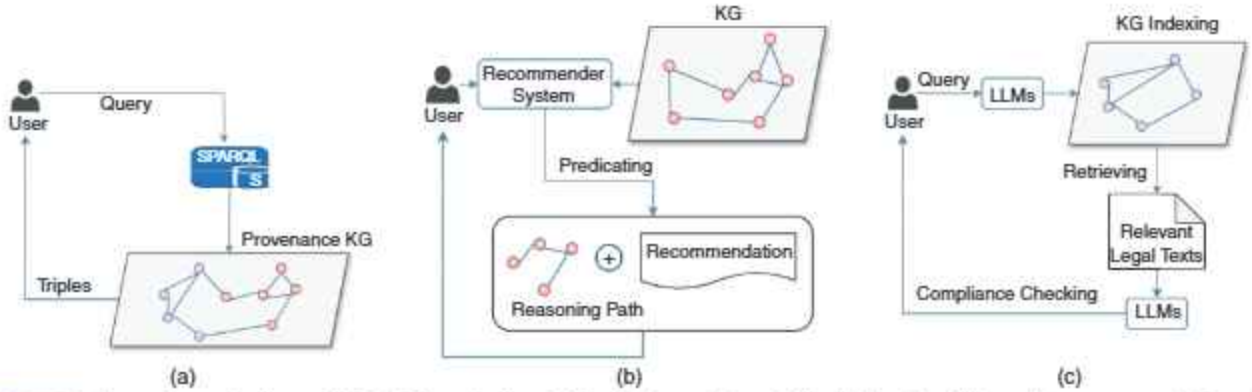


FIGURE 2. Exemplar applications of KGs in RAI realization. (a) Monitoring and traceability. (b) Explainability and transparency. (c) Compliance checking.

collecting address information from the recorded start date to the end date. As shown in Figure 1(a), with this provenance information, users can monitor the system's processing and take necessary actions to mitigate deviations. Involving LLMs further simplifies the process. Users can directly ask questions in natural language, and LLMs will convert them to SPARQL queries. With the provenance information recorded on AI systems and legal requirements, the system can be continuously monitored as the information is queried and updated.

Explainability. The recommender system showcases the benefits of KGs for realizing the transparency RAI value. As shown in Figure 1(b), the KG captures attributes of items and relationships between items. By embedding the KG, the recommender system can extract the reasoning path that leads to the final recommended items.[7] The explanation, represented as the path in the KG, provides transparency on how the AI models make predictions and enhances user trust in the system.

Compliance checking. With LLMs' incredible capabilities in natural language understanding and reasoning, legal texts, such as regulations or acts, serve as trustworthy external knowledge to enhance LLM-based compliance checking.[8] Using legal texts, a KG is generated as an index. Based on the processing activities' facts, LLMs retrieve applicable legal context information using this KG indexing. Subsequently, LLMs conduct legal analysis based on the retrieved facts and legal norms. The compliance checking results are then directly presented in natural language for the end user, as illustrated in Figure 1(c).

KG-enhanced legal compliance: insights from ChatLaw implementation

A research team has developed ChatLaw,[8] a system that exemplifies how the integration of KGs can improve legal compliance verification and advisory services. Their implementation demonstrates that incorporating structured knowledge representation can substantially enhance the precision and dependability of automated legal services while minimizing the tendency of language models to generate unfounded information.

KG integration in system architecture. The KG structure in ChatLaw specifically emphasizes the organization of legal consultation knowledge. The system builds its knowledge representation through a carefully designed workflow that transforms multisource legal data into structured KGs. The KG implementation serves a crucial role in consultation processes. Under expert guidance, the system abstracts legal consultation procedures into a professional KG framework. This framework begins with determining consultation types and their corresponding entity sets, which form the foundation of the knowledge structure.

The practical application of the KG manifests in the consultation workflow. When users initiate a consultation, the legal assistant component selects relevant predefined entity clusters from the KG and systematically populates information nodes. For incomplete information, the system integrates these gaps into new inquiries, guiding users to provide additional relevant details that expand the KG. The legal assistant selects the predefined entity cluster according to user queries and systematically explores key nodes. Once these nodes contain complete information, the system transfers this comprehensive knowledge to subsequent processing stages.

Implications for KG-enhanced compliance. The integration of KGs in ChatLaw demonstrates measurable improvements in legal consultation accuracy and reliability. The system's KG-driven approach contributes significantly to its performance in standardized evaluations, where it surpassed existing models in legal reasoning tasks. The implementation of KGs particularly strengthens the system's ability to gather comprehensive information during consultations. Through its structured clusters of entities and node relationships, the system ensures systematic coverage of all relevant legal factors. This systematic approach helps prevent information gaps that could lead to incomplete or inaccurate legal analysis.

The KG architecture in ChatLaw proves especially valuable in maintaining consultation accuracy by guiding information collection through predefined entity relationships. When encountering unclear or incomplete information, the system uses its KG structure to identify specific information needs and generate targeted follow-up questions. This structured approach helps ensure that all necessary legal factors are considered before generating advice.

Beyond immediate performance metrics, KG implementation establishes a framework for more reliable legal compliance systems. By abstracting legal consultation processes into structured knowledge representation, the system creates a reproducible methodology for handling complex legal queries. This

abstraction enables systematic information gathering while maintaining the logical relationships between different legal concepts and requirements.

In conclusion, the system's ability to maintain consistency in legal consultations through KG guidance addresses a fundamental challenge in legal technology: the need for systematic and thorough information gathering. The predefined entity clusters and relationship patterns in the KG ensure that consultations follow established legal frameworks while adapting to specific case requirements. This structured approach helps maintain the quality and reliability of legal consultations in diverse cases and scenarios.

RESEARCH OPPORTUNITIES

Figure 1 summarizes the contributions of KGs with respect to ethical principles and regulation compliance. With natural language processing-based information extraction, KGs are able to represent requirements enforced in regulations. In addition, much research has been done on developing ethical AI algorithms, and these algorithms embed RAI values by design.[1]

Regulation compliance

While previous work focused primarily on representing GDPR requirements through KGs, the integration with emerging LLMs opens new opportunities.[11] Unlike traditional rule-based representations that were limited in expressiveness, KGs can now serve as reliable external knowledge to ground LLMs and address their challenges of limited domain knowledge, hallucination risks, and need for accurate information processing.[11,12] This is particularly critical in compliance checking, where accuracy and reliability are paramount. To address these challenges, KGs serve as a reliable external knowledge source to ground and enhance LLMs' performance in legal tasks. By providing structured, machine-readable representations of legal concepts and relationships, KGs can help address the issues of factual inconsistencies and missing crucial information in LLMs' outputs.[13] This grounding mechanism is particularly important for ensuring the safety and legality of LLM-generated outputs.

The value of this integration extends to the modeling of complex legislative interconnections. By capturing the intricate relationships between laws, articles, and their broader legislative context, KGs create a comprehensive framework for understanding legal dependencies.[12] This structured representation of legal relationships enhances the ability to perform thorough compliance checking by considering the full context of relevant laws and regulations. Furthermore, the structured nature of KGs contributes to enhanced domain adaptation of LLMs in legal applications. When integrated into the pretraining process, legal KGs help address the challenge of limited domain-specific knowledge in general-purpose LLMs.[11] Their lower noise levels and higher knowledge density compared to raw legal texts make them particularly effective for specializing LLMs for legal tasks.

One of the most significant advantages lies in their flexibility for handling the dynamic nature of legal knowledge. Unlike LLMs that require complete retraining to incorporate new laws, KGs can be continuously updated to reflect the latest legislative changes.[12,13] This adaptability is crucial in the legal domain, where new laws and regulations are published daily, ensuring that compliance checking systems remain current without the need for frequent model retraining.

Ethical principles

KGs offer transformative potential for enhancing privacy protection in AI systems through advanced capabilities beyond existing data structuring approaches.[5] Their ability to represent complex relationships while maintaining data anonymization opens new possibilities for privacy-preserving AI development. KGs can support sophisticated data abstractions and aggregations that enable valuable insights without compromising sensitive information.[5] This advancement is especially significant for domains requiring strict privacy measures while maintaining data utility, such as health care and financial services.

KGs present innovative opportunities for fairness in AI systems by enabling nuanced representation of societal and technical relationships beyond current frameworks. Their potential extends to developing comprehensive fairness evaluation systems that can handle complex intersectional considerations.[9,10] KGs' reasoning capabilities offer promising directions for automated bias detection and mitigation across diverse applications. The technology's ability to integrate multifaceted social contexts with technical parameters suggests potential applications in areas like employment, education, and social services, extending beyond health care.[10] These capabilities could revolutionize how organizations approach fairness in automated decision-making systems while ensuring scalability and consistency.

KGs also demonstrate significant potential in advancing AI transparency

through their unique semantic structuring capabilities.[7] Their ability to capture and represent complex reasoning chains suggests possibilities for developing more sophisticated explainability mechanisms across various AI applications. The semantic modeling potential of KGs indicates promising directions for creating more intuitive and user-friendly AI systems that can explain their decisions effectively to diverse stakeholders, from technical experts to end users.[7] This capability could transform how AI systems communicate their decision-making processes across sectors like autonomous systems, financial services, and public administration.

RESEARCH CHALLENGES

In general, most challenges of constructing KGs to represent unstructured data for RAI come from the heterogeneity when dealing with domain knowledge. Existing work performs this task by extracting domain-specific entities (a.k.a. named entity recognition) and linking these entities through relations (a.k.a. relation extraction) based on predefined ontology.[15] However, RAI requires continuous monitoring throughout the lifecycle of AI system design, development, and deployment.[16] Information and knowledge are dynamic in the time series and heterogeneous in format. Updating rapidly developed new knowledge on RAI and maintaining the quality of the KG[17] put challenges on this predefined framework.

Regulation compliance

Natural language sentences in legal texts can express similar relationships and meanings with different expressions. Current work constructs a GDPR KG through ontology-based information extraction.[15] It tends to align these heterogeneous expressions with predefined relations, which compromises the expressiveness. By contrast, open information extraction (Open IE)[18] is able to extract entity and relation information directly from natural text, which enriches the expressiveness of the output KG. However, information extracted from Open IE is heterogeneous.

The granularity of expressions on regulations is at a coarse-grained level as abstract concepts, while particular AI systems utilize fine-grained level expressions as instances. The mappings between the two granularity levels can be challenging. It is necessary to figure out which level of granularity can represent requirements in legal documents and the corresponding information in AI systems. In regulations, requirements are written in a human-readable way, which means one article or even one sentence actually contains multiple requirements. It is easy for humans to confirm the compliance with a certain article by satisfying all the requirements listed in that article. However, for KG processing, checking each requirement and then aggregating them to claim compliance is challenging. For KG queries, proficiency in SPARQL query language skills is required to retrieve information from the KG. It becomes even more complex to write SPARQL query syntax when querying across multiple graphs.

The LLMs provide alternative approaches for compliance checking, as identified in the "Research Opportunities" section. Despite the promising potential of KGs in legal applications, several significant challenges exist in their effective integration with LLMs. A fundamental challenge lies in the complexity of knowledge integration across different formats. As highlighted in Pan et al.,[19] the process of combining unstructured information like text passages with structured data from KGs presents substantial technical hurdles. The traditional similarity-based retrieval methods often fall short in capturing the intricate logical relationships needed for sophisticated legal reasoning, potentially leading to suboptimal performance when LLMs attempt to process and reason with the retrieved information.

Another critical challenge emerges in the construction and maintenance of domain-specific KGs for legal applications. The work by Colombo[11] emphasizes that the inherent unstructured nature of legislative texts makes it particularly difficult to develop accurate and reliable KGs. This challenge is compounded by the need to ensure that the resulting KG captures not just the surface-level information, but also the complex web of relationships and dependencies that characterize legislative frameworks. The process requires sophisticated approaches to transform unstructured legal texts into well-structured, machine-readable formats while preserving the nuanced legal context and interconnections.

Furthermore, the integration challenge extends beyond mere technical implementation. While KGs can theoretically enhance LLMs' reasoning capabilities, achieving effective reasoning across different knowledge representations remains problematic. As noted by Pan et al.,[19] LLMs often struggle to properly account for the complex relationships between different statements, even when provided with explicit knowledge. This limitation suggests that simply having

access to a KG is insufficient; the challenge lies in developing mechanisms that enable LLMs to effectively utilize the structured relationships and dependencies represented in the KG for sophisticated legal reasoning tasks.

These challenges collectively point to the need for innovative approaches that can bridge the gap between structured knowledge representation and LLM-based reasoning, while maintaining the accuracy and reliability essential for legal applications.

Ethical principles

Privacy challenges in KG implementations stem largely from the complexities of managing entity identities and their relationships. A fundamental obstacle lies in the difficulty of correctly identifying and normalizing entities that share similar characteristics or representations.[6] This becomes particularly concerning for privacy protection since any misidentification can create incorrect linkages between sensitive data points, potentially compromising individual privacy. Another significant challenge emerges from the temporal nature of knowledge management—while current frameworks can handle static snapshots and immediate updates, they struggle with effectively managing rapidly evolving information.[6] This limitation poses substantial risks to privacy preservation, as it complicates the implementation of consistent and reliable privacy controls across different temporal states of the KG.

The implementation of fairness in KGs encounters several technical hurdles, particularly in the realm of knowledge acquisition and structural development. Despite significant technological progress in language processing capabilities, the automated extraction of well-structured knowledge remains a substantial challenge.[6] This limitation has direct implications for fairness, as any gaps or biases in the knowledge extraction process can lead to systemic inequities in the resulting knowledge representations. The challenge becomes more pronounced when scaling KGs, as the necessary combination of different extraction methodologies—ranging from manual to automated approaches—can introduce varying degrees of bias into the system.[6] Each method may capture different aspects of reality with varying degrees of accuracy, potentially leading to uneven representation across different demographic groups or concept categories.

In terms of transparency, KGs face distinct challenges in achieving effective explainability despite their inherently interpretable nature. A key challenge lies in establishing meaningful connections between the technical components of AI systems and their KG representations.[20] This includes the complex task of mapping neural network features to ontological elements in a way that supports clear explanation. Additionally, the field needs to move beyond static explanatory approaches toward more dynamic, interactive systems that can engage with users meaningfully.[20] This challenge is further complicated by the difficulties in extracting and integrating knowledge from diverse data sources,[6] as the quality of explanations directly depends on how well the knowledge structure captures and represents the underlying domain complexity.

In this work, we investigate the roles of KGs on RAI from two perspectives: research opportunities and challenges. The "Research Opportunities" section identifies areas that KGs can facilitate, realizing ethical values to build RAI. To facilitate research on this specific topic, we identify research challenges and propose a concrete use case to illustrate the potential solutions on leveraging KGs for RAI.

REFERENCES

1. H. Liu et al., "Trustworthy AI: A computational perspective," 2021, *arXiv:2107.06641*.
2. Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a roadmap on software engineering for responsible AI," 2022, *arXiv:2203.08594*.
3. T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.
4. K. P. Joshi, L. Elluri, and A. Nagar, "An integrated knowledge graph to automate cloud data compliance," *IEEE Access*, vol. 8, pp. 148,541–148,555, 2020, doi: 10.1109/ACCESS.2020.3008964.
5. C. Karalka, G. Meditskos, M. Papoutsoglou, and N. Bassiliades, "Towards a generic knowledge graph construction framework for privacy awareness," in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 700–705, doi: 10.1109/CSR61664.2024.10679399.
6. N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done," *Queue*, vol. 17, no. 2, pp. 48–75, 2019, doi: 10.1145/3329781.3332266.
7. E. Rajabi and K. Etminani, "Knowledge-graph-based

ABOUT THE AUTHORS

XIANG LI is a Ph.D. candidate at University of Tasmania, Hobart 7005, Australia. Her research interests include responsible AI, knowledge graphs, and information extraction. Li received her master's degree in information and communication technology from the University of Tasmania. Contact her at x.li@utas.edu.au.

QING LIU is a senior research scientist of CSIRO Data61, Hobart 7005, Australia. Her research interests include knowledge graphs, responsible AI, and big data analysis. Qing received her Ph.D. in database indexing and approximate query processing from the University of New South Wales. Contact her at Q.Liu@data61.csiro.au.

QUAN BAI is an associate professor in the School of Information & Communication Technology, University of Tasmania, Hobart 7005, Australia. His research mainly focuses on machine learning, knowledge management, and agent-based modelling for complex systems. Bai received his M.Sc. from the University of Wollongong. Contact him at quan.bai@utas.edu.au.

XIWEI (SHERRY) XU is a principal research scientist at CSIRO Data61 and a conjoint senior lecturer at the University of New South Wales, Sydney, NSW 2015 Australia. Her research interests include software engineering for AI-based systems (SE4AI) and blockchain applications. Xu received her Ph.D. in RESTful business process from the University of New South Wales. She is a Senior Member of IEEE. Contact her at xiwei.xu@data61.csiro.au.

explainable AI: A systematic review," *J. Inf. Sci.*, vol. 50, no. 4, pp. 1019–1029, 2024, doi: 10.1177/01655515221112844.

8. J. Cui et al., "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2024, *arXiv:2306.16092*.
9. J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, and D. L. McGuinness, "An ontology for fairness metrics," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2022, pp. 265–275.
10. T. Shang et al., "Integrating social determinants of health into knowledge graphs: Evaluating prediction bias and fairness in healthcare," 2024, *arXiv:2412.00245*.
11. A. Colombo, "Leveraging knowledge graphs and LLMs to support and monitor legislative systems," in *Proc. 33rd ACM Int. Conf. Inf. Knowl. Manage.*, 2024, pp. 5443–5446.
12. J. Shi, Q. Guo, Y. Liao, Y. Wang, S. Chen, and S. Liang, "Legal-LM: Knowledge graph enhanced large language models for law consulting," in *Proc. Int. Conf. Intell. Comput.*, Singapore: Springer-Verlag, 2024, pp. 175–186.
13. G. Hannah, R. T. Sousa, I. Dasoulas, and C. d'Amato, "A prompt engineering approach and a knowledge graph based framework for tackling legal implications of large language model answers," 2024, *arXiv:2410.15064*.
14. H. J. Pandit, D. O'Sullivan, and D. Lewis, "Queryable provenance metadata for GDPR compliance," *Procedia Comput. Sci.*, vol. 137, pp. 262–268, Jan. 2018, doi: 10.1016/j.procs.2018.09.026.
15. D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *J. Inf. Sci.*, vol. 36, no. 3, pp. 306–323, 2010, doi: 10.1177/0165551509360123.
16. Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A multivocal literature review," 2022, *arXiv:2209.04963*.
17. E. Filtz, "Building and processing a knowledge-graph for legal data," in *Proc. Eur. Semantic Web Conf.*, Cham, Switzerland: Springer-Verlag, 2017, pp. 184–194.
18. J. L. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo, "Information extraction meets the semantic web: A survey," *Semantic Web*, vol. 11, no. 2, pp. 255–335, 2020, doi: 10.3233/SW-180333.
19. J. Z. Pan et al., "Large language models and knowledge graphs: Opportunities and challenges," 2023, *arXiv:2308.06374*.
20. G. Futia and A. Vetrò, "On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three challenges for future research," *Information*, vol. 11, no. 2, 2020, Art.no. 122, doi: 10.3390/info11020122.

Rethinking Software Testing for Modern Development

Anurag Saxena, Red Hat, Inc.

This article explores the shift from manual to automated testing, emphasizing the role of artificial intelligence and machine learning in enhancing efficiency and quality assurance in the software development lifecycle, highlighting innovative security testing for open source software and examining artificial intelligence's impact on testing frameworks.

Software testing is a critical component of the software development lifecycle. It ensures that a product meets the required standards and functions as intended. Over the decades, software testing has evolved significantly, moving from manual testing to more sophisticated automated testing methods. This study examines the transformation of software testing from manual to automated techniques, focusing on the shift from rigid frameworks to adaptable, iterative methods aligned with agile practices. It analyzes trends and innovations to highlight effective strategies for addressing modern software quality assurance challenges, including reducing time to market, lowering costs, and applying critical testing techniques from other sectors.

This article aims to evaluate recent advancements in testing metrics and methodologies for open source software (OSS), focusing on object-oriented versus nonobject-oriented software and optimized testing techniques for artificial intelligence (AI) and machine learning, (ML) including proof of concept for AI applications.

The objectives of this study are threefold: To provide a historical overview of software testing methodologies,

Digital Object Identifier 10.1109/MC.2025.3554094
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

to analyze current practices and their effectiveness, and to project future trends and recommendations aiming to optimize testing strategies.

SIGNIFICANT ADVANCES IN TESTING

In the past decade, the most significant advance in software testing has been the widespread adoption of automated testing powered by AI and ML, which has significantly increased efficiency, test coverage, and the ability to rapidly identify issues throughout the development process; this shift is often associated with the integration of testing into DevOps pipelines, enabling continuous testing and faster release cycles. Figure 1 shows software testing evolution in last few decades. Key points about these advancements are the following:

- *Shift-left testing:* Conduct tests earlier in development to quickly identify and resolve issues, preventing larger challenges.
- *Continuous testing:* Use automated testing in the development pipeline for regular quality feedback.
- *Automation tools like Selenium:* A popular framework for automating web application testing, improving speed and thoroughness.
- *Mobile automation testing:* Automate testing of mobile apps to ensure functionality across devices and operating systems.
- *Security testing:* Focus on security testing to identify and mitigate software vulnerabilities.
- *QA Ops:* Integrate quality assurance with DevOps to improve testing efficiency.
- *Performance engineering:* Enhance application performance throughout the development lifecycle, beyond standard testing.
- *AI and ML in testing:* Use AI to create test cases, evaluate outcomes, and identify potential issues.
- *Scriptless test automation:* Allow nontechnical users to create automated tests without coding, using frameworks like Cucumber.

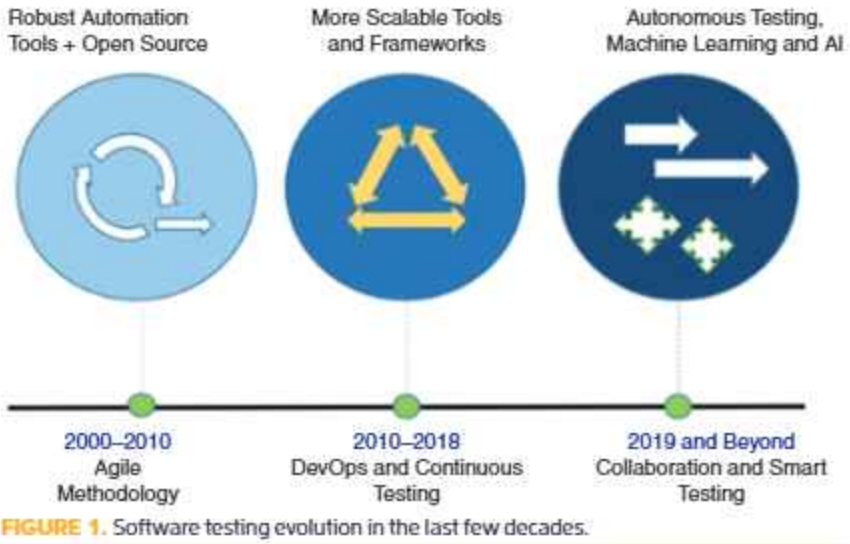


FIGURE 1. Software testing evolution in the last few decades.

The current "state of the art" in automated testing primarily involves leveraging AI and ML to create more intelligent, self-healing test suites, including features like automated test case generation, intelligent test data creation, flakiness detection, visual testing, and seamless integration with continuous integration/continuous delivery (CI/CD) pipelines, allowing for faster feedback loops and improved test coverage across various platforms and devices. A few key aspects of state-of-art testing automated testing are as follows:

- Using AI algorithms to autonomously generate test cases from requirements and application behavior, reducing manual development time.
- Creating realistic and diverse test data with ML to cover various edge cases and scenarios.
- Identifying and fixing "flaky" tests with ML to improve test reliability.
- Implementing systems for automated tests to self-correct minor user interface (UI) changes, reducing maintenance efforts.
- Integrating automated tests into the CI/CD pipeline for quick feedback on code changes.

The transition to automated testing, aided by intuitive interfaces and AI, has significantly reduced testing costs and increased efficiency, prompting organizations to allocate more budget to quality assurance. This trend

underscores the importance of early-stage testing to further cut costs and improve software quality.

MINIMIZING TIME TO MARKET WITHOUT COMPROMISES

In a rapidly evolving technological landscape, the demand for high-quality software products has never been greater. Organizations are under increasing pressure to deliver software solutions that not only meet consumer expectations but do so in a timely manner.

To minimize time to market without compromising testing phases, prioritize automation in testing processes, implement CI/CD, focus on developing a minimum viable product with core functionalities, utilize parallel testing strategies, and prioritize critical test cases while optimizing test environments to identify issues early and quickly in the development cycle. Key strategies to achieve this balance are as follows:

- Automated testing tools should be used for regression, UI, and performance tests, enabling manual testers to focus on more complex scenarios.
- Identify key functionalities and prioritize testing in those areas for faster feedback cycles.
- Run multiple tests simultaneously in different environments to speed up the process.
- Integrate automated tests into the development pipeline to catch bugs early and enable rapid feedback loops.
- Write tests before writing code to ensure functionality is thoroughly tested from the start.
- Focus on developing a product with essential features for early user feedback and quick iterations.
- Create testing environments that mimic real-world scenarios to identify potential issues early.
- Use tools that replicate real-world scenarios to speed up testing without full system integration.
- Create a comprehensive test strategy with clear goals, priorities, and timeframes.
- Ensure all stakeholders are aligned on testing priorities and expectations.

When pursuing key strategies, it's crucial to conduct risk assessments, promote skill development, identify bottlenecks, and maintain quality over speed. These practices can significantly reduce time to market while ensuring thorough testing and high-quality product delivery.

TESTING WITH FORMAL METHODS AND WHEN TO STOP TESTING

Software testing is essential in engineering for defect identification and reliability assurance. Traditional methods often struggle with modern software complexities, prompting the use of advanced techniques like formal methods. These mathematical frameworks improve specification, development, and verification, allowing for early issue detection and lower bug-fix costs. This is especially important in safety-critical sectors like avionics and medical devices, where small errors can have severe consequences. Integrating formal methods into testing enhances traditional approaches, promoting a proactive design strategy for high reliability in critical systems. Key aspects of formal methods in software testing are as follows:

- Formal methods can be applied during requirements gathering and system design, unlike most testing techniques used later, allowing for earlier identification of potential issues.
- Formal methods use mathematical notation to create clear representations of system behavior, enabling better analysis and verification.
- Model checking systematically examines all possible states of a system to identify potential errors or inconsistencies.
- Theorem proving uses logical reasoning to mathematically validate system properties.

Formal methods enhance conventional software testing by systematically identifying issues that standard techniques may miss. By using mathematical models and precise specifications, they uncover hidden defects, improve software reliability, and help create targeted test cases. This proactive approach allows developers to focus on critical aspects, resulting in more robust software.

Finalizing software testing relies on predefined exit criteria that assess deployment quality. Effective communication among team members and stakeholders is essential for aligning expectations. Continuous evaluation of software performance and user feedback is vital for maintaining quality. By establishing clear exit criteria and promoting transparency, organizations can make informed decisions that enhance software quality and stakeholder satisfaction, leading to successful project delivery.

SECURITY VULNERABILITIES

As the world becomes more interconnected, rising security vulnerabilities pose significant risks to organizations and individuals. Cyberattacks are increasingly sophisticated and frequent, leading to financial losses, reputational damage, and legal issues. It is crucial for software developers, cybersecurity professionals, and policymakers to understand and address these vulnerabilities. Traditional software security methods, which focus mainly on basic functionality testing, may no longer suffice in today's complex landscape.

Some ways to reduce security vulnerabilities are as follows:

- Identify system, application, and network weaknesses vulnerable to exploitation.
- Track vulnerability resolution progress and highlight areas for improvement.
- Detect potential threats early and effectively address vulnerabilities.
- Evaluate vulnerabilities and deficiencies in systems, networks, and procedures.
- Use a secure coding checklist to prevent vulnerabilities; implement Bug Bounties to find significant software flaws.
- Apply the principle of least privilege to access rights for devices and systems.
- Create protocols to prevent accidental disclosure of sensitive information in error messages.
- Identify, categorize, and address system vulnerabilities.

Other ways to reduce the risk of vulnerabilities include keeping software and systems up to date, implementing strong access controls, providing regular cybersecurity training, implementing effective incident response procedures, and using encryption and other security technologies.

LATEST DEVELOPMENTS IN TESTING METRICS

The integration of AI and ML in testing metrics is transforming quality assurance by automating test case generation and enhancing pattern recognition in test data. This leads to more efficient problem identification and streamlines testing across software applications. AI-driven testing improves quality assurance by identifying key test scenarios, creating test cases, and optimizing test suites, reducing the need for manual intervention. Additionally, predictive analytics uses historical data to anticipate potential issues, enabling proactive quality improvements. Advanced test case design employs AI algorithms to create cases that capture complex user behaviors and edge cases, enhancing test coverage and overall quality.

In the contemporary landscape, the ongoing advancement and proliferation of AI have rendered ethical considerations paramount in discussions regarding its implementation. A primary concern in this context is the safeguarding of privacy.[1,2,3] As AI technologies become increasingly integrated into various societal functions, they process vast quantities of data, including sensitive and personal information, thereby heightening the risk of privacy violations.[4,5] Consequently, it is imperative to employ AI algorithms that not only honor privacy but also actively protect it.

Emerging testing metrics are reshaping software development. Shift-left testing emphasizes early integration of testing, leading to faster feedback and defect detection. Incorporating testing into the DevOps pipeline enables continuous feedback and accelerates delivery. Codeless automation testing makes automated tests accessible to nontechnical users through graphical interfaces. The rise of Big Data and the Internet of Things increases the need for metrics tailored to extensive datasets and interconnected devices.

When adopting new testing metrics, it's essential to align them with business objectives, ensuring they effectively measure quality and user value. Maintaining data quality and accuracy is crucial, as metrics must be based on reliable data reflecting real-world scenarios. Continuous monitoring and analysis are necessary to keep testing metrics relevant to evolving project needs.

EXPLORING INNOVATIVE SECURITY TESTING FOR OSS

As technology becomes integral to our lives, the significance of software development has surged, with OSS leading in transparency, collaboration, and community-driven innovation. However, this growth brings security challenges, highlighting the need for effective testing methodologies, particularly in security testing and autogenerated code detection. Modern OSS security testing utilizes automated fuzzing, AI-enhanced static analysis, behavioral analysis, and community-driven vulnerability scanning to identify flaws, outperforming traditional methods like manual code reviews by leveraging the collaborative nature of open source development to quickly address vulnerabilities.

To effectively assess the security risks of OSS, consider the following tips:

- Stay up to date with security advisories and bug tracking systems for the software you utilize.

- Regularly update and patch OSS to address any known vulnerabilities.
- Evaluate the reputation and track record of the project's community support.
- Review the codebase for quality, adherence to security best practices, and evidence of ongoing maintenance.
- Engage with the broader open source community to discuss and share security concerns and best practices.

Innovative open source security testing improves vulnerability detection through key components. Advanced static analysis methods, such as taint analysis, track data flow to identify injection vulnerabilities, like structured query language (SQL) injection and XSS. Semantic analysis uncovers complex security issues, while tailored rules target specific architectural weaknesses. Dynamic analysis uses fuzzing frameworks to test applications with random inputs, enhancing critical vulnerability detection through coverage-guided fuzzing. Interactive application security testing allows real-time vulnerability detection during development, and runtime analysis monitors application behavior. Community-driven tools like "Nmap," "Wapiti," and "Zed Attack Proxy" (ZAP) support vulnerability identification and foster collaboration among developers and security researchers.

Advanced security testing can be effectively performed using open source tools. Static analysis tools, like Coverity, SonarQube, and Fortify on Demand, identify code vulnerabilities before deployment. For dynamic testing, fuzzing frameworks, such as American Fuzzy Lop and Peach Fuzz expose weaknesses through random data injection. Web application security is assessed with scanners like OWASP ZAP and Burp Suite, while Nmap detects active devices and services. Sqlmap is essential for identifying SQL injection vulnerabilities. Additionally, tools like Snyk, Checkmarx, SpotBugs, and Checkstyle analyze code patterns to improve accuracy and reduce false positives in autogenerated code.

TRADITIONAL SOFTWARE SECURITY METHODS, WHICH FOCUS MAINLY ON BASIC FUNCTIONALITY TESTING, MAY NO LONGER SUFFICE IN TODAY'S COMPLEX LANDSCAPE.

INDUCTION OF TEST DATA REDUCTION, TEST TIME REDUCTION, AND MODEL-BASED TESTING

Induction in software testing encompasses methods designed to enhance the efficiency, effectiveness, and overall quality of tests, with a focus on three key areas: test data reduction, test time reduction, and model-based testing.

Test data reduction aims to minimize the volume of test data while ensuring adequate coverage of the system under test, which conserves computational resources and time. Inductive approaches for this include automatic generation, search-based methods, and constraint solving, utilizing data mining and ML techniques to eliminate redundant inputs.

Test time reduction focuses on decreasing the duration required to run a test suite, crucial for large software systems. Inductive strategies here involve prioritizing test cases based on previous results, selecting critical test cases for execution, and employing parallel testing across multiple environments.

Model-based testing leverages models to automatically generate test cases that reflect the expected behavior of the system. Inductive methodologies in model-based testing include automated model learning, model refinement, and test generation from these models, enhancing precision and adaptability.

Integrating these techniques enhances testing efficiency by using model-based testing for test generation, applying data reduction to limit necessary test cases, and employing time reduction methods to prioritize tests informed by coverage and risk. Ultimately, these inductive approaches strive to optimize the testing process, utilizing data-driven techniques to deliver higher-quality tests while conserving time and resources.

STEPS TO CONDUCT AN EXPERIMENT ON TESTING TECHNIQUES

To evaluate various software testing methodologies, a structured experiment will assess five approaches—manual

testing, automated testing, model-based testing, randomized testing, and search-based testing—based on their effectiveness, efficiency, and coverage.

The experiment will use a real-world application as the software under test (SUT), which should have diverse functionalities for meaningful comparisons. Each methodology will be applied to the same version of the SUT to ensure comparable execution times. Evaluation metrics will include the number of defects found, test coverage, execution duration, and reproducibility.

The experiment will consist of five phases: initial manual testing, followed by automated, model-based, randomized, and search-based testing, with results documented after each phase. It is expected that automated and model-based testing will identify as many or more defects than manual testing, with model-based testing achieving the highest coverage. Manual testing is anticipated to be the slowest, while automated and model-based testing will likely be more efficient.

After data collection, an analysis will focus on defects per time unit, the relationship between coverage and defects found, and resource utilization, leading to conclusions about the most effective testing methodology for the SUT.

THE CONSUMER TESTING LANDSCAPE: DEFECTS AND IMPACTS

Fault-detecting software defects is crucial for maintaining high-quality applications and can be achieved through various techniques. Static analysis methods, such as code reviews and inspections, allow for manual examination of source code to identify logical errors and security vulnerabilities before execution. Automated static code analysis tools, like SonarQube, Coverity, and Checkstyle, analyze the code without running it, pinpointing issues like syntax errors and compliance violations.

Dynamic analysis techniques, executed during the software's runtime, include various levels of testing, such as unit, integration, system, and acceptance testing, as well as fuzz testing, which involves providing random or malformed inputs to identify unexpected crashes or vulnerabilities.

Additionally, fault injection is used to deliberately introduce errors to evaluate software resilience, and profiling tools, such as Dynatrace, Prometheus, or Jaeger, help track software behavior and detect runtime anomalies. Moreover, AI and ML-based defect detection methods, including anomaly detection models and predictive analytics, enhance the ability to identify abnormal behaviors and predict defect-prone modules based on historical bug data.

The impacts of software defects can be severe, affecting performance and reliability, leading to system crashes, memory leaks, and slow response times. Security risks include potential data breaches, privilege escalation, and denial-of-service attacks due to vulnerabilities. Financially, defects can result in revenue loss due to downtime, legal liabilities from compliance failures, and damage to reputation from poor software quality, which in turn affect customer trust.

Furthermore, software development is impacted by increased maintenance costs, as more defects necessitate additional time and resources for debugging and fixing, and delayed releases when defects are discovered late in the development cycle can push back deployment schedules.

To mitigate risks, organizations should implement a comprehensive testing strategy that includes unit, integration, system, and user acceptance testing. This approach enables early problem detection, ensuring software quality before launch. Gathering user feedback through beta testing and surveys is also essential for enhancing functionality and user satisfaction. By prioritizing thorough testing and user input, companies can reduce risks, improve quality, and build customer loyalty.

BRIDGING THE GAP: TESTING STRATEGIES FOR AI/ML

The rapid evolution of AI and ML technologies has fundamentally transformed various sectors, including health care and finance. As these systems become increasingly integrated into critical applications, the need for rigorous testing methodologies is more crucial than ever.

Traditional software testing approaches often fall short when applied to AI/ML applications due to their unique characteristics, such as reliance on large datasets, the probabilistic nature of outputs, and potential for adaptive learning. Effective and innovative testing methods for AI/ML systems include data quality analysis, model explainability techniques, like LIME and SHAP, adversarial testing, cross-validation, continuous monitoring of performance metrics, and human-in-the-loop evaluation. These methods address the distinct features of AI/ML models, particularly their dependence on data and the possibility of unexpected behavior in edge cases.

A thorough data-quality assessment is essential for identifying biases

and inconsistencies that can impact model performance. Employing model explainability techniques allows for insight into decision-making processes, helping to pinpoint biases and areas for improvement.

Adversarial testing is crucial for identifying vulnerabilities, while cross-validation enhances model evaluation by assessing generalization ability across different data subsets. Continuous performance monitoring is necessary to detect any degradation over time, and incorporating human judgment in the evaluation of model outputs is vital for complex tasks.

Additionally, performance testing evaluates the model's efficiency under varying load conditions, and canary deployments facilitate gradual rollouts for performance monitoring before wider deployment.

The limitations of conventional testing methods in assessing AI/ML systems underscore the necessity for these specialized methodologies,[6] as traditional methods rely on deterministic behavior and predefined input/output pairs, which are inadequate for the dynamic and probabilistic nature of AI models. Both approaches can complement each other. While AI-driven testing excels in handling complexity and scale, conventional methods are indispensable for tasks requiring human insight or deterministic testing.

Evaluating the impact of technologies like AI on security frameworks is essential due to emerging threats. The rise of AI and ML enhances threat detection and response through trend analysis and automation, but also enables more sophisticated attacks. A cohesive, proactive strategy is needed to integrate technology, policy, and ethics for a secure digital future.[4]

AI PROOF OF CONCEPT, APPLICATIONS AND CHALLENGES

An AI proof of concept (PoC) is a prototype designed to demonstrate the feasibility and benefits of an AI solution for a specific problem. Its main goal is to validate that a particular AI approach can achieve the desired results before investing in full development.

Key steps in developing an AI PoC are as follows:

1. *Define the problem:* Clearly outline the specific problem the AI will address, including business challenges and potential AI solutions.
2. *Identify the AI technology:* Select suitable AI technologies and algorithms for the PoC, such as ML, deep learning, natural language processing, or computer vision.
3. *Data collection:* Gather necessary data for training or evaluating the AI model, including historical data, user interactions, or sensor data.
4. *Model development:* Create a preliminary AI model by training it with the collected data to perform the designated task.
5. *Implementation:* Implement the AI model in a real or simulated environment to assess its performance and practical benefits.
6. *Performance evaluation:* Measure the model's effectiveness using metrics like accuracy, efficiency, and user feedback to demonstrate its value.
7. *Results and recommendations:* Evaluate the AI solution's broader feasibility or the need for further research based on PoC results.

Common applications for AI PoCs are:

- *Customer service:* A chatbot PoC can demonstrate how AI can handle basic customer inquiries, reducing workload on human agents.
- *Predictive maintenance:* AI models that predict equipment failure or maintenance needs based on sensor data and historical trends.
- *Health care:* AI PoC for analyzing medical images or predicting patient outcomes from historical data.

Challenges to an AI PoC are as follows:

- *Data quality:* AI models are only as good as the data they are trained on, and poor-quality data can lead to inaccurate results.
- *Complexity:* Developing an AI PoC may require specialized skills and resources, making it challenging for some organizations.
- *Scalability:* A PoC may succeed in a limited setting, but scaling it to production can present challenges.

An AI PoC is a crucial initial step in developing AI solutions, showcasing the technology's effectiveness and potential value before full-scale development.

EXPERIMENT: VALIDATION OF SOFTWARE TESTING METHODOLOGIES VERSUS AI-DRIVEN TESTING

An e-commerce platform is preparing for a major sales event, prompting the

development team to prioritize reliability, performance, and user experience under increased traffic. The focus is on evaluating traditional software testing methods, including manual and automated testing, against AI-based techniques. This analysis targets key features, like user authentication, product exploration, shopping cart and checkout, payment processing, and order management.

Key focus areas for evaluation during this major sales event include performance and load testing, which is essential for confirming the system's capability to manage increased traffic. Scalability testing is necessary to assess the effectiveness of autoscaling features. It is crucial to conduct thorough testing of checkout and payment processing to avoid failures during peak transaction periods. Additionally, search and recommendation engines must be evaluated to ensure that product searches and dynamic pricing adjustments function properly. Finally, security testing is vital for detecting fraudulent activities and bot interactions.

Three testing approaches will be evaluated: manual testing by human testers, traditional automated testing using tools like Selenium or Cypress, and AI-driven testing that dynamically generates and executes test cases using advanced tools, like Applitools, Test.ai, or Mabl.

When evaluating a testing methodology, key metrics include the defect detection rate, which measures the percentage of identified defects, and test execution time, reflecting the duration to complete tests. Test coverage assesses the extent of functionality tested, while false positives and false negatives indicate testing accuracy. Cost per defect is calculated by dividing total testing expenses by the number of defects identified, and scalability measures the methodology's ability to handle complex scenarios. Ease of maintenance considers the effort needed to update test cases. Together, these metrics provide a comprehensive assessment of testing performance.

The experimental design entails the preparation of the SUT by incorporating 50 identified defects, which encompass functional issues, UI problems, and edge cases. A uniform set of test cases is developed for both manual and automated testing, concentrating on the features of the SUT. The scenarios include logging in with both valid and invalid credentials, searching for products using various filters, adding items to the shopping cart, and processing payments. Additionally, the testing replicates concurrent user activities and evaluates the system's responses to invalid inputs. Teams of equal size are designated for manual testing, automated testing, and AI-driven testing to maintain equity. The results and metrics gathered during this testing phase are compiled in a summary (see Table 1).

TABLE 1. Results and metrics collected.

Metric	Manual testing	Automated testing	AI-driven testing
Defect detection rate	65% (33/50)	80% (40/50)	92% (46/50)
Test execution time	15 h	4 h	1.5 h
Test coverage	70%	85%	92%
False positives	5%	10%	2%
False negatives	10%	8%	5%
Cost per defect	US$200	US$80	US$40
Ease of maintenance effort	High	Moderate	Low
Scalability	Low	Moderate	High

The observations above indicate that AI-driven testing outperforms both manual and automated testing across key metrics. It achieves a 92% defect detection rate, identifying 46 of 50 defects, compared to 80% for automated and 65% for manual testing. AI testing completes in 1.5 h, while automated takes 4 h and manual takes 15 h. It also offers 92% test coverage, surpassing automated's 85% and manual's 70%. With a 2% false-positive rate and 5% false negatives, AI testing is more reliable than automated testing (10% false positives, 8% false negatives) and manual testing (5% false positives, 10% false negatives). The cost per defect is US$40 for AI, US$80 for automated, and US$200 for manual testing. AI requires minimal maintenance, while automated needs moderate effort and manual requires significant resources. Finally, AI testing is highly scalable, outperforming the moderate scalability of automated and low scalability of manual testing.

A few key aspects to note here are as follows:

- Conventional load testing relies on a fixed number of simulated users, limiting flexibility, while AI-driven testing uses dynamic traffic simulation that adapts to real-time demand.
- Conventional UI and functional testing use static scripts that often fail, while AI-driven methods utilize self-healing scripts that autonomously detect and fix UI issues.
- Conventional checkout and payment testing follows established test cases, while AI-driven testing proactively identifies potential failures and adapts dynamically.
- Conventional fraud detection uses static rules that can be easily evaded, whereas AI-based methods adapt to evolving fraud patterns.
- Finally, traditional downtime prediction is reactive, identifying issues post factum, while AI-driven solutions focus on proactive failure prevention through advanced forecasting techniques.

Figure 2 shows an AI-generated python test script using GPT and Selenium for automated checkout testing. AI can improve this test script by involving the intelligent selection of test steps based on past checkout failures. It utilizes self-healing automation to automatically adjust to modifications in UI components. Furthermore, AI can generate multiple variations customized for different product categories, payment methods, and user behaviors.

A comprehensive strategy is vital for ensuring software quality, involving usability, manual, regression, and automated testing. Usability testing identifies user experience issues, while manual testing provides detailed feature evaluations. Regression testing ensures that new code doesn't harm existing functions, and automated testing improves efficiency, especially with frequent updates. Additionally, AI-driven testing offers self-healing capabilities that adapt to code changes, enhancing edge case detection and accuracy. In summary, AI-enhanced testing is a cost-effective approach that streamlines the testing process while maintaining high quality.

```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
import time

# AI-powered adaptive test case
def ai_checkout_test():
    driver = webdriver.Chrome()
    driver.get("https://www.example-ecommerce.com")

    try:
        # AI identifies test steps based on previous failures
        search_box = driver.find_element(By.NAME, "search")
        search_box.send_keys("Laptop")
        search_box.send_keys(Keys.RETURN)
        time.sleep(2)

        # AI detects UI elements dynamically
        product = driver.find_element(By.XPATH, "//a[contains(text(),'Gaming Laptop')]")
        product.click()
        time.sleep(2)

        add_to_cart = driver.find_element(By.ID, "add-to-cart")
        add_to_cart.click()
        time.sleep(2)

        checkout = driver.find_element(By.ID, "checkout")
        checkout.click()
        time.sleep(2)

        # AI inserts dynamic payment scenarios
        payment_option = driver.find_element(By.ID, "payment-credit-card")
        payment_option.click()
        time.sleep(2)

        confirm_order = driver.find_element(By.ID, "confirm-order")
        confirm_order.click()
        time.sleep(3)

        print("Test Passed: Checkout completed successfully")

    except Exception as e:
        print(f"Test Failed: {e}")

    finally:
        driver.quit()

# Run AI-powered checkout test
ai_checkout_test()
```

FIGURE 2. AI-generated python test script using GPT and Selenium for automated checkout testing.

ABOUT THE AUTHOR

ANURAG SAXENA is a lead principal engineer focused on OpenShift Networking at Red Hat Inc., Boston, MA 02210 USA. His research interests include software-defined networking, intelligent cloud networking, and artificial intelligence. Saxena received an M.S. in telecommunications networking from Northeastern University, Boston, MA, in 2011. He is a fellow of both the Institution of Electronics and Telecommunication Engineers and Hackathon Raptors and a Senior Member of IEEE. Contact him at saxenaanurag84@gmail.com.

The evolution of software testing from manual to automated techniques has significantly reduced testing costs and increased efficiency. The integration of AI and ML in testing metrics has transformed quality assurance by automating test case generation, enhancing pattern recognition in test data, and enabling the creation of intelligent, self-healing test suites. Furthermore, the innovative security testing for OSS has leveraged automated fuzzing, AI-enhanced static analysis, and community-driven vulnerability scanning to identify vulnerabilities, outperforming traditional methods.

The rise of AI and ML in security testing has not only improved the detection and response to emerging threats, but has also enabled more sophisticated attacks, necessitating a cohesive, proactive strategy to integrate technology, policy, and ethics for a secure digital future. In addition, the integration of AI and ML in testing metrics has revolutionized the identification of potential issues throughout the development process.

The ongoing advancement and proliferation of AI have rendered ethical considerations paramount in discussions regarding its implementation. In summary, the integration of AI and ML in testing metrics has revolutionized the software testing landscape, leading to the creation of more intelligent, self-healing test suites and improving the efficiency and accuracy of test processes.

REFERENCES

1. P. Radanliev, "Digital security by design," *Secur. J.*, vol. 37, no. 4, pp. 1640–1679, 2024, doi: 10.1057/s41284-024-00435-3.
2. D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cogn. Sci.*, vol. 9, no. 1, pp. 147–169, 1985, doi: 10.1016/S0364-0213(85)80012-4.
3. I. Bartoletti, "AI in healthcare: Ethical and privacy challenges," in *Proc. Conf. Artif. Intell. Med. Europe (AIME)*, D. Riaño, S. Wilk, and A. ten Teije, Eds., Cham, Switzerland: Springer-Verlag, 2019, pp. 7–10, doi: 10.1007/978-3-030-21642-9_2.
4. M. Zheng, D. Xu, L. Jiang, C. Gu, R. Tan, and P. Cheng, "Challenges of privacy-preserving machine learning in IoT," in *Proc. 1st Int. Workshop Challenges Artif. Intell. Mach. Learn. Internet Things*, 2019, pp. 1–7, doi: 10.1145/3363347.3363357.
5. R. Ahmad and I. Alsmadi, "Machine learning approaches to IoT security: A systematic literature review," *Internet Things*, vol. 14, Jun. 2021, Art. no. 100365.
6. M. Islam, F. Khan, S. Alam, and M. Hasan, "Artificial intelligence in software testing: A systematic review," in *Proc. IEEE Region 10 Conf. (TENCON)*, Chiang Mai, Thailand, 2023, pp. 524–529, doi: 10.1109/TENCON58879.2023.10322349.
7. P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and responsible AI deployment," *Frontiers Artif. Intell.*, vol. 7, Mar. 2024, Art. no. 1377011, doi: 10.3389/frai.2024.1377011.
8. M. N. Zafar, W. Afzal, E. P. Enoiu, Z. Haider and I. Singh, "Optimizing model-based generated tests: Leveraging machine learning for test reduction," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation Workshops (ICSTW)*, Toronto, ON, Canada, 2024, pp. 44–54, doi: 10.1109/ICSTW60967.2024.00020.
9. "Open-source software: Risks and benefits for your business." FasterCapital. Accessed: Jun. 26, 2024. [Online]. Available at https://fastercapital.com/content/Open-Source-Software--Risks-and-Benefits-for-Your-Business.html
10. K. Ganesan, "AI development vs. traditional software engineering: Distinguishing the differences," *OPINOSIS Analytics*, 2024. [Online]. Available at https://www.opinosis-analytics.com/blog/ai-vs-software-engineering/
11. "Software testing trends: 2020 and beyond," *BIP. Monticello Consulting Group*, Jan. 31, 2020. [Online]. Available: https://www.monticellocg.com/blog/2020/1/31/software-testing-trends-2020

From Search Engines to Action Engines

Suman Nath, **Ryen W. White**, **Fazle E. Faisal**, **Morris E. Sharp**, **Robert W. Gruen**, and **Lenin Ravindranath Sivalingam**, Microsoft Research

With generative artificial intelligence (AI), there is progress in moving from search results to AI-generated answers that synthesize and summarize content. Research on AI agents and artificial capable intelligence aims to reach the next frontier in information access: task completion.

Search engines revolutionized the way we access information on the web. When users search, they get a list of web pages that might answer their questions. However, these search results are only starting points, and users still need to browse through them to extract the required information or to answer their query. Research on orienteering, postquery trailfinding, and teleportation (to popular trail destinations) (for example, White et al.[1]) has examined ways to support user activity beyond search engine result pages (SERPs) but still requires significant user engagement (which has its advantages in terms of, for example, cognitive development but can also be onerous).

BEYOND INFORMATION ACCESS

Fueled by recent advancements in generative artificial intelligence (GenAI), information systems have been adding capabilities that also make them *answer engines*. These engines can take a natural language question as input and generate an answer using knowledge from foundation models such as GPT-4o[a] and Gemini,[b] grounded in information found in search results related

Digital Object Identifier 10.1109/MC.2025.3556643
Date of current version: 29 May 2025

[a]https://openai.com/index/hello-gpt-4o.
[b]https://deepmind.google/technologies/gemini.

 0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

to the prompt. This evolution from results to answers saves users the trouble of having to formulate effective keyword queries and manually sift through one or more web pages that traditional search engines would retrieve in response.[2]

But what if the user needs to use the answer to perform some tasks online? Suppose Alice needs to register for a summer camp for her daughter. She 1) searches for summer camps around her location; 2) browses the camp websites to understand what is available and then picks a camp that best matches her needs; and 3) finally registers for the camp by visiting its website, navigating to its registration page, and completing and submitting a registration form. Traditional search engines would help Alice perform step 1, while answer engines would also help her achieve step 2 without needing to browse many camp websites. However, she still needs to perform step 3 on her own, which can take significant time.[c]

To address this challenge, we envision *action engines*, a further AI-enabled expansion of the capabilities of information systems beyond answer engines that would automatically perform step 3 (as well as prerequisite steps 1 and 2) on Alice's behalf. Given a natural language task description, the action engine would automatically perform the task and make any necessary decisions, engaging (for example, sharing progress, getting feedback, and raising exceptions) with Alice as necessary and based on her preferences and what the system knows about her (for example, from prior similar tasks).

Over the past few decades, we have seen remarkable progress in search engines that help us find, learn, and investigate (a past and present focus); in answer engines using GenAI for synthesis, summarization, and content generation (an emerging focus); and now in action engines that assist with task completion via automation (a required future focus). We envisage that all three will coexist in information systems, providing value for different tasks.

Task completion is the next frontier in information interaction.[3] The rise of action engines is driven by progress in task automation, AI agents, and agentic workflows. Designers of these engines will leverage AI technology to expand engine capabilities, moving beyond mostly supporting thought, as in search engines and answer engines, to supporting both thought *and* action. As AI continues to evolve and improve, it will become increasingly adept at understanding and executing complex tasks across a range of applications and websites, making action engines more widespread in the near future. This will revolutionize how we interact with the web and other applications, making it more efficient and more task- and action-centric. Users can go directly to using action engines for well-defined tasks, first explore alternatives via search/answer engines and then execute the task via an action engine, or trust the action engine to handle all aspects of the task. All these systems, working seamlessly together with users, will revolutionize how we complete tasks, making our daily lives more efficient and productive, freeing more time for activities that we enjoy and value.

SEARCH ENGINES TO ANSWER ENGINES

Search engines offer only limited support for task completion, such as 1) *inline answers* for tasks requiring information fragments, such as stock quotes and weather updates; 2) *knowledge cards* that provide immediate fact-based answers on topics such as celebrities, history, and health; 3) *deep links* to specific pages within a website that provide functionality, such as a store locator, tool rentals, and latest deals at a popular hardware store (note that this partial, navigation-only task support still requires user browsing actions thereafter); and 4) *query autocompletions* (and related searches, shown on SERPs) to suggest potential query formulations that may be popular with others. Even simple tasks that just require information for completion can take time and effort because users have to create queries, review results, and follow links after the search.

To help address these inefficiencies in search engine interaction, *answer engines* have been developed in recent years that are powered by GenAI. These engines provide quick answers to user questions, dynamically generate bespoke user experiences adapted to the task and/or the user, and auto-organize search results on SERPs for more optimal information consumption. Examples of answer engines can be found in mainstream web search engines, for example, in Google Generative Search,[d] or in separate conversational experiences such as Perplexity.[e]

Results are provided by answer engines dynamically and may draw attention away from clickable links on the SERP unless the user experience is carefully designed. To maintain a healthy information ecosystem, which is mostly funded by revenue from user engagement with source content, at least per the predominant economic

[c]Some of that time can be offset by web browser autocompletion functionality that people trust to use their browser history to generate relevant URL completions.

[d]https://blog.google/products/search/generative-ai-google-search-may-2024.
[e]https://www.perplexity.ai.

TABLE 1. A summary of some of the key differences among search engines, answer engines, and action engines.

Dimension	Search engine	Answer engine	Action engine
Input	Search query	Question	Complex task
Input format	Engine-dependent query language	Natural language (questions, clarification)	Natural language (tasks, clarification)
Target	Websites, web page content	Foundation model plus retrieval-augmented generation (RAG) (retrieved result content)	Functionalities of existing websites and applications
Output	Search results	AI-generated answer	Completion of task (result/effects)
Core technology	Crawling, indexing, ranking	Foundation model, RAG	Foundation model, agents, tools
Completion support	Low (instant answers)	Moderate (AI answers, content creation)	High [performs task, leveraging user's own applications and data (if needed)]
Human interaction	Frequent (queries, clicks)	Frequent (dialog, clicks)	Infrequent (feedback, unblocking)
AI	Traditional AI (neural networks, etc.)	GenAI	GenAI, agentic AI
Autonomy	Low [specific algorithms and rules, granular user engagement (queries, clicks)]	Moderate (needs user prompts/guidance)	High [executes tasks end to end, human involvement determined by task, user preferences, and task progress (errors)]
Learning	Rules/human feedback	Data-driven (using existing data)	Improves though experience

model on the Internet, search system designers need to retain links to that content on SERPs. This can be done through inline references in AI-generated answers. Ensuring the provenance of information also helps boost user trust in the system output.

Answer engines use the query and conversation context to generate an answer using a foundation model, enhanced by information from the search index and search results. They take several steps that users typically have to do themselves in interacting with search engines: 1) decomposing goals into subgoals or subqueries, 2) examining SERPs and selecting and reading web pages, and 3) aggregating and synthesizing relevant knowledge from retrieved information. Answer engines focus primarily on knowledge acquisition and offer only limited support for task completion, for example, planning and content creation. Table 1 presents some of the differences between search engines and answer engines along several key dimensions.

More generally, answer engines are also *copilots*, Microsoft's nomenclature for an application that uses modern AI to assist with complex tasks.[f] Copilots offer more general task assistance than information finding alone and also serve as companions and coaches. Users engage with copilots via conversational user interfaces (UIs) that let them interact over multiple turns via natural language and/or multimodal inputs. Copilots are powered by foundation models and are extensible with skills, tools, and plug-ins. Copilots exist as standalone chatbots or as chat embedded in other applications, for example, in development environments such as Visual Studio Code as an AI pair programmer. As foundation models become commoditized over time, the differentiation in AI applications will soon be in the user experience. That may need to be significantly reimagined from text-based prompting alone to more seamlessly integrate AI and support more users and more types of tasks with alternative input mechanisms.[4]

Copilots provide planning capabilities, for example, decomposing complex tasks, and some action capabilities. They are currently focused on content

[f] http://copilot.microsoft.com.

creation across text, images, and video. Their output can contain errors (hallucinations), and they can be impacted by biases in the data used to train them. They create new opportunities for human learning (for example, they can promote thought and prompt learners to think critically, unpack problems, and understand underlying concepts), but any increased automation can also increase metacognitive demands,[5] offer less human control over the information seeking process than users are accustomed to in search engines, and lead to less critical engagement and fewer opportunities to develop domain expertise. Automation has advantages, but copilot designers should only increase automation if they also have ways to help users retain control and responsibility.[6] Human agency and system automation are in tension in interactive systems that promise *intelligence augmentation*, and discussions about the thoughtful design of the user experience to accommodate both requirements in this age of AI are necessary.[7]

Looking ahead, there are several frontiers for answer engines and copilots more generally, including the following:

1. *Memory*: expansive memory for a seamless experience beyond the current session
2. *Personalization*: dynamically adapting to a user's preferences and habits (which depend on the existence of memory)
3. *Actions and planning*: taking actions on behalf of users to help them attain their task objectives
4. *Safety*: ensuring that any actions are taken responsibly, that information shared with users is accurate, and that sensitive data are handled with utmost care.

We focus our discussion in this article on frontier 3 (but also cover aspects of the other frontiers), given the emergence of GenAI-powered agents and the imminent and significant opportunity for humans and AI to work together to reimagine task completion on a global scale and the many implications of doing so. Doing this well, starting with digital but soon expanding into the physical realms through rapid progress in embodied AI, will involve combining world, enterprise, and user knowledge with real-world capabilities (for example, vision, audio, biosensing, spatial awareness, environment monitoring, and robotic actuation). Availability of data streams across applications, devices, and the cloud will be crucial for enabling richer task modeling, providing better complex task support, and helping more users complete more tasks. AI can already help complete repetitive tasks such as making reservations on users' behalf. For such tasks, action transformers, such as ACT-1,[g] trained on digital tools, and "tasklets" (scripts) learned from websites[8] have proven to be reasonable starting points as actuators of the digital world.

The recent (re)emergence of AI *agents* has the potential to transform how copilots affect the world. The tradeoffs between direct manipulation and interface agents has been a discussion in the human–computer interaction community for decades.[9] Agents can automate long-running business processes, reason over actions and user inputs, leverage memory to bring in context, learn based on user feedback, and record exception requests and ask for help from humans. Reasoning and acting (ReAct) agents such as UFO[10] are more robust than scripting solutions and can reason over observations (for example, interface updates) to determine agent actions. Going forward, more copilots will integrate one or more specialized agents to help perform complex tasks, and people will create custom agents via lightweight interactive environments, much like they do with document authoring today.

[g]https://www.adept.ai/blog/act-1.

ARTIFICIAL CAPABLE INTELLIGENCE

In an attempt to frame progress in the fast moving world of AI, OpenAI recently outlined five aspirational stages of AI development: 1) *Chatbots*, AI with conversational language; 2) *reasoners*, with problem-solving abilities; 3) *agents*, which can take actions; 4) *innovators*, AI that can aid in invention; and 5) *organizations*, AI that can perform the work of an entire organization.[h] As mainstream research and development in AI moves from the second to the third stage, artificial capable intelligence (ACI)[11] may represent the next attainable frontier in AI. Unlike the audacious pursuit of artificial general intelligence (humanlike intelligence) or superintelligence (smarter than humans), ACI focuses on achieving practical, actionable outcomes such as task completion. AI agents imbued with ACI are designed to tackle complex, multistep tasks given only a high-level description of user goals (intent) with minimal supervision, for example, plan and run a vacation, develop safer and more energy-efficient battery technologies, or even strategize to win an election.

The Turing test, proposed in 1950,[12] is a long-standing benchmark for whether a machine can exhibit humanlike intelligence. The test checks if a human user

[h]https://www.bloomberg.com/news/articles/2024-07-11/openai-sets-levels-to-track-progress-toward-superintelligent-ai.

can tell whether they are interacting with a machine or another human. It does not show if the machine understands the task or can create effective plans, which are key aspects of natural intelligence. In his recent book, *The Coming Wave*,[11] Mustafa Suleyman, CEO of Microsoft AI and DeepMind cofounder, proposes a modern Turing test for AI that captures those elements and more. It would involve giving an AI agent an open-ended, complex goal that requires interpretation, judgment, creativity, decision making, and action across multiple domains, for example, make a million dollars from an initial hundred thousand dollar investment, through various e-commerce activities. This test scrutinizes an AI agent's prowess in independently navigating the entire process of researching, planning, sourcing manufacturers, and selling products, thereby evaluating its effectiveness in accomplishing intricate objectives with minimal human oversight.

Agentic AI is characterized by its multistep, iterative process involving user instructions, planning, allocation, execution, and feedback.[1] Agentic systems possess a degree of autonomy, allowing them to make decisions, plan actions, and learn from experiences to achieve specific goals set by their human creators. The adaptability of these systems to achieve goals with limited supervision in complex environments is known as the *agentic property*. This shift toward agenticness is transforming AI and software systems, making agents the unit of programming in these workflows. Agentic workflows have a place on the web, in action engines, as a way to support users in tackling complex tasks. There are many challenges, including learning and surfacing the correct workflows (*task plans*, as we call them) for a user intent and generating them efficiently at a scale needed to be useful given the billions of websites and billions of users on the web. Additionally, these challenges include expanding beyond websites to also include applications, tools, etc., that people also use to tackle tasks and making them robust and resilient to system exceptions (for example, to challenge–response tests or to changes in the content of the resources they use) and adversarial activities from malevolent actors who may try to exploit action engine capabilities for nefarious purposes.

The principles of agentic AI include natural language input and output, modular task decomposition, interoperability with existing tools and platforms, continuous improvement through reinforcement learning, and a human-centric approach built around feedback. There is no one-size-fits-all architecture for agentic AI; instead, various design patterns are used to build agents, such as reflection, tool use, planning, and multiagent collaboration. Recent case studies, for example, AutoGen,[13] demonstrate the use of these patterns to, for example, develop stateful AI assistants, among many other possible applications.

The ecosystem of agentic AI is rapidly expanding, with many companies and startups entering the field. GenAI-powered applications, such as Amazon's Q, Anthropic's Computer use, and Google's Project Astra, are heralding a shift from knowledge-based to action-based systems. Startups such as Please, Imbue, and Snowflake are focusing on tasks such as workplace automation, personal assistance, and data analytics, showcasing the diverse potential of agentic AI. Action engines can leverage these third-party AI agents and other tools to address specific task needs during task automation, while keeping users in the loop to the extent that they need to be and prefer to remain involved.

TASK AUTOMATION AND ACTION ENGINES

An action engine is a system that can automatically execute complex tasks on existing websites (and applications), based on natural language descriptions from the user. The action engine interprets the user's intent and uses one or more foundation models, AI agents, or tools to accomplish the task. To accomplish a complex task, the action engine may break it down into a sequence of actions and perform them on relevant websites and applications, either by interacting with their UIs or underlying application programming interfaces (APIs) or by employing external domain-specific agents and tools. Sometimes, the action engine may need minimal user interaction to handle tasks beyond its capabilities, such as solving a captcha or obtaining additional information about that task that is not included in the initial task description, to deal with system exceptions when the output is unexpected or task progress is slow or to give users more control over the action engine's operation (for example, the specific website, application, or tool it uses to make task progress).

Two key components of an action engine are a planner and an executor. Figure 1 shows a high-level architecture of an action engine, and Table 1 presents some characteristics of action engines in relation to both search engines and answer engines. The planner receives the task the user wants to complete as input and generates a plan of actions that are required to complete the task. The planner may use the user's profile, preferences, and task history of completed

[1]https://blogs.nvidia.com/blog/what is agentic ai.

tasks and leverage foundation models or specialized planning algorithms. For instance, a user may ask the action engine to "renew all checked-out library books for one more month." The action engine then creates a plan of actions that includes 1) opening the user's default library website (found from the user's profile), 2) logging in if necessary and if the user has given login consent, 3) navigating to the "checked-out books" section of the website, 4) selecting all books, 5) clicking on the "renew" button, and finally, 6) verifying that the final page contains the confirmation text "all books renewed." The executor carries out each step in the action plan by interacting with the target app or website UI or underlying APIs. The executor may leverage foundation models to determine how to perform an action. For example, to execute the action "click the 'renew' button," the executor can use a language model to locate the target button on the current website screen. The executor may also use the user's profile and preferences (for example, to log into their account or to choose an option), external tools, or even AI agents. For example, if the library website has an AI agent that can do the renewal task, the action engine may interact with it to perform the task. The planner and the executor may work together: the planner may generate a few steps that the executor can perform; based on the outcome of these steps, the planner may generate more steps for the executor. The loop may also include the user, to clarify the user's intent, to handle exceptions, to help the agent perform new tasks, or to complete a complex task with human–agent collaboration. Human inputs and feedback can be leveraged to fine-tune or retrain the foundation model to improve its accuracy.

Note that the action engine can utilize existing APIs to perform tasks, but APIs are not mandatory for task execution. A significant advantage of our design is that the action engine can accomplish tasks by interpreting and interacting with application or website UIs, in the same way humans do. The strong emphasis on safely and robustly performing these actions *at web scale* is what distinguishes action engines from typical AI assistants, which often target specific domains or select websites.

We can leverage existing technology such as large language models (LLMs) and web automation frameworks (for example, Playwright[j]) to build the planner and the executor. However, to make them useful and reliable, we need to overcome several novel challenges. First, the planner has to infer the user's implicit and explicit intent; ask the user for any missing or disambiguating information; and use the user's profile, preferences, and task history to generate a complete or partial action plan. To generate the plan, the planner has to understand the set of possible actions, as well as their semantics, that can be performed on different websites or applications. The plan has to be accurate and safe, meaning that it not only achieves the user's goal but also avoids any unintended side effects. The plan also may be personalized, taking into account the user's long-term and short-term interests, as well as the user's feedback and behavior over time. Personalization can be achieved by first mapping the identified intents to task plans of other users with similar interests and preferences (for example, as has been proposed in search engines[14]) and then refining the plan based on the individual user's characteristics and preferences.

Second, the executor has to perform each step of the plan by identifying the correct interaction target (for example, UI control, API, tool, or agent) and verifying that each action is performed as expected. The executor has to ensure that the task execution is safe, reliable, secure, and (preferably, as much as is permissible/necessary) personalized. Speed may also be preferable for some tasks, depending

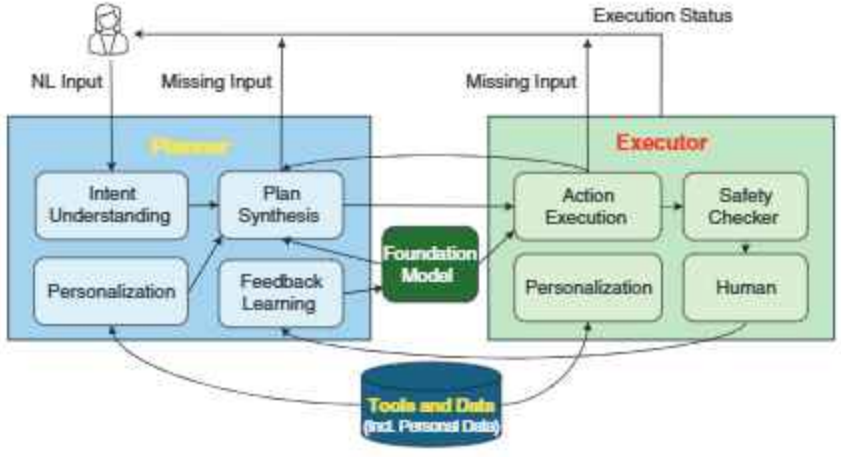


FIGURE 1. An example architecture for an action engine, with the user playing a central role various ways in the task completion process, for example, to initiate the task, receive status updates, provide feedback, and handle exceptions. NL: natural language.

[j] https://playwright.dev/.

on the task and the circumstances. It should also be able to involve humans for any critical steps, such as unblocking an operation with a login or providing additional information. The planner and the executor must evolve over time, improving based on their experiences and feedback across multiple websites or applications, and improving their performance and robustness. Reinforcement learning can be useful here, using, say, task outcome as the reward function.

Recent advances in foundation models, such as large action models[15] and AI agents,[10,16] show promise to address some of the above challenges. For example, several models have demonstrated impressive capabilities in understanding websites and their screenshots to identify actions, but their accuracy is still below human level. One possible way to improve their accuracy is to leverage website annotations from website developers. Similar to how websites today are annotated with accessibility information, we envision that future websites will also be annotated with information targeting AI agents (similar to Apple App intents,[k] where developers can specify application capabilities to the system). Standardizing such agent-friendly website annotations, akin to WAI-ARIA standards,[l] can foster the adoption and utilization of this information by action engines. An agents.txt file listing the actions is supported by a website that can also be hosted in the root directory of the site in a similar way to how robots.txt informs the operation of search engine crawlers. Additionally, there is an ongoing effort for llms.txt,[m] which is focused on helping LLMs not only answer questions about a website but also navigate and understand documentation, such as libraries. While llms.txt targets answer engines, agents.txt is designed for action engines.

The future is bright and holds great promise for action engines. Together with search engines and answer engines, they will enhance information systems with a wide array of capabilities, enabling them to handle (in cooperation with users but also on their own in some cases) a variety of tasks, from simple information retrieval to complex knowledge work and real-world actions.

BRIGHT FUTURES

As we look toward a bright future for action engines, it is important to consider the implications from their development.

Human and technology implications

To build action engines and implement them effectively at the web scale, we need to consider human and technology implications.

One area is *intent understanding*, which involves determining the task at hand from a natural language description plus any available additional context (for example, recently accessed websites or enterprise content), decomposing it as needed into achievable subtasks, and representing it accurately in the system for downstream processing. The initial task description may be imprecise, so the system may need to identify critical missing information and iterate with the user to develop a task model that more fully represents the task and the user's intentions in completing it.

Another area is *scalability*. Running foundation models at scale (for example, to interpret UIs and devise task plans) can be costly and inefficient. Ideally, an action engine should only do this once for every intent →website/application pairing until the task plan is no longer valid and needs to be regenerated. Hence, we need to explore scalable caching solutions that can store generated plans and retrieve them on demand per the task at hand. Many of the highly optimized web-scale technologies for search engines (for example, for crawling, indexing, filtering, and ranking web content) could be adapted and applied for this purpose.

Customization of the action engine, primarily to different users,[n] includes memory and related concepts of user authentication and personalization.

The *user experience* in action engines is fundamental to their adoption and continued use over time. Action engines can operate synchronously (where users can observe and validate their every step) and asynchronously (when the user is not present) depending on the task complexity, the severity of action consequences should execution go awry, user preferences, and so on. An effective action engine experience includes providing users with ways to track progress in tasks, supervising tasks as they are completed by action engines, guiding users and assisting as appropriate, and enabling users to provide feedback across various device modalities (small screens, no screens, etc.). When proactively engaging users, the high cost of consulting them needs to be balanced against the risks of not (for example, costly system errors). Explainability is important for users in building trust given that agent reasoning can be complex and the actions of action engines may be difficult to debug and understand. Principles for the design of

[k] https://developer.apple.com/documentation/appintents/app-intents.
[l] https://www.w3.org/WAI/standards-guidelines/aria.
[m] https://llmstxt.org.

[n] This could also cover adaptation to different tasks, domains, industries, and so on.

mixed-initiative interfaces[17] will help designers combine automated services with direct manipulation in the right ways for users and their tasks.

Human feedback on task success from agent execution, similar to result clickthrough and online advertising conversions in search engines, alongside other means of tracking action engine performance at scale will be useful for auditing, verification, and validation (for example, to detect errors in system operation and flag them to users, to designers, and to the action engine system for triggering the reconstruction of any cached task plan). Initially, users will likely also welcome opportunities to teach the system, communicate their preferences and specify policies to the action engine via direct feedback, and supervise its actions, even if they are not directly involved in their execution.

Safety is another critical area, including validation, robustness, and risk mitigation. Action validation ensures that actions represented in the task plan can still be performed on the site or application, while risk assessment is needed to involve the user for more risky actions where their feedback or supervision may be necessary. Policies could also be specified that help limit the extent of the risk (for example, not making purchases over a certain amount or not performing irreversible actions, such as sending an e-mail, without first consulting with the user). Initially, users may only be comfortable with the action engine working on rudimentary tasks that are of low risk. Over time, as the system shows a low propensity to make mistakes and learns more about the user (for example, preferred websites and applications or brand preferences), users may trust the action engine to execute more tasks on their behalf, even asynchronously, as mentioned above, when the user is not available to handle exceptions immediately.

Action engines also need to be able to engage with the task *ecosystem*. Some tasks will require the action engine to use external tools. To do so, the foundation model must know about those tools, how they operate, and the data needed for their operation. This also presents several challenges, such as the need for trusted execution to safeguard security, data, and privacy, especially as data move among websites, applications, APIs, and models. Models can run locally without sending data to untrusted third-party AI providers, achieving faster inference, enhanced governance, and improved data security.

The physical *environment* may also form a future frontier for action engines, expanding their task completion capabilities from digital settings (for example, using applications), to the real world (for example, using physical artifacts), with progress in areas such as embodied AI, robotics, and Internet of Things providing vital enabling technologies. Imbuing action engines with physical capabilities may also require another layer of protection, security, and audit and ways of locating and using objects in the physical world, an area where there has already been some early research.[18]

Evaluation is another critical area. Metrics and benchmarks are needed to measure various aspects of action engines and progress over time as updates are made. Existing benchmarks for autonomous agents are useful, but there is a need for more complex and diverse benchmarks. These benchmarks also focus on the accuracy of task outcomes, but we need to also measure the reliability, safety, and performance of action engines via new benchmarks with injected faults. The benchmarks also need to look beyond static trajectory matching to also consider *execution-based evaluation*,[19] focused on task outcomes, not agent processes. The focus on binary task outcomes (complete/incomplete) could also be expanded to task progress (for example, the fraction of complex task steps that have been completed) and partial task success (for example, the fraction of complex tasks with the correct outcomes).

Societal and economic implications

The broader implications of action engines are significant, spanning economics, ethics, and the web ecosystem, among other areas.

Online advertising is the major economic driver of the web. Just as with answer engines, action engines need to employ a suitable incentive model for web developers to adopt the action engine without discouraging the organic growth of the Internet. The "paradox of reuse"[20] suggests that fewer visits lead to less content being created and in turn could lead to lower-quality foundation models. We expect that action engines will disrupt the current Internet economy by significantly altering the online advertising model as they let users complete tasks on a website without visiting it or seeing advertisements. Business models for agents and task completion could revolve around a combination of subscriptions and revenue sharing with the websites or applications that the agents use. Action engines could also receive payments for using specific task providers. There may also be advertising pass-throughs from utilized websites to reports generated by action engines (for example, for exceptions or signifying task completion) or "action advertisements," similar to search advertisements but surfacing in the action engine UI after user intent specification.

Moving beyond a caching solution for a single action engine, independent action marketplaces could emerge that host collections of task plans for later reuse, creating a robust ecosystem. In doing so, we may need to look beyond the web and into application repositories (for example, app stores and agent repositories such as the GPT Store[9]) and develop tooling and interface standards that enable interface elements across a range of non-web applications to be exposed for downstream processing and use by foundation models. Agents must also be capable of connecting together several applications and/or websites to support multifaceted complex tasks. This necessitates considerations for interoperability and standardization to ensure seamless integration.

Action engines, which automatically perform tasks on behalf of users, inherently carry higher risks than answer engines, which merely provide information. They can be brittle, and autonomy is a threshold where risk increases significantly. To mitigate these risks, the system must incorporate robust safeguards to ensure safety and reliability. Before executing potentially harmful or risky actions, such as those with financial or legal consequences or actions that are irreversible and might misalign with user intent, the system must reliably identify them and involve users. It should adhere to users' policies, distinguishing among tasks that can be automated and those requiring explicit consent. Furthermore, the system must respect users' privacy by obtaining consent prior to using any private information.

The main ethical considerations around search engines and answer engines include ensuring accuracy, preventing misinformation, protecting user privacy, avoiding bias, and maintaining transparency. Many of these issues also apply to action engines, with additional issues such as the potential for misuse. Given the scale and the dynamics of the web, issues such as robustness, efficiency, and scalability/cost will also play a significant role in design decisions related to how action engines are rolled out. For example, they are not applicable for all tasks, and we may want to start with tasks where we can be more confident that task plans can be executed seamlessly by action engines (for example, sites with fewer changes in their user experiences over time, sites where

[9]https://openai.com/index/introducing-the-gpt-store.

ABOUT THE AUTHORS

SUMAN NATH is a partner research manager at Microsoft Research, Redmond, WA 98052 USA. His research interests include performance and reliability of distributed and AI systems. Nath received a Ph.D. in computer science from Carnegie Mellon University. He is Fellow of IEEE and an ACM Distinguished Member. Contact him at suman.nath@microsoft.com.

RYEN W. WHITE is a partner research director and general manager at Microsoft Research, Redmond, WA 98052 USA. His research interests include human–machine collaboration in search engines and AI assistants. White received a Ph.D. in computer science from the University of Glasgow. He is an ACM Fellow and an IEEE Senior Member. Contact him at ryenw@microsoft.com.

FAZLE E. FAISAL is a senior research engineer at Microsoft Research, Redmond, WA 98052 USA. His research interests include architecture, efficiency, and reliability of agentic AI systems. Faisal received a Ph.D. in computer science and engineering from the University of Notre Dame. Contact him at fafaisal@microsoft.com.

MORRIS E. SHARP is a senior research engineer at Microsoft Research, Redmond, WA 98052 USA. Sharp received a Ph.D. in computational chemistry from the University of Chicago. Contact him at morrissharp@microsoft.com.

ROBERT W. GRUEN is a principal research engineer in the Office of the Chief Technology Officer, Microsoft, Redmond, WA 98052 USA. Gruen is currently helping design Microsoft's AI platform of the future. Contact him at robgruen@microsoft.com.

LENIN RAVINDRANATH SIVALINGAM is a researcher and innovation strategist at Microsoft Research, Redmond, WA 98052 USA. His research interest include systems, programming frameworks, and artificial intelligence. Sivalingam received a Ph.D. in computer science from the Massachusetts Institute of Technology. Contact him at lenin@microsoft.com.

metadata are explicitly shared with the action engine by site designers, and tasks with minimal or no financial impact). Social and cultural norms around how work gets done and tasks get tackled will influence how action engines are launched and adopted and their suitability for different markets.

Ultimately, we want information systems to do more than just provide information; we want them to also help action tasks and reduce the execution burden on people, improving task outcomes and giving people time back to focus on activities that they value. Setting goals and having systems help attain them is a key aspect of progress toward ACI and ultimately a world where humans can focus on joyful, creative activities while systems handle administrative toil. Just as answer engines have expanded the search frontier beyond recall to other cognitive activities such as synthesis and analysis, action engines are poised to take these information systems even farther, toward action, task automation, and task completion, working collaboratively with humans to build up trust so that action engines can execute some tasks (especially menial and repetitive ones) on their own and keep extending the task frontier outward to support more complex tasks. To ensure success, we need robust controls, safeguards, and guardrails and systems that are transparent enough to build that trust, with a firm focus on enabling greater human learning and critical thinking in addition to more automation and completion. ■

REFERENCES

1. R. W. White, M. Bilenko, and S. Cucerzan, "Studying the use of popular destinations to enhance web search interaction," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 159–166, doi: 10.1145/1277741.12777.
2. S. Suri et al., "The use of generative search engines for knowledge work and complex tasks," Microsoft, Redmond, WA, USA, Tech. Rep. MSR-TR-2024-9, Mar. 2024.
3. R. W. White, "Opportunities and challenges in search interaction," *Commun. ACM*, vol. 61, no. 12, pp. 36–38, 2018, doi: 10.1145/3195180.
4. M. R. Morris, "Prompting considered harmful," *Commun. ACM*, vol. 67, no. 12, pp. 28–30, 2024, doi: 10.1145/3673861.
5. L. Tankelevitch et al., "The metacognitive demands and opportunities of generative AI," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 2024, pp. 1–24, doi: 10.1145/3613904.3642902.
6. B. Shneiderman, *Human-Centered AI*. Oxford, U.K.: Oxford Univ. Press, 2022.
7. J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," *Proc. Nat. Acad. Sci.*, vol. 116, no. 6, pp. 1844–1850, 2019, doi: 10.1073/pnas.1807184115.
8. Y. Li and O. Riva, "Glider: A reinforcement learning approach to extract UI scripts from websites," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1420–1430, doi: 10.1145/3404835.3462905.
9. B. Shneiderman and P. Maes, "Direct manipulation vs. interface agents," *Interactions*, vol. 4, no. 6, pp. 42–61, 1997, doi: 10.1145/267505.267514.
10. C. Zhang et al., "UFO: A UI-focused agent for windows OS interaction," 2024, *arXiv:2402.07939*.
11. M. Suleyman, *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*. New York, NY, USA: Crown, 2023.
12. A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.
13. Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversations," in *Proc. 1st Conf. Lang. Model.*, 2024.
14. J. Teevan, M. R. Morris, and S. Bush, "Discovering and using groups to improve personalized search," in *Proc. 2nd ACM Int. Conf. Web Search Data Mining*, 2009, pp. 15–24, doi: 10.1145/1498759.149878.
15. J. Zhang et al., "xLAM: A family of large action models to empower AI agent systems," 2024, *arXiv:2409.03215*.
16. S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," 2023, *arXiv:2210.03629*.
17. E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 1999, pp. 159–166, doi: 10.1145/302979.303030.
18. K. Kaneda, S. Nagashima, R. Korekata, M. Kambara, and K. Sugiura, "Learning-to-rank approach for identifying everyday objects using a physical-world search engine," *IEEE Trans. Robot. Autom.*, vol. 9, no. 3, pp. 2088–2095, Mar. 2024, doi: 10.1109/LRA.2024.3352363.
19. R. Bonatti et al., "Windows agent arena: Evaluating multi-modal OS agents at scale," 2024, *arXiv:2409.08264*.
20. N. Vincent, "The paradox of reuse, language models edition," *Mataroa Blog*, Dec. 2, 2022. Accessed: Nov. 17, 2024. [Online]. Available: https://nmvg.mataroa.blog/blog/the-paradox-of-reuse-language-models-edition/

General and Agentic AI, and the Challenges of Explainable Reliability

Angelos Stavrou, Virginia Tech University

Jeffrey Voas, IEEE Fellow

A relationship between artificial intelligence and traditional reliability models is explored.

Recent advances in generative artificial intelligence (GenAI) and large language models (LLMs) have given rise to tremendous potential and expectations for industrial and everyday applications. GenAI applications have begun to appear in market verticals, ranging from manufacturing to medicine to software engineering. GenAI applications promise to automate and disrupt the status quo.[1] However, as with every disruptive technology in its infancy, it is necessary to understand its advantages and use cases while reducing operational risks. While there is a lot of excitement, engineers are skeptical about using GenAI without guarantees of operational reliability, security, and, in some cases, safety.

How can reliability engineers benefit from the application of AI? Traditional reliability engineering techniques, including root cause analysis, failure mode and effect analysis, physics of failure, and condition-based monitoring, rely heavily on data analysis methods to proactively identify potential failures and implement preventative measures. Many of these approaches depend on field data analysis. That involves collecting and analyzing data from operational systems to identify trends and patterns, with the goal being to identify and forecast the risk of failures. In many cases, the field data are incomplete, lacks the necessary quality, or can be erroneous.

GenAI represents a class of machine learning algorithms capable of sifting through massive amounts of data by searching for patterns and trends across multiple siloed data feeds. Further, GenAI can automate arduous manual tasks, visualize results crucial to understanding reliability risks, and guide decision-making processes. As pointed out in Defense – SCSP,[2] "Presently, the most promising aspect of generative AI models is as a decision aid, or what we would term a cognitive copilot." GenAI differs from prior machine learning approaches in that it can address the challenge of missing field data for reliability analysis by creating synthetic data points that mimic the patterns of existing, missing, or erroneous data. GenAI can

DISCLAIMER

The authors are completely responsible for the content in this column article. The opinions expressed here are their own.

Digital Object Identifier 10.1109/MC.2025.3555462
Date of current version: 29 May 2025

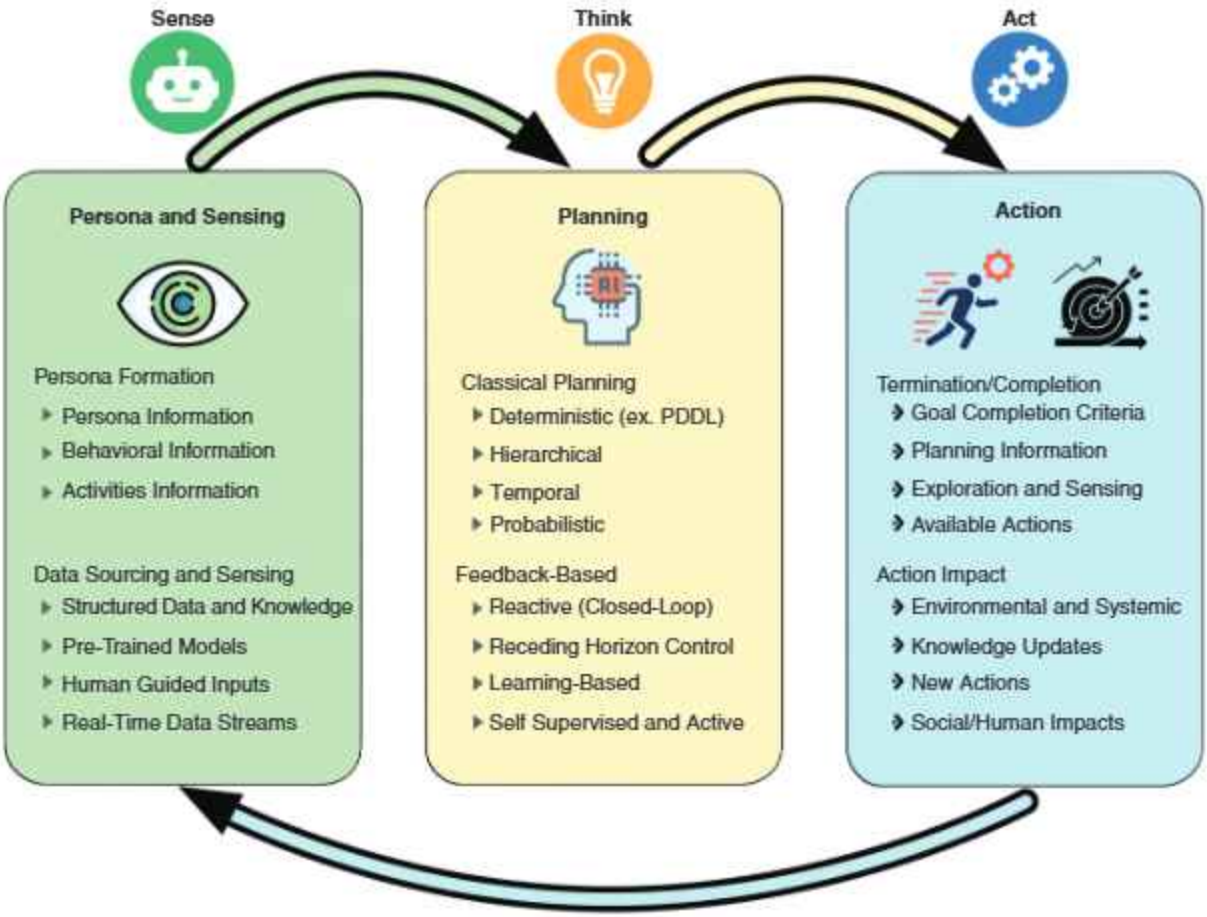
0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

effectively fill in gaps in datasets with values based on learned relationships between variables, empowering reliability engineers to assess reliability risks when complete data are unavailable. Thus, by leveraging continuous learning and adaptation when new real and new synthetic data are produced, GenAI has the potential to revolutionize system reliability through predictive maintenance strategies to anticipate and prevent potential failures before they occur.

Software engineers were among the first to benefit from generative AI to produce new source code and modify existing code. AI plugins were created for popular software integrated development environments to facilitate software development and code maintenance. AI code generation streamlines development by eliminating repetitive tasks, for example, crafting boilerplate functions or configuring project templates, thus saving time and boosting productivity. Meanwhile, intelligent process automation (IPA) takes efficiency a step further by seamlessly coordinating complex tasks such as: 1) incident management, 2) updates to infrastructure-as-code, and 3) the orchestration of continuous integration/continuous deployment workflows. This reduces downtime and lightens the load of operational maintenance. Additionally, multimodality empowers AI to juggle diverse data types such as text, images, audio, and video, thus fostering a smoother fusion of design, development, and documentation processes for more cohesive and dynamic workflows.

AI has a transformative potential for modern engineering systems, especially large software and hardware designs that may feel like rigid and unyielding relics with translational errors (when aligning human aspirations with computational execution). Generative AI can turbocharge the "old-school" software development grind, streamlining the leap from concept to code, for example, GitHub Copilot, which has slashed coding time by up to 55% in studies.[3]

Enter *living software systems*[4] that are fueled by generative AI and promise to tackle this core computing (code development) challenge head on. Typically, crafting software is a clunky process, riddled with flawed handoffs. Business needs get distilled into requirements that then get morphed into code through layers of interpretation that leave systems brittle and ill-equipped to pivot as user and business demands or circumstances change. Generative AI, specifically LLMs, can be a game changer, that is, a near-magical interpreter bridging the gap between what people want and what machines can do.



Having automated interpreters that can adapt based on usage opens doors to systems that aren't only static but are dynamic partners that can collaborate and that are attuned to context and capable of evolving alongside user objectives.

Newly developed AI agent-based architectures extend the ability to plan and autonomously execute tasks across multiple steps with minimal or no human interaction. For instance, the NVIDIA blog[5] defines Agentic AI as using sophisticated reasoning and iterative planning to solve complex, multistep problems. At the same time, the TechTarget article[6] emphasizes its ability to make decisions and adjust behaviors autonomously.

While Agentic AI systems employ LLMs to perform tasks that benefit from flexibility and dynamic responses, they leverage traditional programming for strict rules, logic, and performance. This hybrid approach enables AI to be more intuitive and precise.[7] This allows critical processes (such as security or calculations) to rely on deterministic, traditional algorithms. Agentic AI offers the potential to create systems that don't just react but proactively adapt based on their actions and environmental feedback. Autonomous AI agents can analyze data, set goals, and take actions with decreasing human supervision.[7] AI agents can orchestrate workflows on the fly, akin to how IPA reshapes decision-making and dynamic problem-solving and learning (through each interaction). The technical leap and primary difference between GenAI and Agentic AI systems is that GenAI focuses on creating content. In contrast, Agentic AI focuses on taking semi or fully autonomous actions while evolving and adapting on each iteration. Let's look at three types.

- *Simple Reflex Agents*: This is the most basic type, performing one specific task reliably and consistently based on immediate sensory input without memory. They operate on predefined condition-action rules, such as a thermostat turning on heating at a set time every night, as noted in the IBM article.[8] They are effective in fully observable environments but cannot learn or adapt.[7]
- *Model-Based Agents*: These agents can use current perception and draw on memory, enabling them to receive and store new information and perform a broader range of tasks. They maintain an internal state or model of the world, allowing them to handle partially observable environments. An example is a robotic vacuum

TABLE 1. Agentic AI use cases by industry.

Industry	Use Cases	Sources
Customer Service	Automating customer support, digital humans for support, personalized real-time interactions, handling tickets and FAQs	aimultiple.com, uipath.com, thoughtspot.com, moveworks.com, nvidia.com
Healthcare	Medical data analysis, patient care and monitoring, telemedicine, drug discovery and development, virtual caregiving	uipath.com, hbr.org, daffodilsw.com, ibm.com, nvidia.com
Software Engineering	Automating coding testing debugging code reviews, AI code assistants, building applications and application programming interfaces	aimultiple.com, uipath.com, thoughtspot.com, moveworks.com, aisera.com, nvidia.com
Gaming	AI agents for gameplay testing NPC behavior	aimultiple.com
Supply Chain Management	Optimizing logistics, Inventory management, order placement and production scheduling	uipath.com, hbr.org, ibm.com, aisera.com
Travel Planning	Autonomous trip planning and arrangements	hbr.org
Video Analytics	Video search summarization anomaly detection	nvidia.com
Cybersecurity	Monitoring network traffic detecting threats real-time response	moveworks.com, ibm.com
Finance	Insurance claims processing, underwriting, financial decision making, expense reporting compliance reporting	aimultiple.com, uipath.com, daffodilsw.com, moveworks.com
Human Resources	HR assistance recruitment employee engagement, payroll automation, onboarding and training	aimultiple.com, moveworks.com, ibm.com
Retail and E-commerce	Personalized recommendations, customer service	thoughtspot.com
Content Creation	Generating content for marketing	aimultiple.com, nvidia.com
Business Intelligence	Data analysis reporting, sales and marketing insights	thoughtspot.com, aimultiple.com

cleaner that remembers which areas it has cleaned, adjusting its path accordingly, as mentioned on the Restackio page.[9]

- *Learning Agents*: These agents can ingest new data and use it to inform later decisions, improving accuracy over time through learning. They can adapt their behavior based on experience relying, in general, on which is composed of a profiling module, a memory module, a planning module, and an action module (see **Figure 1**).[10] The purpose of the profiling module is to identify the agent's role. The memory and planning modules place the agent into a dynamic environment, enabling it to recall past behaviors and plan future actions. The action module translates the agent's decisions into specific outputs. Within these modules, the profiling module impacts the memory and planning modules, and collectively, these three modules influence the action module.

How far can we extend "living software systems" to enable "living engineered systems"? And where does the boundary of adaptability lie when we try to update old designs and as we forge new designs?

This is an open question that we will answer as AI becomes more integral in our everyday lives. But can we assess the quality of synthetic data and trust the outputs that GenAI systems produce? Unfortunately, the answer is not a resounding "Yes." Instead, a more pragmatic assessment is that, in many cases, AI can reduce arduous tasks that engineers must perform to a smaller set of tasks that they can validate for correctness. To help analysts better understand the outputs from AI, a new branch of research and products called explainable AI (XAI) has evolved. XAI aims to provide context and explain the outputs generated by AI. Although XAI is a means to enhance the trustworthiness of generative AI, it cannot entirely resolve concerns about the accuracy and reliability of AI outputs. XAI is beneficial because generative AI models are often complex, and that makes it challenging for humans to understand how output results were determined. In many cases, outputs are too detailed or abstract for humans to comprehend. To make matters worse, it is also unclear how XAI methods should be evaluated, how different terms should be applied, and how XAI relates to trustworthiness.[11] Therefore, despite advances, breakthroughs, and applications of XAI methods, more research is required before we can take advantage of the full potential of AI for safety-critical applications and complex engineering tasks.

We acknowledge that GAI can produce bad code and invent facts. A grand challenge is *how can these problems be prevented*? And since they're persistent in the current state of the art, will they cause users to lose trust in GenAI outputs? This story is still being written.

To conclude, we believe that the principles of XAI can create a pathway to xplainable reliability. Making reliability theory and models easier to understand and more transparent for humans should be "accomplishable" alongside the near-daily advances in AI. ■

REFERENCES

1. F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *J. Inf. Technol. Case Appl. Res.*, vol. 25, no. 3, pp. 277–304, 2023, doi: 10.1080/15228053.2023.2233814.
2. "Department of defense adoption of generative artificial intelligence," SCSP, Arlington, VA, USA, 2023. [Online]. Available: https://www.scsp.ai/reports/gen-ai/defense/
3. J. Bauer, "Does GitHub Copilot improve code quality? Here's what the data says," *GitHub Blog*, Nov. 18, 2024. [Online]. Available: https://github.blog/2024-11-18-does-github-copilot-improve-code-quality-heres-what-the-data-says/
4. J. White, "Building living software systems with generative & agentic AI," 2024, *arXiv:2408.01768*.
5. E. Pounds, "What is agentic AI?" *NVIDIA Blog*, Oct. 22, 2024. [Online]. Available: https://blogs.nvidia.com/blog/what-is-agentic-ai/
6. L. Craig. "What is agentic AI? Complete guide." TechTarget. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/agentic-AI
7. D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18,912–18,936, 2025, doi: 10.1109/ACCESS.2025.3532853.
8. C. Stryker. "Agentic AI: 4 reasons why it's the next big thing in AI research." IBM. [Online]. Available: https://www.ibm.com/think/insights/agentic-ai
9. "Simple reflex agent and model-based reflex agent." Restack. [Online]. Available: https://www.restack.io/p/agent-architecture-answer-simple-reflex-model-based-cat-ai
10. L. Wang et al., "A survey on large language model based autonomous agents," *Frontiers Comput. Sci.*, vol. 18, no. 6, Mar. 2024, Art. no. 186345, doi: 10.1007/s11704-024-40231-1.
11. L. Longo et al., "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Inf. Fusion*, vol. 106, Jun. 2024, Art. no. 102301, doi: 10.1016/j.inffus.2024.102301.

ANGELOS STAVROU is a professor of computer science and the entrepreneurship leader at the Innovation Campus at Virginia Tech University, Alexandria, VA 22305 USA. He is a Senior Member of IEEE. Contact him at angelos@vt.edu.

JEFFREY VOAS, Gaithersburg, MD 20899 USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.

CALL FOR SPECIAL ISSUE PROPOSALS

Computer solicits special issue proposals from leaders and experts within a broad range of computing communities. Proposed themes/issues should address important and timely topics that will be of broad interest to *Computer*'s readership. Special issues are an essential feature of *Computer*, as they deliver compelling research insights and perspectives on new and established technologies and computing strategies.

Please send us your high-quality proposals for the 2025–2026 editorial calendar. Of particular interest are proposals centered on:

- 3D printing
- Robotics
- LLMs
- AI safety
- Dis/Misinformation
- Legacy software
- Microelectronics

Proposal guidelines are available at:
www.computer.org/csdl/magazine/co/write-for-us/15911

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3562303

From Data to Action: Building Healthy and Sustainable Open Source Projects

Dawn Foster, CHAOSS

This article provides advice and resources for proactively using metrics to improve open source project health and sustainability before a crisis occurs to make software more sustainable and reliable for everyone.

Open source software has become ubiquitous and can be found in almost every codebase,[1] proprietary and open source alike, but sustaining open source projects and communities over the long term can be a challenge. Project leaders, maintainers, and contributors don't always have the time or experience to focus on sustainability. Using metrics is one way to help open source projects more quickly identify potential issues and areas where they can improve to make their projects more sustainable over the long term. Within the open source CHAOSS[a] project, metrics definitions and software exist to help people collect metrics for their open source projects, which has been described in more detail in previous articles. Goggins et al.[2] described how CHAOSS plays an integral role in the automation of key measures to make the state of open source readily observable using a CHAOSS tool called Augur[2] (see Figure 1). Gonzalez-Barahona et al.[3] took a slightly different approach by describing how people fitting several personas might use CHAOSS's GrimoireLab tools for data analysis of open source software.[3] Both of these articles are consistent with the approach that the CHAOSS project has taken in the past to provide tools and metrics to help gather data—but stopping short of providing advice about how to take action on the data and make improvements within open source projects. However, over the past couple of

Digital Object Identifier 10.1109/MC.2025.3530556
Date of current version: 29 May 2025

[a]CHAOSS (Community Health Analytics for Open Source Software) is an open source project under The Linux Foundation https://chaoss.community/.

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

EDITOR **DIRK RIEHLE**
Friedrich Alexander-University of Erlangen Nürnberg,
dirk.riehle@fau.de

FROM THE EDITOR

Welcome back to the "Open Source" column! This month, Dawn Foster takes a look at open source community metrics. Anyone interested in setting up their open source project for community collaboration rather than commercial exploitation is well advised to dig into the metrics that work by Dawn and her colleagues at the CHAOSS project unearthed. Happy collaborating, everyone, and be healthy and happy!—*Dirk Riehle*

FIGURE 1. OSSF Scorecard security assessment and general information for a repository. OSSF: Open Source Security Foundation. (Source: Image generated using Augur.[e])

years, providing advice has gradually started to change at the CHAOSS project (see Figure 2).

The CHAOSS project has learned that not everyone has the experience or skills required to know how to interpret metrics and use those learnings to make improvements within an open source project and community. This is why the CHAOSS project began working on a series of MIT-licensed Practitioner Guides.[b] The goal of these guides and this article is to help practitioners, who may not be experts in data analysis or open source, understand how to interpret the data about an open source project and develop insights that can help to improve the health of that project. Open Source Program Offices, project leads, community managers, maintainers, and anyone who wants to better understand project health and take action on what can be learned from metrics will benefit from this article and the CHAOSS practitioner guides.

Measuring project health is complex with a complex array of aspects to consider.[4] One of the best places to start isn't actually with the metrics but by spending some time understanding the overall goals for a project in question[c] and talking to the people who participate in and maintain that project.[5] One reason that the CHAOSS project has avoided providing specific advice in the past is because there is no one-size-fits-all approach to using metrics to measure open source project health. Every open source project is a little different, and metrics should always be interpreted with the needs of that project and its context taken into account (Goggins et al.[2]). This is why it's important to look at trends in the data over time and think about whether other factors might be influencing those trends (for example, conferences, release timing, and vacation season). However, it is still possible to provide advice about certain topics that are common across open source projects, like contributor sustainability, responsiveness, organizational participation, and security.

CONTRIBUTOR SUSTAINABILITY

Many open source projects struggle to find enough people to sustain them.[6] If there are too few contributors and maintainers to sustain a project, the risk that the project will fail increases,[7] which creates a variety of challenges for the users and other projects that depend on that project. With respect to open source project sustainability, the relationship between contributors and maintainers is important to understand, and the Contributor Sustainability Practitioner Guide[d] helps in this regard. For example, bringing on new contributors increases the maintainer load because those maintainers will need to provide feedback on and merge contributions from

[b]https://chaoss.community/about-chaoss-practitioner-guides/.

[c]https://chaoss.community/practitioner-guide-introduction/.

[d]https://chaoss.community/practitioner-guide-contributor-sustainability/.

[e]https://github.com/chaoss/augur.

those new contributions. Promoting existing established contributors into maintainer roles to handle that increased load is key because projects require enough maintainers to handle incoming requests.[8] By focusing on recruiting and retaining contributors and subsequently promoting those contributors to maintainers, projects can help proactively prevent sustainability crises later. In this regard, there are several CHAOSS metrics that can help to understand the contributor, and related maintainer, sustainability of a project.

By starting with the Contributor Absence Factor metric, the risk to the project if one or more key contributors/maintainers decide to leave can be assessed while also better understanding which people are making the most contributions. The Contributors metric looks broadly at who contributes to a project to help understand how many contributors are active along with how many have increasing or decreasing activity over time. Because there are so many ways to contribute to an open source project, the Types of Contributions metric can help build a more holistic understanding of where and how people are contributing.

> The CHAOSS project has learned that not everyone has the experience or skills required to know how to interpret metrics and use those learnings.

If it has been determined, via these metrics, that a project would benefit from improvements to contributor sustainability, there are a number of actions that can be taken. A good place to start is by looking for ways to reduce maintainer load through better contribution documentation. Projects may also benefit from taking a phased approach to recruiting new maintainers and reducing the scope that they will be responsible for (for example, a subproject or a portion of the codebase) and creating reviewer roles to help people build the skills they need as a maintainer while still allowing someone more experienced to oversee contributions before merging them. Maintainers can also use mentoring[9] and/or shadowing to more quickly teach people how to engage in maintainer work and help them learn to perform tasks that they'll need to do to become a maintainer. One reason to look at the Types of Contributions metric is that it can help to identify opportunities to promote people into maintainer roles to be responsible for activities that take up time from maintainers but that might be more effectively done by someone with more specialized expertise (for example, community management, marketing, and technical writing). Finally, having a written succession plan can also provide better sustainability if something happens to one or more of the existing maintainers.

RESPONSIVENESS

Responsiveness metrics[f] are an important part of assessing project health[8] since responsiveness is one of the most important factors in attracting newcomers[10] and retaining existing contributors to a project. New and existing contributors can become discouraged when they don't receive a timely and appropriate response to their contribution but can be encouraged when they get a quick and helpful resolution to their contribution. When projects are responsive, it can make people want to contribute more or continue contributing. Timely, thoughtful, and kind responses to contributors indicate that their work is appreciated.

By looking at Time to First Response, Time to Close, and Change Request Closure Ratio metrics together, a project can get a sense of whether contributors are getting a timely response and whether maintainers are keeping up with contributions by closing change requests (for example, pull requests/merge requests). For example, large numbers of open change requests can indicate that maintainers aren't particularly attentive to the project.[11] It can be tempting to put pressure on existing maintainers to respond more quickly,

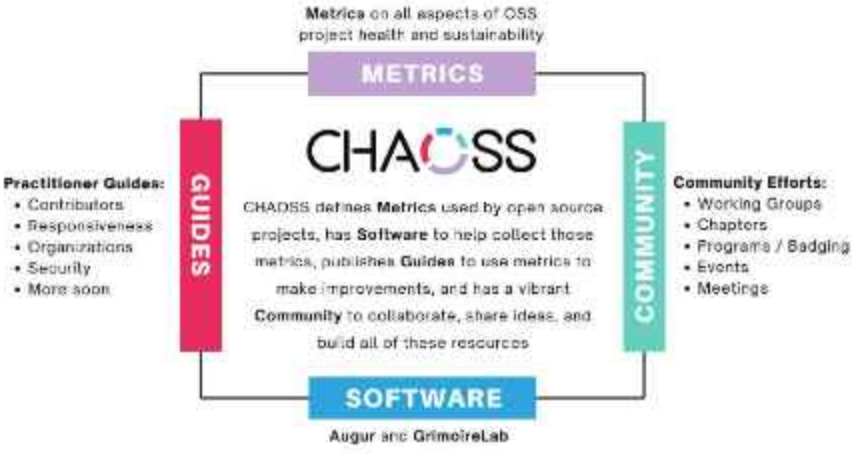


FIGURE 2. The CHAOSS community produces metrics, software, and guides to improve project health and sustainability.

[f] https://chaoss.community/practitioner-guide-responsiveness/.

but this rarely solves the long-term problem. It might result in short-term gains but can result in maintainer burnout if the underlying problems that are causing the lack of responsiveness are not resolved.

Like Contributor Sustainability, it can help to promote more contributors into maintainership roles so that more people can help respond, particularly into roles that free up time from code maintainers (for example, community management and documentation maintenance). Projects can also set clear expectations about when someone can expect a response, including delayed responses during busy times or holiday breaks. Using issue and pull request templates can further help people make better contributions the first time to reduce the reviewer load later.

ORGANIZATIONAL PARTICIPATION

Organizations can have a significant impact on the health and sustainability of an open source project,[g] especially when they come together under foundations to collaborate with other organizations.[12] On the one hand, organizations can help sustain open source projects by employing people to work on the open source projects that they use or by contributing other resources to those projects.[6] However, if all or most of the contributions are from the employees at a single organization, what happens when that organization is no longer willing or able to continue contributing at that same level?

From a metrics standpoint, a good starting point is looking at the Elephant Factor metric to determine how the work is distributed among multiple organizations along with the Organizational Diversity metric to look at which organizations are making contributions. Finally, it's also important to think about Organizational Influence metrics to understand which organizations have employees in leadership or other decision-making positions.

> There are several CHAOSS metrics that can help to understand the contributor, and related maintainer, sustainability of a project.

If a need to improve organizational diversity has been identified, how to accomplish this depends on whether or not some of the contributors work for the dominant organization. If an organization is dominant, a good first step is to improve transparency and make sure that open source project work is being done in the open. It can also help to use professional connections to other organizations that are using the project and discuss ways for them to contribute. If another organization is dominant, make sure that contributions from others are welcome since, unfortunately, some organizationally dominated projects aren't particularly welcoming to contributions from outside of the leading organization. If contributions are welcome, other companies can dedicate time from employees to work within the project to provide more organizational diversity and act as a catalyst to show other organizations that their employees are welcome.

SECURITY

Open source software packages can be found in almost all software, so the security[h] of open source projects can have wide-reaching implications for other projects, their users, and the broader software ecosystem. Security is only as strong as its weakest link, so the security of any software component is only as good as the security of its dependencies.[13]

Security is a complex topic, but there are a few key metrics that can be used as a starting point. First, the OpenSSF Best Practices Badging criteria create a good engineering foundation that incorporates basic security practices. Second, using outdated dependencies results in projects that are four times as likely to have security issues,[14] so using the Libyears metric can help to understand if dependencies are kept up to date. Third, the Release Frequency metric helps gauge whether security fixes and other updates are incorporated in a release in a timely manner so that users can benefit from those security updates.

To improve the security of an open source project, securing the code repository and creating a detailed security policy document, often in a SECURITY.md file, is a solid place to start. Using automated tools (for example, Dependabot) can help keep dependencies as up to date as possible. On an ongoing basis, projects are likely to find or receive reports about security vulnerabilities that will need to be fixed, so those security fixes should be clearly documented and released in a timely fashion.

Finally, using some of the OpenSSF tools and resources can help find areas within a project where security practices could be improved. The OpenSSF Scorecard (see Figure 1) can help identify areas to improve, and working through the OpenSSF Best Practices Badge criteria is a good way to continue to make security improvements for open source projects.

The CHAOSS Practitioner Guides provided the inspiration for this article because contributor

[g] https://chaoss.community/practitioner-guide-organizational-participation/.

[h] https://chaoss.community/practitioner-guide-security/.

sustainability, responsiveness, organizational participation, and security are all key topics as open source projects work to improve sustainability. The guides are MIT licensed and can be used as-is, or they can be forked from the CHAOSS Data Science Working Group repository[i] and modified to meet other needs.

Building sustainable open source projects over the long term can be a challenge. Project leaders, maintainers, and contributors are busy people who don't always have the time to focus on growing a community along with maintaining their software. Using metrics is one way to help identify potential issues and areas where a project can be improved to make it more sustainable over the long term. Metrics are best used if they aren't used once and never again. By monitoring the data over time, projects can understand trends that might indicate areas for improvement as well as see if those improvements are having the desired effect.

Being proactive about improving sustainability before it becomes a crisis can help make open source software more sustainable and reliable for everyone, but this requires work. The CHAOSS project is addressing these issues now with metrics, software, guides, and community collaboration, but ongoing work is needed from all of us to maintain and build on these resources while also using these resources to make open source projects more sustainable over time.

ACKNOWLEDGMENT

This work was funded by the Alfred P. Sloan Foundation (Grant 10384).

[i]https://github.com/chaoss/wg-data-science/tree/main/practitioner-guides.

REFERENCES

1. "Open source security and risk analysis report," Black Duck, Burlington, MA, USA, 2024. [Online]. Available: https://www.blackduck.com/resources/analyst-reports/open-source-security-risk-analysis.html
2. S. P. Goggins, M. Germonprez, and K. Lumbard, "Making open source project health transparent," *Computer*, vol. 54, no. 8, pp. 104–111, Aug. 2021, doi: 10.1109/MC.2021.3084015.
3. J. M. Gonzalez-Barahona, D. Izquierdo-Cortazar, and G. Robles, "Software development metrics with a purpose," *Computer*, vol. 55, no. 4, pp. 66–73, Apr. 2022, doi: 10.1109/MC.2022.3145680.
4. J. Linåker, E. Papatheocharous, and T. Olsson, "How to characterize the health of an Open Source Software project? A snowball literature review of an emerging practice," in *Proc. 18th Int. Symp. Open Collaboration*, 2022, pp. 1–12, doi: 10.1145/3555051.3555067.
5. A. Casari, J. Ferraioli, and J. Lovato, "Beyond the repository: Best practices for open source ecosystems researchers," *Queue*, vol. 21, no. 2, pp. 14–34, 2023, doi: 10.1145/3595879.
6. N. Eghbal, *Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure*. New York, NY, USA: Ford Foundation, 2016.
7. G. Avelino, E. Constantinou, M. T. Valente, and A. Serebrenik, "On the abandonment and survival of open source projects: An empirical investigation," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 1–12, doi: 10.1109/ESEM.2019.8870181.
8. N. Eghbal, *Working in Public: The Making and Maintenance of Open Source Software*. San Francisco, CA, USA: Stripe Press, 2020.
9. F. Fagerholm, A. S. Guinea, J. Münch, and J. Borenstein, "The role of mentoring and project characteristics for onboarding in open source software projects," in *Proc. 8th ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas.*, 2014, pp. 1–10, doi: 10.1145/2652524.2652540.
10. F. Fronchetti, I. Wiese, G. Pinto, and I. Steinmacher, "What attracts newcomers to onboard on OSS projects? TL;DR: Popularity," in *Proc. IFIP Int. Conf. Open Source Syst.*, 2019, pp. 91–103.
11. L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in GitHub: Transparency and collaboration in an open software repository," in *Proc. ACM 2012 Conf. Comput. Supported Cooperative Work*, 2012, pp. 1277–1286.
12. D. Riehle, "The innovations of open source," *Computer*, vol. 52, no. 4, pp. 59–63, Apr. 2019, doi: 10.1109/MC.2019.2898163.
13. N. Imtiaz, A. Khanom, and L. Williams, "Open or sneaky? Fast or slow? Light or heavy?: Investigating security releases of open source packages," *IEEE Trans. Softw. Eng.*, vol. 49, no. 4, pp. 1540–1560, Apr. 2023, doi: 10.1109/TSE.2022.3181010.
14. J. Cox, E. Bouwers, M. Van Eekelen, and J. Visser, "Measuring dependency freshness in software systems," in *Proc. IEEE/ACM 37th IEEE Int. Conf. Softw. Eng.*, vol. 2, Piscataway, NJ, USA: IEEE Press, 2015, pp. 109–118, doi: 10.1109/ICSE.2015.140.

DAWN FOSTER is the director of data science at CHAOSS, London, U.K. Contact her at dawn@dawnfoster.com.

EDITOR **CHRISTOF EBERT**
Vector Consulting Services;
christof.ebert@vector.com

Innovation Turns Smart and Green

Christof Ebert, Vector Consulting Services

The 2025 Consumer Electronics Show boasted innovations across industries and spanning the globe. Learn here about the major trends, what companies we saw and talked to, expert opinions, and what might be in it for your own advancement.

The Consumer Electronics Show (CES) is the world's largest technology trade fair. Each year in January, it takes place in Las Vegas. This year, CES had over 4,500 exhibitors with industry giants such as Nvidia, Samsung, and Siemens but also 1,400 start-ups. There were 141,000 visitors, with an increasing share from Asia. One third of exhibitors came from China, with a clear growing trend. Since it was founded in 1967 in New York with 250 exhibitors and 17,500 visitors, the technology trade fair has steadily grown. Today it is the leading trade fair for all kinds of "consumer electronics," from smart homes, ecology, and medical technology to data centers and robots, industry automation, vehicles, and mobility.

MAJOR TRENDS AT CES 2025

Three major themes and trends can be grouped, which we will dive into in this article:

- consumer, health, and home
- industry and automation
- automotive and mobility.

Artificial intelligence (AI) was clearly leading the innovation headlines and expert talks at CES, ranging from software development copilots to smart home and health gadgets with AI up to industry robots with latest AI technology (Figure 1). The high proportion of Chinese companies and visitors was surprising. Almost a third of the exhibitors came from China. Europe was lagging, as were, in general, automotive companies, which until last year almost dominated CES with their innovations. The current economic decline was tangible, even from just measuring the loudness in the halls. Consumer goods

Digital Object Identifier 10.1109/MC.2025.3540533
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

FROM THE EDITOR

The 2025 Consumer Electronics Show in Las Vegas is the global innovation hub. This overview shows key trends covering consumer, home, and health; industry and automation; and automotive and mobility. Innovation turns smarter and greener. Artificial intelligence is at its peak hype, being present across the halls. Convergence of once separate domains and technologies is another megatrend. Yet our expert interviews also emphasize uncertainties related to the economy, cybercrime, and global trade restrictions. —*Christof Ebert*

were colorful and loud, while the automotive booths in the west hall of the Las Vegas Convention Center were rather quiet and sober.

CONSUMER, HEALTH, AND HOME: AI AND SUSTAINABILITY

Consumer electronics, the initial starting point of CES, is dominated by AI and sustainability. AI is increasingly integrated into everyday consumer products. Products range from AI-powered devices and smart homes to health-monitoring wearables, laundry machines, and a huge amount of home entertainment, such as the latest high-definition TV screens. Augmented reality is converging with AI. Innovations such as AI-enhanced smart glasses are capable of real-time translation and augmented reality overlays. Consumer experiences become almost daily more immersive and interactive. The Korean companies Samsung and LG introduced AI-driven home appliances that adapt to user habits, optimize energy efficiency, and enhance convenience.

In healthcare, Internet of Things (IoT)-enabled wearables are used for everyday health surveillance from fitness tracking to real-time health monitoring and early disease detection. Medical devices highlight, for instance, smart insulin pumps and AI-assisted diagnostic tools. AI-assisted diagnostic tools now analyze voice patterns, eye movements, and skin conditions to detect early signs of diseases such as Parkinson's and cardiovascular disorders. Remote patient-monitoring solutions are also gaining traction, reducing the need for frequent hospital visits. AI-powered medical IoT devices allow doctors to track patients' vitals in real time, improving healthcare accessibility, especially in remote areas. However, concerns over patient data security and regulatory approval processes continue to shape the adoption of these technologies.

Connectivity is everywhere with software updates and IoT. Cybersecurity thus is a growing issue in consumer electronics, especially as AI becomes more prevalent. With smart devices continuously collecting data, experts warned about potential risks of data misuse. The good news is that consumers increasingly demand transparency in how their personal data are handled, pushing manufacturers to implement stricter security protocols and encryption measures. The integration of AI and IoT in healthcare raises ethical and regulatory concerns, with experts emphasizing the need for stringent data security measures to protect patient privacy.

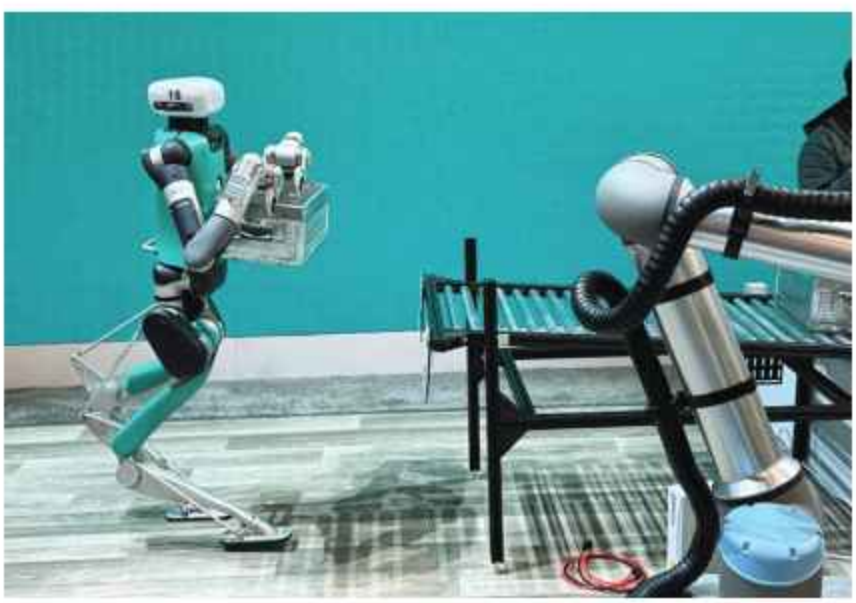


FIGURE 1. AI shapes the technology landscape.

INDUSTRY AND AUTOMATION: CONNECTIVITY AND CONVERGENCE

Connectivity drives the entire industry domain with IoT highlighting advancements in industrial automation, healthcare, and smart city technologies. Edge computing solutions and edge AI will reduce latency and improve efficiency by processing data closer to the source. This is particularly relevant in manufacturing, where real-time data analytics enhance predictive maintenance and minimize downtime.

IoT, once thought to bridge industry domains with the many organically grown proprietary protocols and stacks, seems to rather propagate old habits. Interoperability among

IoT devices remains a persistent challenge. Latest robot technology is visible at many booths, including a tasty AI-driven ice cream machine at one of the booths, yet operate within proprietary ecosystems. Seamless integration across different platforms, while promoted in flyers and headlines, remains wishful thinking and slows down innovation. Standardization efforts have been underway for decades, but the lack of universal protocols continues to hinder widespread IoT adoption.

With billions of connected edge devices, machines, and entire supply chains, vulnerabilities in industrial IoT systems can lead to significant disruptions. To mitigate, companies are investing in AI-driven security solutions capable of identifying and mitigating cyberthreats in real time, as we learned from a Siemens executive. An increasing amount of governance and compliance even punishes those companies who try to obey rules and behave correctly.

AUTOMOTIVE AND MOBILITY: SMART TRANSPORTATION

The CES 2025 west hall emphasized rapid advancements in electric vehicles (EVs), autonomous driving, and AI-powered mobility solutions (Figure 2). Because of the current uncertainties, many carmakers that have traditionally had huge booths were just absent. Instead, their engineering and management teams just wandered the halls and held major meetings in adjacent hotels.

Hyundai had one of the biggest booths, packed with augmented reality, AI, and the latest car technology. Deere promoted the latest harvesting machines, which converge classic agriculture with backbone software systems to integrate with enterprise software systems and maintenance. Automakers unveiled next-generation EVs with improved battery efficiency, reduced charging times, and extended range capabilities. BMW, Hyundai, and Tesla showed concept cars with solid-state batteries that are expected to revolutionize energy storage in the mobility industry. AI-driven in-car assistants and augmented reality dashboards are major trends, enhancing driver experience and safety. They offer real-time traffic analysis, personalized entertainment, and identification of hazards. Potential overreliance on AI in driving decision making emerges as a new risk along with cybersecurity, as we learned in our interviews with automotive experts.

Autonomous driving, last year still a major hype, is now in the valley of disillusion. The Japanese manufacturer Suzuki was represented at CES for the first time. Among other things, it presented an autonomous electric platform in a small car format, which is intended to counteract the shortage of drivers in logistics. Companies demonstrated AI-powered self-driving systems that rely on enhanced sensor fusion and machine learning algorithms. These systems promise safer and more efficient transportation but face regulatory and ethical challenges. The deployment of fully autonomous vehicles remains limited due to unresolved legal and liability concerns.[1]

WHERE DO WE GO FROM HERE?

CES 2025 highlighted several challenges facing the technology industry. Regulatory compliance has become a central issue, with governments worldwide implementing stricter AI and data privacy laws. These evolving regulations, mostly driven by Europe but fast adopted in countries such as Korea and Japan, impact how companies develop and deploy AI-driven solutions, demanding greater transparency and adherence to mushrooming AI standards.

Sustainability must connect ecology, economy, and social standards. Manufacturers thus show sustainability end to end. Examples include eco-friendly materials such as textiles, energy-efficient products, and longer-lasting batteries. Companies such as Siemens, Sony, and Dell showcase devices made from recycled materials. The smartphone industry promises to extend software

FIGURE 2. CES remains the place to be for the latest trends.

GUIDE TO GROW

TAKEAWAYS FROM INDUSTRY TRENDS

- *AI everywhere:* From robots and automation to smart home and automotive systems, artificial intelligence (AI)–powered personalization and automation dominate technology evolution. Improved AI models allow augmented reality, content creation, and numerous personal assistants.
- *Sustainability:* Companies emphasize energy-efficient devices, recycled materials, and extended lifespans to reduce e-waste.
- *Autonomous and electric mobility:* Although impacted by the current economic decline, electric vehicles with novel battery systems, AI-driven driver assistance, and enhanced charging infrastructure will further stimulate mobility.
- *Healthcare:* AI-assisted diagnostics, wearable health monitors, and remote patient monitoring will further grow digital health markets.
- *Cybersecurity:* Increased AI adoption raises concerns over deepfakes, data privacy, and potential misuse of generative AI.
- *Regulatory uncertainty:* Growing standards and laws with many regional variants challenge businesses to comply especially. A lack of universal Internet of Things standards makes seamless device integration difficult.
- *AI divide:* High-tech solutions need to be more inclusive and user-friendly to reach broader audiences across countries and cultures.

TRANSFER QUESTIONS

- How can you integrate AI-driven personalization into our products to enhance user experience and efficiency?
- What strategies from CES innovations in cybersecurity and data privacy should you adopt to strengthen the security of our applications?
- How can you align software development with sustainability goals, such as optimizing energy efficiency or extending product lifecycles?

updates for less electronic waste. However, the push for sustainability faces challenges, particularly in balancing technological advancements with environmental impacts (see Guide to Grow).

The widespread use of AI and connected devices further grows energy demands such as data centers and training of language models. According to experts we talked to, the industry must address electronic waste, responsible sourcing of raw materials, and energy-efficient manufacturing processes to achieve long-term sustainability goals. European initiatives with focuses on green energy were esteemed. Bosch, for instance, devoted its booth almost entirely to eBike and sustainable energy usage, away from the once omnipresent automotive gadgets. Yet, experts are also concerned when looking to China and the United States, who are rather increasing their ecologic footprint with an increasing number of data centers fueled by carbon energy.

The digital divide is further growing, and we already see an AI divide. User adoption and accessibility also emerged as significant themes. High-tech solutions attract early adopters, but widespread adoption depends on price and usability. With AI advancements, ethical concerns persist regarding the misuse of generative AI, particularly in creating deepfake content and misinformation. Companies are working on developing AI-generated content verification tools to combat these risks.

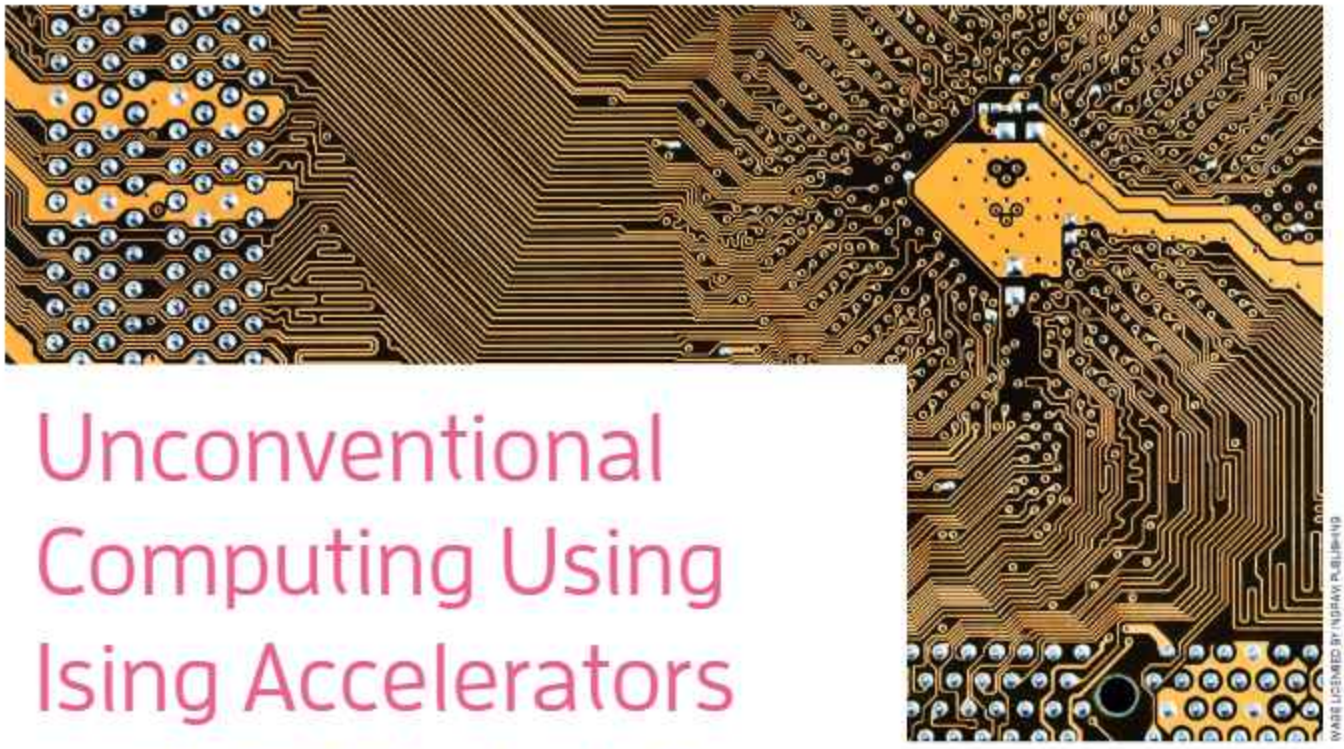
CES 2025 showcased a future driven by AI, connectivity, and sustainability. Companies and countries must collaborate to address these issues, ensuring that innovation progresses responsibly and bridging the many technology divides. The famous politician Winston Churchill remarked that a pessimist sees the difficulty in every opportunity, while an optimist sees the opportunity in every difficulty. As industries continue to evolve, CES is the platform for shaping the technological landscape. The next CES is scheduled for 6–9 January 2026.

REFERENCE

1. C. Ebert and M. Weyrich, "Validation of autonomous systems," *IEEE Softw.*, vol. 36, no. 5, pp. 15–23, Sep./Oct. 2019, doi: 10.1109/MS.2019.2921037.

CHRISTOF EBERT is the managing director of Vector Consulting Services, 70499 Stuttgart, Germany. He is a Senior Member of IEEE. Contact him at https://www.linkedin.com/in/christofebert or christof.ebert@vector.com.

EDITOR CONRAD JAMES
Sandia National Laboratories, USA;
cdjame@sandia.gov



Unconventional Computing Using Ising Accelerators

Jaydeep P. Kulkarni, Siddhartha Raman Sundara Raman, Shanshan Xie, and Chieh-Pu Lo, The University of Texas at Austin

Recent advancements in tackling complex computational problems have focused on drawing inspiration from nature. One notable example is the use of Ising machines to solve nondeterministic polynomial-time hard combinatorial optimization problems.

Combinatorial optimization problems (COPs) play a critical role in various real-world applications, such as semiconductor supply chain management,[1] financial index tracking,[2] and optimizing mRNA sequences[3,4] for COVID-19 vaccines. These problems are computationally nondeterministic polynomial-time hard, making brute-force approaches highly resource-intensive and impractical for scaling to larger problem sizes. A more efficient approach involves mapping COPs onto nature-inspired Ising models.[5] These models emulate the spin dynamics of ferromagnets, wherein spins naturally align to the lowest ensemble energy state, thereby representing the optimal solution to the COP.[6]

HAMILTONIAN COMPUTE

An Ising model could be mathematically mapped to a Hamiltonian function, which commonly represents the total energy of a physical system. The optimal COP solution is found by iteratively minimizing the total energy of the Hamiltonian function, defined by pairwise spin couplings

$$H = -\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \quad (1)$$

where σ_i is the target spin, σ_j are neighboring spins, h_i is the external field of target spin, J_{ij} represents interaction coefficients (ICs) between σ_i and σ_j, and i, j denote node

Digital Object Identifier 10.1109/MC.2025.3544798
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

pairs in an Ising network graph (see Figure 1). Minimizing H involves updating each spin iteratively based on its interactions with neighbors

$$H_s = \sum -J_{ij} * \sigma_j - h_i. \quad (2)$$

The probable spin update is carried out based on the sign of H_s[7]

$$\sigma = \begin{cases} -1, & \text{if } H_s > 0 \\ +1, & H_s < 0 \\ +1/-1 & H_s = 0. \end{cases} \quad (3)$$

computation method used in optimizing COPs, have led to different implementations of Ising machines. These are broadly classified as a physical Ising machine or iterative Ising machine.[5]

Ideally, Ising machines do not require iterative updates, as they reach the minimal energy state in an extremely parallel manner, similar to physical Ising machines. However, depending on the implementation, an annealing mechanism may be applied iteratively to find the optimal COP solution.

These models emulate the spin dynamics of ferromagnets, wherein spins naturally align to the lowest ensemble energy state, thereby representing the optimal solution to the COP.

However, local spin updates can trap H in a local minimum. Simulated annealing is then performed, and uses probabilistic spin flips to reach the global minimum (see Figure 2).

ISING MACHINES AND COPs

Ising machines leverage the concepts of Ising models to efficiently represent spins and ICs and solve optimization problems. In these machines, variables and constants in COPs are encoded as spins and ICs, respectively. The Hamiltonian energy function is used as a heuristic to identify optimal solutions to COPs.

Architecturally, variations in how spins and ICs are represented, and the

PHYSICAL ISING MACHINES

Prior hardware implementations of solving Ising models has focused on physics-based behavior to emulate spin dynamics using superconducting quantum bits (quantum annealers), lasers (optical annealers), or coupled oscillators. The D-Wave quantum annealer[8] uses quantum bits (qubits) to encode states of 0, 1, or both simultaneously, and qubit–qubit coupling to encode interaction coefficients. To reduce thermal noise and enhance qubit coherence time, the quantum annealer operates at cryogenic temperatures (15 mK) and requires high cooling power (25 kW), making it a cost-inefficient approach. Similarly, optical Ising machines[9] determine spin energy minima primarily via a combination of binary phase modulators and feedback mechanisms. These encode spin/interaction coefficients using light polarization/phase. However, their reliance on bulky components and specialized fabrication requirement to integrate photonic chips hinder scalability for large COPs. CMOS-coupled oscillators[10] offer room-temperature operation and leverage semiconductor miniaturization to minimize energy. Their steady-state phase dynamics determine the Ising model's ground state. However, they suffer from high dynamic power consumption due to constant node toggling and are prone to process variations and spurious couplings, leading to ground state fluctuations.

ITERATIVE ISING MACHINES

An alternative approach abstracts spin dynamics into a Hamiltonian model, iteratively updating spins to minimize the Hamiltonian, which is then implemented in CMOS hardware as an Ising machine. Compared to physical Ising machines, CMOS-based Ising machines offer a scalable, cost-efficient, and energy-efficient alternative.

For instance, Yamaoka et al.[11] introduced a CMOS-based digital annealer that simulates spin interactions to reach the lowest energy state. This design employs dedicated arithmetic circuits (XORs, switches, majority voting) for Hamiltonian computation and

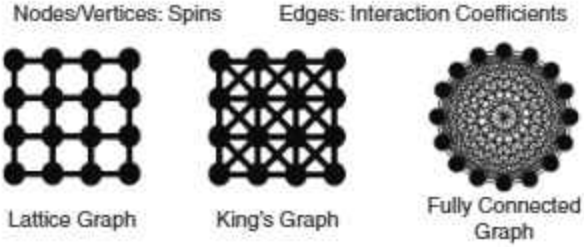


FIGURE 1. Commonly used Ising network graphs include lattice, king's, and fully connected graphs. Each graph presents varying levels of complexity for solving different problems.

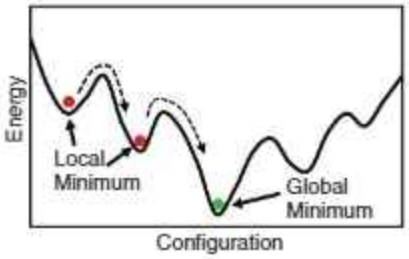


FIGURE 2. Illustration of Hamiltonian energy with possible trap from local minimum and global minimum energy as an optimal COP solution.

uses standard static random access memory (SRAM) bit cells to store spins and interaction coefficients. However, this iterative process involves frequent memory read-and-write operations. On conventional von Neumann architectures [Figure 3(a)], such operations lead to excessive off-chip memory accesses, causing significant performance and energy overheads.

To mitigate these issues, digital compute/near-memory designs have been proposed, enabling massively parallel computations near-memory bit cells. For example, the digital annealing processor (AP)[7] employs a compute-in-memory (CIM) spin operator and register-based spins to enhance energy efficiency and reduce annealing time [see Figure 3(b)]. By eliminating SRAM read/write operations for adjacent spin values and local simulated annealing triggers, it improves computation performance. However, it requires custom digital circuits (transmission-gate, XNORs, and full-adders) per column and four SRAM bit cells, with Hamiltonian computation delayed by partial sum propagation.

Another approach uses a fully connected annealer based on the stochastic cellular automata algorithm,[12] featuring innovations like delta-driven simultaneous spin updates and all-to-all interactions. However, its compute/near-memory design [see Figure 3(c)] relies on dedicated digital logic, reducing memory bit cell density, and increasing energy costs for data movement from SRAM to the annealer. Similarly, a 144-Kb AP chip[13] demonstrated a 9 × 16-k spin system using flip-flop–based structures to connect chips via interchip interfaces. While effective, such custom circuits increase area overhead, degrade memory density, and raise hardware costs. Although these designs achieve significant speedups and lower energy consumption compared to CPUs, they rely on custom digital arithmetic logic integrated near memory segments, resulting in substantial area overheads and a 3−10× reduction in memory density. These limitations significantly increase hardware costs and restrict the adoption of compute/near-memory-based Ising architectures in modern system-on-chips.

Ising machines leverage the concepts of Ising models to efficiently represent spins and ICs and solve optimization problems.

To address these challenges, SACHI[5] repurposes the CPU's L1 cache with minimal near-memory logic for Hamiltonian computation, significantly reducing hardware costs. It introduces stationarity in in/near-memory digital Ising machines and argues against spin-stationary computation for better performance–energy tradeoffs. Instead, it proposes a reuse-aware, mixed-stationary approach to minimize data movement costs and improve energy efficiency for Ising computation. Another approach is where Ising-based CIM [Figure 4(a)][6] efficiently maps Hamiltonian computations onto memory wordline and bitline circuitry, enabling analog bitline operations with minimal peripheral circuit modifications. It performs one-bit spin updates per Hamiltonian iteration with a local write-back mechanism using existing sense amplifiers in the memory array. It acts as a one-bit comparator and eliminates the need for complex data converters. To enhance the complexity of Ising-CIM, a ping-pong CIM-based Ising machine [Figure 4(b)][14] achieves an all-to-all connection design. It restricts single-bit spin data movement within the memory array while keeping interaction coefficients as stationary storage, significantly reducing memory access overhead. The architecture offers flexibility in configuring multibitwidth ICs and various Ising network graphs, enabling the solution of a wide range of complex COPs.

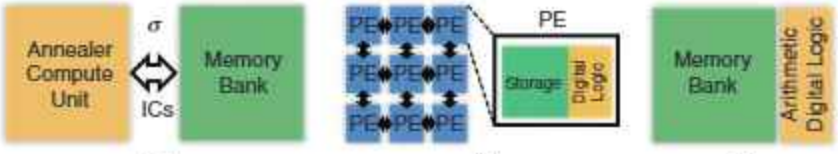


FIGURE 3. Architectures of iterative Ising machine. (a) Conventional von Neumann architecture, (b) multiple process-element (PE) architecture, and (c) compute/near-memory architecture.

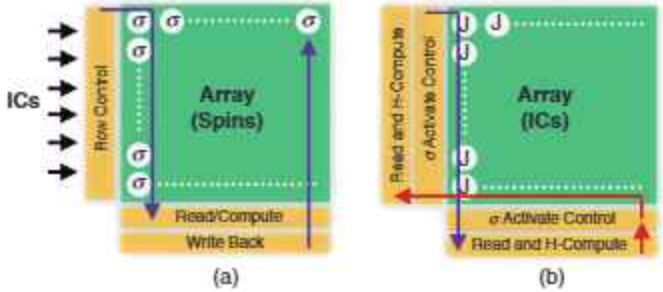


FIGURE 4. Different approach of storing in CIM architecture. (a) Spins are stored and updated within memory array. (b) ICs are stored and stationary within memory array.

To conclude, recent Ising compute architectures, ranging from physical-based to iterative CMOS-based designs, offer a promising and efficient approach to solving complex COP problems. With the rapid growth of accelerators in this artificial intelligence-driven era, more efficient and reliable Ising machine accelerators have been introduced, paving the way for unconventional computing paradigms. C

REFERENCES

1. K. G. Kempf, "Control-oriented approaches to supply chain management in semiconductor manufacturing," in *Proc. Amer. Control Conf.*, Piscataway, NJ, USA: IEEE Press, 2004, vol. 5, pp. 4563–4576, doi: 10.23919/ACC.2004.1384031.
2. Y. Crama, "Combinatorial optimization models for production scheduling in automated manufacturing systems," *Eur. J. Oper. Res.*, vol. 99, no. 1, pp. 136–153, 1997, doi: 10.1016/S0377-2217(96)00388-8.
3. K. Leppek et al., "Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics," *Nature Commun.*, vol. 13, no. 1, 2022, Art.no. 1536, doi: 10.1038/s41467-022-28776-w.
4. N. Pardi et al., "mRNA vaccines—A new era in vaccinology," *Nature Rev. Drug Discov.*, vol. 17, no. 4, pp. 261–279, 2018, doi: 10.1038/nrd.2017.243.
5. S. R. Sundara Raman, L. K. John, and J. P. Kulkarni, "SACHI: A stationarity-aware, all-digital, near-memory, Ising architecture," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archi. (HPCA)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 719–731, doi: 10.1109/HPCA57654.2024.00061.
6. S. Xie, S. R. S. Raman, C. Ni, M. Wang, M. Yang, and J. P. Kulkarni, "Ising-CIM: A reconfigurable and scalable compute within memory analog Ising accelerator for solving combinatorial optimization problems," *IEEE J. Solid-State Circuits*, vol. 57, no. 11, pp. 3453–3465, Nov. 2022, doi: 10.1109/JSSC.2022.3176610.
7. Y. Su, H. Kim, and B. Kim, "CIM-spin: A scalable CMOS annealing processor with digital in-memory spin operators and register spins for combinatorial optimization problems," *IEEE J. Solid-State Circuits*, vol. 57, no. 7, pp. 2263–2273, Jul. 2022, doi: 10.1109/JSSC.2021.3139901.
8. M. W. Johnson et al., "Quantum annealing with manufactured spins," *Nature*, vol. 473, no. 7346, pp. 194–198, 2011, doi: 10.1038/nature10012.
9. D. Pierangeli, G. Marcucci, and C. Conti, "Large-scale photonic Ising machine by spatial light modulation," *Phys. Rev. Lett.*, vol. 122, no. 21, 2019, Art.no. 213902, doi: 10.1103/PhysRevLett.122.213902.
10. T. Wang, L. Wu, and J. Roychowdhury, "New computational results and hardware prototypes for oscillator-based Ising machines," in *Proc. 56th Annu. Des. Autom. Conf.*, 2019, pp. 1–2.
11. M. Yamaoka et al., "A 20k-spin Ising chip to solve combinatorial optimization problems with CMOS annealing," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 303–309, Jan. 2016, doi: 10.1109/JSSC.2015.2498601.
12. K. Yamamoto et al., "STATICA: A 512-spin 0.25 M-weight annealing processor with an all-spin-updates-at-once architecture for combinatorial optimization with complete spin–spin interactions," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 165–178, Jan. 2020, doi: 10.1109/JSSC.2020.3027702.
13. T. Takemoto et al., "4.6 A 144Kb annealing system composed of 9 × 16Kb annealing processor chips with scalable chip-to-chip connections for large-scale combinatorial optimization problems," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 64–66.
14. C.-P. Lo et al., "Data movement-aware, ping-pong Ising machine supporting full connectivity and variable bitwidths," in *Proc. IEEE Eur. Solid-State Electron. Res. Conf. (ESSERC)*, 2024, pp. 721–724, doi: 10.1109/ESSERC62670.2024.10719421.

JAYDEEP P. KULKARNI is an associate professor in the Chandra Department of Electrical and Computer Engineering and a Fellow of Silicon Labs Endowed Chair in electrical engineering at the University of Texas at Austin, Austin, TX 78712 USA. Contact him at jaydeep@austin.utexas.edu.

SIDDHARTHA RAMAN SUNDARA RAMAN is with the University of Texas at Austin, Austin, TX 78712 USA. Contact him at s.siddhartharaman@utexas.edu.

SHANSHAN XIE is with the University of Texas at Austin, Austin, TX 78712 USA. Contact her at sxie@utexas.edu.

CHIEH-PU LO is a Ph.D. student in electrical and computer engineering at the University of Texas at Austin, Austin, TX 78712 USA. Contact him at kcplo@utexas.edu.

IEEE Computer Society Volunteer Service Awards

Nominations accepted throughout the year.

T.Michael Elliott Distinguished Service Certificate

Highest service award in recognition for distinguished service to the IEEE Computer Society at a level of dedication rarely demonstrated. i.e., initiating a Society program or conference, continuing officership, or long-term and active service on Society committees.

Meritorious Service Certificate

Second highest level service certificate for meritorious service to an IEEE Computer Society-sponsored activity. i.e., significant as an editorship, committee, Computer Society officer, or conference general or program chair.

Outstanding Contribution Certificate

Third highest level service certificate for a specific achievement of major value to the IEEE Computer Society, i.e., launching a major conference series, a specific publication, standards and model curricula.

Continuous Service Certificate

Recognize and encourage ongoing involvement of volunteers in IEEE Computer Society programs. The initial certificate may be awarded after three years of continuous service.

Certificate of Appreciation

Areas of contribution would include service with a conference organizing or program committee. May be given to subcommittee members in lieu of a letter of appreciation.

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3562021

Nominations

Submit your nomination at
bit.ly/computersocietyawards

Contact us at
awards@computer.org

The Blue Pill Attack as a Wake-Up Call

Bob Maley, Black Kite

Joanna F. DeFranco, The Pennsylvania State University

Prioritizing Internet of Things security today is crucial to protecting data, lives, economies, and the stability of national infrastructure. Immediate action is necessary to mitigate emerging threats and ensure long-term resilience.

In a world where interconnected devices dominate our personal and professional lives, the recent "Blue Pill" attack reminded us of the vulnerabilities inherent in the Internet of Things (IoT).[1] The Blue Pill is a reference from the 1999 movie *The Matrix*. The "red pill versus blue pill" choice was to take a pill to either reveal the truth (red pill) or be unaware of being controlled (blue pill). This Blue Pill is now considered an advanced cyberattack. This idea was developed by cybersecurity expert Joanna Rutkowska, where a malicious virtual machine gains control of the victim's machine.[2] Affecting more than 35 million routers globally, this attack exploited flaws in consumer-grade devices, allowing attackers to hijack networks and potentially launch more significant, more devastating assaults. This attack can cause network degradation for home users as well as disruption from operations, cybercrime, and reputation damage.[1]

While many dismissed this incident as a consumer issue, the implications are far more profound. IoT devices form the backbone of modern infrastructure, connecting everything from home networks to critical national systems. The Blue Pill attack highlights the urgent need for a comprehensive approach to IoT security to protect not just individual users but also the stability of national infrastructure.

THE PERVASIVENESS OF IOT: FROM HOMES TO NATIONAL INFRASTRUCTURE

Consumer IoT devices

The proliferation of IoT devices in homes has revolutionized convenience. Smart thermostats adjust temperatures based on our preferences, connected security cameras provide peace of mind, and voice assistants simplify daily tasks. However, these conveniences often come with significant security tradeoffs.

Many consumer IoT devices are shipped with default passwords, rarely receive firmware updates, and are designed with minimal security considerations. These vulnerabilities make them easy targets for attackers. Worse, compromised devices can act as gateways, granting attackers access to more extensive networks, including corporate systems, when employees work remotely.

Digital Object Identifier 10.1109/MC.2025.3547171
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

The threat extends beyond individual households, underscoring the importance of securing even the most mundane devices.

Industrial IoT

The Fourth Industrial Revolution brought digital transformation to industries not only through automation but data exchange. The industrial sector's reliance on IoT devices has revolutionized operations and introduced new vulnerabilities. Industrial IoT (IIoT) devices are integral to energy grids, transportation networks, and health-care systems. Smart sensors monitor power distribution, traffic management systems optimize flow, and connected medical devices enable better patient care and management of chronic conditions. However, their critical nature makes them prime targets for attackers.

The 2021 Colonial Pipeline ransomware attack highlighted the importance of cybersecurity, critical infrastructure vulnerabilities, and the growing threats.[3] This attack was a stark example of what's at stake, as it disrupted fuel supplies across the eastern United States, causing widespread panic and economic damage. Such incidents reveal the fragility of critical systems when IIoT devices are compromised. The risks extend beyond operational downtime to encompass public safety and national security.

The IoT ecosystem

The interconnectedness of the IoT ecosystem magnifies the risks. A compromised consumer router, for instance, could serve as a launchpad for attacks on corporate or industrial networks. Similarly, unsecured IIoT devices can be entry points for attackers targeting critical infrastructure. The cascading effects of such vulnerabilities underscore the need for a holistic approach to IoT security: one that considers the entire ecosystem rather than individual devices.

THE GROWING THREAT LANDSCAPE

Large-scale botnets and ransomware synergy

One of the most immediate and visible threats posed by insecure IoT devices is their potential to form botnets. Botnets are networks of compromised devices controlled by attackers, often used to launch distributed denial-of-service attacks. The infamous Mirai botnet, which weaponized IoT devices in 2016,[4] demonstrated the scale of damage such attacks can inflict; taking down major websites and services globally demonstrated the scale of damage such attacks can inflict.

The recent Black Kite Ransomware Report, *State of Ransomware 2024: A Year of Surges and Shuffling*,[5] suggests that ransomware groups are evolving strategies, targeting industries with high economic value and weak security defenses. If attackers can compromise millions of routers, as seen in the Blue Pill attack, they could potentially sell access to ransomware groups, creating an entry point for broader cybercriminal operations. This access-as-a-service model could enable ransomware actors to conduct multistage attacks: first leveraging IoT devices to steal credentials or gain footholds and later deploying ransomware for financial extortion. Figure 1 illustrates how personal consumer devices can become a part of a botnet that contributes to a ransomware attack. This is shown by the possible infection methods (that is, phishing, malware, vulnerabilities), connecting to the botnet and helping to distribute ransomware or launch other attacks.

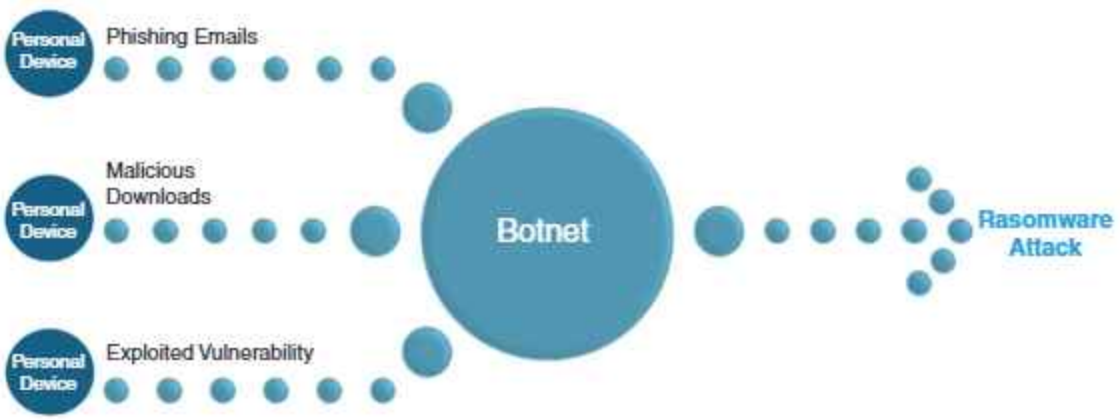


FIGURE 1. A ransomware scenario where personal devices contribute to the spread, execution, or success of ransomware campaigns.

Data exfiltration and espionage

IoT devices are treasure troves of data. Compromised devices can leak sensitive information, from personal details to corporate secrets. The Black Kite report[5] highlights that modern ransomware actors increasingly use data theft before encryption, meaning IoT devices could become entry points for espionage or large-scale data breaches.

The implications are profound. Stolen personal data can fuel identity theft and fraud, while leaked corporate secrets undermine competitive advantages or compromise national security. Organizations must recognize the importance of securing IoT devices like traditional IT systems.

The implications are profound. Stolen personal data can fuel identity theft and fraud, while leaked corporate secrets undermine competitive advantages or compromise national security.

Cyberphysical risks and the national impact

The convergence of digital and physical systems introduces unique risks. When IoT devices are exploited, the consequences often extend beyond cyberspace to the physical world. Examples include disrupting power grids, tampering with medical devices, and manipulating traffic systems.

According to the Black Kite report,[5] the manufacturing, finance, and health-care sectors remain top ransomware targets due to their reliance on IoT and IT systems. The consequences could be life-threatening if a compromised router leads to a ransomware attack on a hospital's network. Similarly, a compromised industrial system could disrupt supply chains or cripple transportation networks.

The rising trend of quick-succession ransomware attacks further amplifies these risks. The report found that multiple ransomware groups hit multiple companies in rapid succession, suggesting that cybercriminals share intelligence on vulnerable targets. A Blue Pill-style attack could expose IoT devices to similar cascading threats, where multiple attackers exploit the same vulnerability before it is patched.

The Blue Pill attack was more than just a wake-up call; it was a stark reminder of the vulnerabilities embedded in our increasingly connected world. From consumer devices to critical infrastructure, IoT security is no longer optional, it is a necessity.

The pervasiveness of IoT demands a unified approach to security. Governments, businesses, and individuals must work together to address vulnerabilities and protect interconnected systems. The Black Kite Ransomware Report[5] underscores the real-world consequences of cyberthreats, and the Blue Pill attack signals the next evolution of how cybercriminals can exploit IoT weaknesses.

With personal devices, vulnerabilities can be reduced by simply strengthening password security and updating default home router passwords. For organizations, it's important to reduce risk with mitigating exploitable vulnerabilities (that is, gaining unauthorized network access), credential leakage (that is, gaining access using leaked username-passwords), server misconfigurations (that is, improperly configured mail servers), and protecting open access points (that is, unprotected IoT devices, such as security cameras).

By prioritizing IoT security today, we can safeguard not just data but also lives, economies, and the stability of national infrastructure. The time to act is now.

REFERENCES

1. D. Winder, "Are you already in the matrix—35 million devices under blue pill attack," *Forbes*, Nov. 27, 2024. Accessed: Feb. 13, 2025. [Online]. Available: https://www.forbes.com/sites/daveywinder/2024/11/27/is-your-router-in-the-matrix-35-million-devices-under-blue-pill-attack/
2. "Blue pill attack." NordVPN. Accessed: Feb. 14, 2025. [Online]. Available: https://nordvpn.com/cybersecurity/glossary/blue-pill-attack/
3. "Alert AA21-131A: DarkSide ransomware: Best practices for preventing business disruption from ransomware attacks." Cybersecurity & Infrastructure Security Agency (CISA). Accessed: Feb. 14, 2025. [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-131a
4. M. Antonakakis et al., "Understanding the Mirai botnet," in *Proc. 26th USENIX Conf. Secur. Symp. (SEC)*, 2017, pp. 1093–1110. Accessed: Feb. 17, 2025. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis
5. "State of ransomware 2024: A year of surges and shuffling," Black Kite Research & Intelligence Team, Boston, MA, USA, 2024. Accessed: Feb. 19, 2025. [Online]. Available: https://blackkite.com/wp-content/uploads/2024/04/BlackKite_Report_Ransomeware-2024_r5-1.pdf

BOB MALEY is the chief security officer at Black Kite, Boston, MA 02199 USA. Contact him at bob@c-ooda.com.

JOANNA F. DEFRANCO is an associate professor of software engineering at The Pennsylvania State University, University Park, PA 16802 USA, and an associate editor in chief of *Computer*. Contact her at jfd104@psu.edu.

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

Why Large Language Models Appear to Be Intelligent and Creative: Because They Generate Bullsh*t!

Daniel M. Berry, University of Waterloo

*This article tries to explain why so many perceive large language models (LLMs), such as ChatGPT, as intelligent and creative. An LLM generates convincingly cogent bullsh*t (BS) in the Frankfurtian sense. This BS is perceived by humans as intelligent and creative.*

Digital Object Identifier 10.1109/MC.2025.3547458
Date of current version: 29 May 2025

These days, despite the careful explanations by tech-savvy artificial intelligence (AI) explainers,[1,2] many are falling for the hype[3,4] and are coming to the conclusion that large language models (LLMs), such as ChatGPT,[5] are truly intelligent and creative,[1,6,7] for example:

> "13 October 2023.[a] ChatGPT changes everything! This and other smooth-talking artificial intelligences will soon be sentient![b] If they're not already!" [a posted image that John Horgan[1] saw]

and

> "... we've reached a momentous point. Large language models, or LLMs, can often seem to wield something close to human intelligence, at least to us nonexperts."[8]

[a]This is the date on which ChatGPT was released to the world.
[b]Google felt that it was necessary to fire an engineer who was claiming that its AIs were sentient.

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

Even *my own* initial reaction on seeing ChatGPT in operation was that, finally, here's an AI that might actually be intelligent! Careful thinking brushed that thought aside.

The question remains: Why do so many, even tech-savvy, people perceive LLMs and their chatboxes to be intelligent and creative?

This article attempts to answer that question by combining two independent observations, one about LLMs and one about bullsh*t (BS). It limits itself to proposing and discussing a possible answer to this question. Until the answer is known for sure, any suggestions offered in this article about fixing LLMs to avoid this phenomenon would be speculation.

LLMs AS BS GENERATORS

In recent years, many have observed that generative AI tools based on LLMs often do not tell the truth, that they generate plausible text that is simply not true. For example, Goghlan, Randell, and O'Boyle asked ChatGPT to tell them about Percy Ludgate, a little known early inventor of a computer. They estimated that 48% of what ChatGPT said about Ludgate was made up and false.[9,10] John Herrman, a tech columnist at *Intelligencer* describes the latest enhancement of ChatGPT, ChatGPT-4o, "o" for "omni," with "ChatGPT is now better than ever at pretending it's something that it's not." [11] Simon Willison admitted that,

> **"We need to tell people ChatGPT will lie to them, not debate linguistics.**
>
> **ChatGPT lies to people.** This is a serious bug that has so far resisted all attempts at a fix. We need to prioritize helping people understand this, not debating the most precise terminology to use to describe it."[12]

Some have even taken advantage of this bug of an LLM's being able to speak nonsense coherently. Evan Ratliff, who created an A.I. voice as a prank to deal with telemarketers for him, after seeing how friends reacted to talking with it, observed that, "A.I. agents are already triggering an avalanche of synthetic conversation, as they are deployed as tireless, unflagging talkers, capable of endless invented chatter."[13]

The cause of this phenomenon is that the LLMs driving the generators are built out of the random data that are available out on the Internet, which is replete with nonsense. The data for the LLMs are not vetted for truth, simply because doing so is neither feasible nor possible.[14,15]

Vint Cerf in email communication with me added that an LLM is trained on only human discourse, but not human common sense,[8] for the simple reason that it is impossible to codify all common sense. The deficiency of common sense in the training data leads to fabrications built out of truths, fabrications that would be ruled out by common sense. Thus, even if an LLM is trained on only truthful data, it will still generate untrue statements.

An early term used to describe the nonfactual output of the generators was "hallucination." For example, Hal Berghel said of the claim by Google's Bard that the first photographs of planets outside the solar system were taken by the James Webb Space Telescope, "Such nonsense generation is labeled by those in the know by the euphemism 'hallucination.'"[14]

Later, after Berghel explains the effect of the lack of vetting on the reader of the generated output, he concludes that "This is precisely what is wrong with nonvetted recommender systems: Without knowing the source of recommendations, there is no way to assign credibility and value to the recommendations."[15] He asks, "Will AIChat (qua automated bloviation) become a displacing technology?"[15]

Later, Berghel explicitly called the nonfactual output of the generators "bloviation," a euphemism for "BS," while giving a succinct description of the bloviated output.

> "The current litter of generative AI tools have reached the apex of automated bloviation. ... With generative AI, it is now possible to create syntactically well-formed content that, while it looks and feels like the product of mature thought to an unprepared mind, it actually involves negligible cognitive investment."[16]

Vint Cerf concluded the email cited previously with the exclamation, "LLMs are bullsh*tters [sic]."

The use of these terms to describe the output of LLMs has been informal, based on observations backed by common sense. However, since 2005, "bullsh*t" has been a technical term that has a definition, that of Harry Frankfurt, that is accepted by many doing research into the general phenomenon of BSing, in literature and in life.[17]

Very recently, Hicks et al. explained how any LLM, such as ChatGPT, can be regarded as a machine that generates soft BS[18] in the Frankfurtian sense. They use Frankfurt's defining feature of BS (p. 340)[19]:

> *"a lack of concern* with truth, or an *indifference to how things really are* [our emphasis]"

They note that an LLM, such as ChatGPT, is designed with the goal of generating cogent text, text that reads as though it was written by a human being. Beyond this goal, there is no requirement that the text the LLM generates bears any relation to the truth since the data from which the LLM learns are not vetted for truth. The LLM is thus indifferent, even careless, as to the truth. That the LLM is indifferent to the truth is freely admitted by the creators of the LLM, when they say that "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers."[5]

As such, beyond the intention for the LLM to produce text that appears to have been written by a human, there is no intention for the LLM to deceive.

Thus, say Hicks et al., an LLM generates at least soft BS[c] and is therefore, a BSer.[d]

Trevisan et al. provide independent empirical corroboration that an LLM-based chatbot, such as ChatGPT, usually produces Frankfurtian BS because the LLM is knowingly trained on data that have not been vetted for truth,[e,20] Basically, they show that, with statistical significance, in two relevant measurable language characteristics, the pair-wise difference between

- 1000 *Nature* articles written by humans and
- 1000 like-sized articles generated by ChatGPT from the titles of the *Nature* articles

is the same as
the pair-wise difference between

- 45 articles from the BNC written by humans and
- 45 BSy political-party manifestos written also by humans

and between

- 50 articles about non-BS jobs written by humans and
- 50 articles about BS jobs written also by humans.

That is, Trevisan et al. were able to demonstrate that their LM can reliably distinguish BS from non-BS in natural human language. Consequently, they reason that the distinction detected by their LM between the 1000 *Nature* articles and the 1000 ChatGPT-generated articles on the same topics must be the presence of BS in the latter. See "Trevisan et al." for more detail.

[c]Generating hard BS requires intent, and it is not clear that an LLM has any intent.

[d]Hicks et al. go on to explain the dangers of regarding an LLM as hallucinating when it makes mistakes. Using such an anthropomorphism leads to ascribing greater intelligence and agency to the LLM than it can possibly have. Describing an LLM as a BSer gives a more accurate picture.

[e]They do argue that with a different design, training the LLM on only data that have been vetted for truth, an LLM-based chatbot can be produced that is not indifferent to truth and, therefore, does not produce Frankfurtian BS.

WHY LLMs ARE PERCEIVED AS INTELLIGENT

Turpin et al., using Frankfurt's definition of BS, explain how, hypothesize that, and prove empirically that, a human's ability to BS convincingly is taken by other humans as an honest signal of the [first] human's intelligence.[21] Humans evidently understand instinctively that to be able to tell lies or nontruths that appear to be true requires intelligence.

From Hicks et al.'s description of an LLM's output as soft BS, it is clear that an LLM's output appears to a human exactly as human-generated BS appears to a human. The output reads as though it is written by a human being, and it may or may not be true. Therefore, the LLM's output serves as an honest sign of the intelligence of whatever generated the output, the LLM, even though in fact, the LLM has no intelligence whatsoever. Therefore, anyone who does not understand the reality of an LLM, feels in er[f] guts that the LLM is truly intelligent.

This effect might be compounded by what appears to be a common misunderstanding of the term "artificial intelligence." It is my observation that many nontechies understand "artificial intelligence" as "an artificial being that is truly intelligent" rather than as "faked intelligence," which is what the techies who coined "artificial intelligence" meant. I came to this observation after repeatedly seeing how nontechies use the term and the tacit inferences they draw from it. This misunderstanding of the term would leave nontechies predisposed to believing that LLMs are intelligent.

Unfortunately, there appear to be some people, such as those in high tech, who, by all expectations, should understand the reality of an LLM, but nevertheless behave as though they ascribe real intelligence to the LLM. If any of these people is not deliberately exaggerating the LLMs abilities, it seems that this person's System I has decided that the LLM is truly intelligent long before er System II is able to conclude that it is really a BSer, and System I's decision prevails.[g,22]

While the explanation of this section of why LLMs are perceived by humans as intelligent is fairly compelling, it would be useful to confirm its validity with an experiment similar to that described by Turpin et al. using some of the documents prepared and tested by Trevisan et al. See "Experiment Details" for more detail.

WHY LLMs ARE PERCEIVED AS CREATIVE

There are many definitions of creativity and of creative ideas, such as given in the *Oxford Handbook of Creativity, Innovation, and Entrepreneurship*.[23] Almost all mention innovativeness as a key element of creativity. I have long considered a creative, innovative idea as an exception, from the norm, that is perceived, in retrospect, to be a good idea after all. Here, an exception may range

- from an inadvertent failure to follow a procedure or rules, a mistake,
- to an intentional deviation from the current conventions or styles, a thinking out of the box.

For example, a composer who decides to deviate from the prevailing style of music, tries sequences of notes until E finds one that sounds good to er ears. If enough concert goers agree with em, the resulting composition is considered a creative innovation. Eventually, this composition becomes part of the norm, possibly later being the norm from which an exception leads to yet another creative innovation.[24]

In brainstorming, the exceptions are intentional. In the idea generation phase, brainstormers are intentionally deviating from the norm, that is,

[f]"E," "em," and "er" are gender nonspecific third-person singular pronouns in subjective, objective, and possessive forms, respectively.

[g]As Kahneman explains in *Thinking, Fast and Slow*, the older part of a human brain, System I, comes to its instinctive decisions from stimuli that resemble past events much more rapidly than the newer part of the human brain, System II comes to its reasoned, logical decisions from the same stimuli.[22]

TREVISAN ET AL.

Trevisan et al.'s work was an attempt to determine if BS can be reliably detected using computational methods.[20] They built an LM (learned machine) identifies and measures differences in two language characteristics between 1) the contents of 1,000 published articles in *Nature* and 2) the contents of 1,000 corresponding articles generated by ChatGPT when prompted with the titles of the *Nature* articles with instructions to mimic the style of *Nature*. The two characteristics were 1) the frequencies of characterizing words and 2) traces of characterizing contexts. They found that each of the two characteristics distinguished the corresponding pairs of articles with 100% accuracy with at least 99.84% confidence. The question is whether what distinguishes the corresponding articles is BS. If it is, then this LM could function as a BS meter.

Trevisan et al. tested whether the distinctions that the LM identifies and measures are BS by running the LM on two constructed samples of BS articles paired with non-BS articles, all written by humans.

- Based on Orwell's observation that what political parties say is BS,[S1] Trevisan et al. ran the LM to compare 45 manifestos published by U.K. political parties from 1945 through 2005 with 45 similarly sized texts from the British National Corpus (BNC) of everyday language in the United Kingdom, texts that did not include transcripts of political meetings and news reports, which often include political commentary.

  The LM showed, with statistical significance, that the distinctions between the everyday discourse, and the party manifestos were the same as for the *Nature* articles and the ChatGPT-generated articles.
- Based on Graeber's description of BS jobs,[S2] Trevisan et al. ran the LM again to compare 50 texts from online sources that were likely to have been written by people employed in such BS jobs with 50 similarly sized text from online sources that were likely to have been written by people employed in other than such BS jobs. None of either set of texts were written by scientists, and none were deemed after careful examination to have been written by ChatGPT.

  The LM showed, with statistical significance, that the distinctions between the non-BS-job texts and the BS-job texts were the same as for the *Nature* articles and the ChatGPT-generated articles.

Trevisan et al. were thus able to demonstrate that the LM can reliably distinguish BS from non-BS in natural human language.

Therefore, the distinction that the LM detected between the 1,000 *Nature* articles and the 1,000 corresponding ChatGPT-generated articles must have been the presence of BS in the latter.

REFERENCES

S1. G. Orwell, "Politics and the English language," in *The Collected Essays, Journalism and Letters of George Orwell*, vol. 4, S. Orwell and I. Angus, Eds., London, U.K.: Secker & Warburg, 1969, pp. 127–140.

S2. D. Graeber, *Bullshit Jobs: A Theory*. New York, NY, USA: Simon & Schuster Paperbacks, 2018.

EXPERIMENT DETAILS

Build a collection of lists of, say, 20 articles, from among the 1,190 articles written by humans and the 1,000 articles generated by ChatGPT used by Trevisan et al. Each list consists of, in randomized order, 1) five non-BS articles written by humans, 2) five BS articles written by humans, 3) five non-BS articles written by ChatGPT, and 4) five BS articles written by ChatGPT. Obtaining the non-BS and BS articles written by ChatGPT will require vetting the 1,000 articles generated by ChatGPT for truth.[n] In the experiment, half of the subjects are tech savvy and half are not. Each subject is asked to rate, in a Likert scale, the intelligence of the author of each of the 20 articles in a list randomly chosen from the collection. It will be interesting to see which authors are rated as more intelligent, humans or ChatGPT, and it will be interesting to see how the ratings of the subjects who are tech savvy compare with the ratings of the subjects who are not.

[n] I hope that some of the ChatGPT-generated articles have no untrue sentences in them.

making exceptions, to rapidly generate as many ideas as possible, under the regime of quantity before quality. Later in the pruning phase, the brainstormers evaluate the ideas to keep only those ideas, the exceptions, that turned out to be good ideas after all.

An LLM is a BSer. A good fraction of what it produces are exceptions. Some of these are seen by humans, upon examination, to be good ideas after all. Ergo, the LLM is perceived as creative.

Others have come to a similar conclusion about LLMs and are parlaying LLMs to speed up breakthroughs. For example, Broad observes

> "Artificial intelligence often gets criticized because it makes up information that appears to be factual, known as hallucinations. In science, however, AI hallucinations can be remarkably useful. Smart machines are dreaming up riots of unrealities that help scientists track cancer, design drugs, invent medical devices, uncover weather phenomena, and even win the Nobel Prize."[25]

He then describes the process of an AI's hallucinating proteins that do not exist in nature but are possible according to the data on which the AI has been trained. Some of these hallucinations are recognized by human scientists as possibly useful, and are then tested to verify that they do what they are thought to do.

An LLM is designed to generate text that appears to have been written by a human from many data that may or may not be true, with no concern as to the truth of what it generates. Therefore, an LLM is a BSer in the Frankfurtian sense. Since humans perceive the ability to BS convincingly as a signal of intelligence, humans that do not know the true nature of LLMs perceive an LLM as intelligent. Since a creative idea can be an exception that humans perceive to be a good idea after all, humans perceive some of the possibly false ideas generated by an LLM as creative. These perceptions are despite that the LLM was designed only to be cogent and not specifically to be intelligent or creative. Ultimately, the intelligence and creativity of LLMs are entirely in the eyes of the beholder.

ACKNOWLEDGMENT

Daniel Berry's work was supported in part by a Canadian NSERC Grant RGPIN-2023-03584. The author thanks Vint Cerf, Ahmed ElShatshat, Alessio Ferrari, Eddy Groen, Andrea Herrmann, Cliff Jones, Jay Judkowitz, John Mylopoulos, Vicky Sakhnini, Richard Schwartz, Joe Slater, and Jeff Voas for their comments on previous drafts of this article.

REFERENCES

1. J. Horgan, "Cutting through the ChatGPT hype," *John Horgan (The Science Writer)*, Oct. 13, 2023. [Online]. Available: https://johnhorgan.org/cross-check/cutting-through-the-chatgpt-hype
2. T. Steiner. "Is ChatGPT intelligent?" LinkedIn. Accessed: Mar. 12, 2025. [Online]. Available: https://www.linkedin.com/pulse/chatgpt-intelligent-thomas-steiner-3j4ie
3. M. Saltzman. "The artificial intelligence boom: Understanding the hype (and concern) over ChatGPT." Everything Zoomer. Accessed: Mar. 12, 2025. [Online]. Available: https://everything-zoomer.com/money/2023/06/06/the-artificial-intelligence-boom-understanding-the-hype-and-concern-over-chatgpt/
4. "Understanding the hype of ChatGPT: An in-depth look at the AI-powered chatbot revolution." First Line Software. Accessed: Jul. 26, 2024. [Online]. Available: https://firstlinesoftware.com/blog/understanding-the-hype-of-chatgpt-an-in-depth-look-at-the-ai-powered-chatbot-revolution/
5. "Introducing ChatGPT." OpenAI. Accessed: Mar. 12, 2025. [Online]. Available: https://openai.com/index/chatgpt/
6. O. Atkins. "ChatGPT: A creative tool or a threat to human creativity?" Creative Salon. Accessed: Mar. 12, 2025. [Online]. Available: https://creative.salon/articles/features/qotw-chatgpt-creative-tool-threat-to-creativity
7. J. Wei et al., "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.
8. P. Greenwood, "Will AI ever have common sense? Transcript of 'The Joy of Why' podcast of Steven Strogatz's interview of Yejin Choi," *Quanta Mag.*, Jul. 18, 2024. [Online]. Available: https://www.quantamagazine.org/will-ai-ever-have-common-sense-20240718/?mc_=%20cid=3D6520b06369&mc_eid=3D66f5e73daf
9. B. Coghlan, B. Randell, and N. O'Boyle, "ChatGPT's astonishing fabrications about Percy Ludgate," 2023. [Online]. Available: https://www.scss.tcd.ie/SCSSTreasuresCatalog/miscellany/TCD-SCSS-X.20121208.002/ChatGPTs-AstonishingFabrications-aboutPercyLudgate-CoghlanRandellOBoyle-20230424-1434.pdf
10. B. Randell, and B. Coghlan, "ChatGPT's astonishing fabrications about Percy Ludgate," *IEEE Ann. Hist. Comput.*, vol. 45, no. 2, pp. 71–72, Apr./Jun. 2023, doi: 10.1109/MAHC.2023.3272989.
11. J. Herrman, "GPT-4o is OpenAI's plan to win friends and influence people," *Intelligencer, NY Mag.*, May 23, 2024. [Online]. Available: https://nymag.com/intelligencer/article/gpt-4o-is-openais-plan-to-win-friends-and-influence-people.html#
12. S. Willison, "We need to tell people ChatGPT will lie to them, not debate linguistics," *Simon Willison's Weblog*, Apr. 7, 2023. Accessed: Mar. 12, 2025. [Online]. Available: https://simonwillison.net/2023/Apr/7/chatgpt-lies/

13. E. Ratliff, "I created an A.I. voice clone to prank telemarketers. But the joke's on us," *New York Times*, Oct. 10, 2024. [Online]. Available: https://www.nytimes.com/2024/10/10/opinion/ai-voice-telemarketers.html
14. H. Berghel, "ChatGPT and AIChat epistemology," *Computer*, vol. 56, no. 5, pp. 130–137, 2023, doi: 10.1109/MC.2023.3252379.
15. H. Berghel, "Fatal flaws in ChatAI as a content generator," *Computer*, vol. 56, no. 9, pp. 78–82, 2023.
16. H. Berghel, "Generative artificial intelligence, semantic entropy, and the big sort," *Computer*, vol. 57, no. 1, pp. 130–135, 2024.
17. H. G. Frankfurt, *On Bullshit*. Princeton, NJ, USA: Princeton Univ. Press, 2005.
18. M. T. Hicks, J. Humphries, and J. Slater, "Correction: ChatGPT is bullshit," *Ethics Inform. Technol.*, vol. 26, no. 3, 2024, Art. no. 46, doi: 10.1007/s10676-024-09785-3.
19. H. G. Frankfurt, "*Reply to Cohen*," *in The Contours of Agency: Essays on Themes from Harry Frankfurt*, S. Buss and L. Overton, Eds. Cambridge, MA, USA: MIT Press, 2002.
20. A. Trevisan, H. Giddens, S. Dillon, and A. F. Blackwell, "Measuring bullshit in the language games played by ChatGPT," 2024, *arXiv:2411.15129*.
21. M. H. Turpin, M. Kara-Yakoubian, A. C. Walker, H. E. K. Walker, J. A. Fugelsang, and J. A. Stolz, "Bullshit ability as an honest signal of intelligence," *Evol. Psychol.*, vol. 19, no. 2, 2021, Art. no. 14747049211000317, doi: 10.1177/14747049211000317.
22. D. Kahneman, *Thinking, Fast and Slow*. London, U.K.: Penguin, 2012.
23. J. Zhou, Ed., *The Oxford Handbook of Creativity, Innovation, and Entrepreneurship* (Oxford Library of Psychology Series). Oxford, U.K.: Oxford Univ. Press, 2016.
24. T. Chakraborty and S. Masud, "The promethean dilemma of AI at the intersection of hallucination and creativity," *Commun. ACM*, vol. 67, no. 10, pp. 26–28, 2024, doi: 10.1145/3652102.
25. W. J. Broad, "How hallucinatory A.I. helps science dream up big breakthroughs," *New York Times*, Dec. 23, 2024. [Online]. Available: https://www.nytimes.com/2024/12/23/science/ai-hallucinations-science.html

DANIEL M. BERRY is a professor of computer science and software engineering at the University of Waterloo, Waterloo, ON N2L 3G1, Canada. Contact him at dberry@uwaterloo.ca.

IEEE Security & Privacy magazine provides articles with both a practical and research bent by the top thinkers in the field.

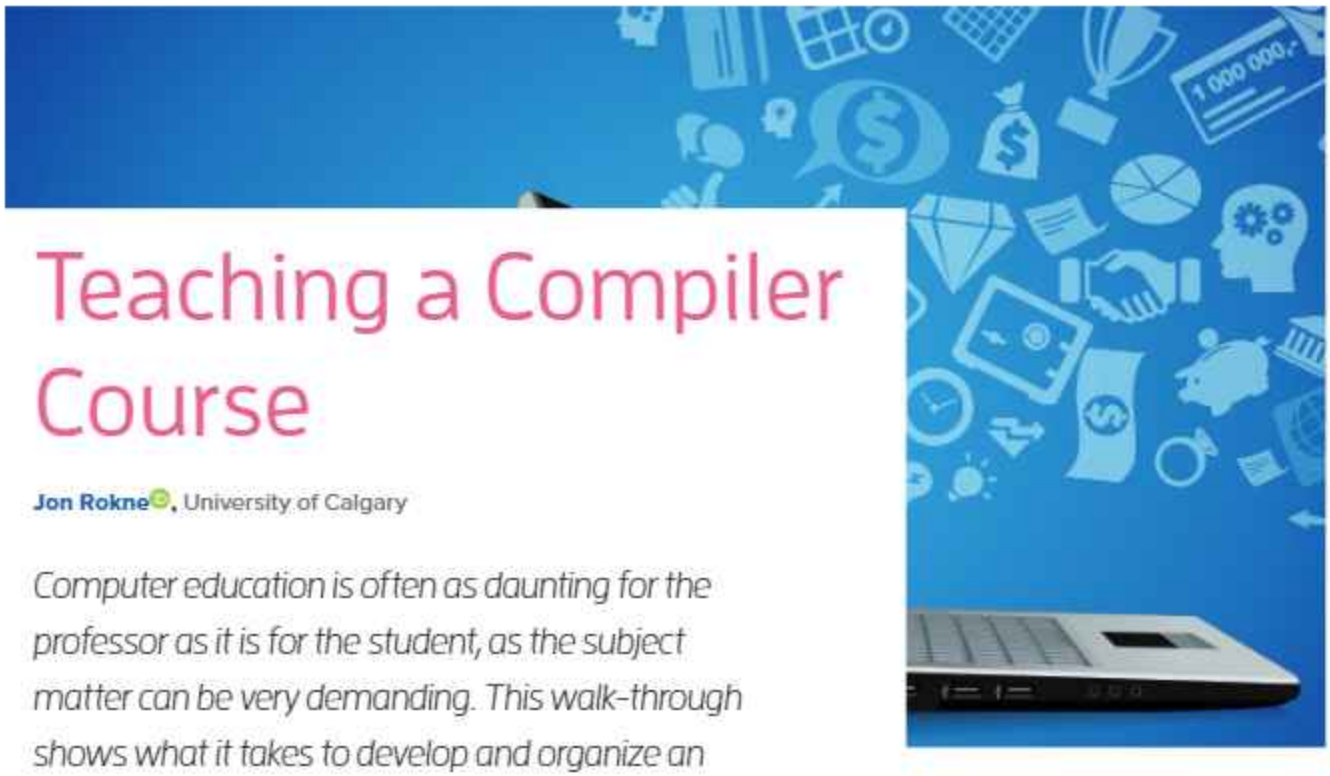
- stay current on the latest security tools and theories and gain invaluable practical and research knowledge,
- learn more about the latest techniques and cutting-edge technology, and
- discover case studies, tutorials, columns, and in-depth interviews and podcasts for the information security industry.

computer.org/security

IEEE COMPUTER SOCIETY IEEE

Digital Object Identifier 10.1109/MC.2025.3562022

EDITORS
GEORGE HURLBURT U.S. Federal Service (Retired), USA; gfhurlburt@gmail.com
SOREL REISMAN California State University, USA; sreisman@computer.org



Teaching a Compiler Course

Jon Rokne, University of Calgary

Computer education is often as daunting for the professor as it is for the student, as the subject matter can be very demanding. This walk-through shows what it takes to develop and organize an advanced computer course from scratch.

I was asked to teach a third-year computer science course, Computer Science 411, "Introduction to Compiler Construction," since the regular instructor was on an extended medical leave. I had about one month to prepare the course. I accepted the challenge. The department chairs stated it this way:

> "You have been instrumental, in stepping in at the last minute when our colleagues needed help, even if it was outside your area of expertise."

The content of the course was definitely outside my area of expertise. My formal background in the subject was scant, consisting of a 1963 Algol programming course in Norway. To further quote, the course was intended to consist of the following:

1. twenty-six teaching hours where the following topics were treated: Metalanguage, Backus normal form, representation of numbers, and so on
2. demonstration of the electronic calculator—GIER
3. four obligatory exercises involving writing a program, punching the program on a Friden Flexowriter, and so on.

Further background included writing programs in computer languages such as: Prolog, Lisp, Mathematica, PL1, Algol, and C. It turned out that some knowledge of C was the most useful.

Essentially compiling a program written in some computer language for a given computer consists of translating the statements in the computer language into a new language that is executable by the computer. This is much like translating text written in one human language into another. One advantage of computer languages is that they are generally more straightforward and more rigorously specified than human languages. In both cases, the allowable constructs of the languages are governed by grammars, that is, the rules specifying the allowable sentences in the languages.

A blog by Stuntz[1] argues that writing a compiler is an important exercise for software professionals.

Compiler construction is helpful to nearly all software engineers, especially for those who will not build compilers as part of their daytime job.

Digital Object Identifier 10.1109/MC.2025.3551196
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

TEACHING COMPILER CONSTRUCTION

The first step toward developing a course in compiler construction was to search for material on teaching the subject. One article, titled "Is Compiler Construction a Dead Subject?"[2] did not bode well. However, the article, which considered the subject from the point of view in India, turned out to be thoughtful. It stated:

> "First, compilers are complex programs, and a subject called 'compiler construction' should be formally established. Second, and more importantly, compilers are useful software programs that can revolutionize the art of computer programming."

The article also noted that when a team at IBM led by John W. Backus was asked to develop a compiler for FORTRAN for the IBM 704, the team expected the work to take six months. In reality, it took 18 man-years of effort over 2.5 years. This established that developing a compiler for a programming language was a nontrivial task.

Therefore, I had to assume the expectation of the work required to present a compiler course that covered the essentials of compiler construction would be a significant effort.

Teaching the course turned out to be more challenging than expected. First of all, I was not provided any course materials. The instructor stepping out of the teaching did not offer any course material. A previous instructor did provide me with some slightly useful information. Essentially, I felt I was dealing with a case of tabula rasa.

SELECTING THE TEXT

When preparing for any course, the first step is acquiring source material in books and papers. The first frequently mentioned book for compiler design was "Compilers: Principles, Techniques, and Tools."[3] It has been known as the "Dragon Book" to generations of computer scientists, as its cover depicts a knight and a dragon in battle, representing a metaphor for conquering complexity."[4] However, it, and several other texts, suffered from overkill. They had from 800 to 1,000 or more pages of text. While I could easily "read" 1,000 pages, it was impossible to absorb the information and ideas contained in these pages in a short time frame.

> One advantage of computer languages is that they are generally more straightforward and more rigorously specified than human languages.

A blog post summed it up:

> "Don't read the dragon book if you're interested in compilers. It's lex and parse heavy, and is not up to date with more recent best practices on codegen. Further, the lex and parse end of it is focused on the kind of theory you need to build tools like lex and yacc, rather than stuff you need to know if you want to write a compiler."[5]

This led me to continue the search for more suitable texts. I found a list of 63 compiler books.[6] One of the available texts seemed reasonable—"Engineering a Compiler."[7] A second book, that was not on the list, was "Introduction to Compilers and Language Design" by Thain.[8] Thain's book had the advantage of being much shorter than the average compiler text. It was also less expensive for the students, while still covering the critical aspects of writing a compiler.

An Internet search for supplementary materials provided a wealth of helpful information in the form of papers and web pages dealing with various subjects applied to compiler construction.

When reading the books and the information on the webpages, it became clear that compiler writing was fraught with two significant deficiencies.

1. The language used to describe objects, concepts, and algorithms in the compiler world was inconsistent. This made it difficult to understand the exact meaning of what was written. When discussing the scanning of input code: "That process of going character by character can be wrapped up into a routine-also called a function, a method, a subroutine, or component.[9] Little in computing has a single, reliable name, which means everyone always argues over semantics."
2. More importantly, the lack of standard notation was an issue. Whereas in the established subjects such as mathematics, notation is fairly well standardized, in the case of compilers 2 3+4 could mean anything depending on the source language and the compiler. This problem is discussed further by Iverson.[10]

After having absorbed a certain amount of compiler knowledge, I decided to use the text by Thain, "Introduction to Compilers and Language Design," as the course text. It was easily available, not expensive, and had a freely available online pdf version.

SELECTING THE LANGUAGE

Moving forwards, there was obviously a need for some way to implement the steps of the compilation. Ideally, it would be using the set of hardware instructions for a computer. Since, as

noted previously, this would require man-years of work, this approach was clearly impractical. A simple computer language that students could use to implement the steps of compilation as part of their assignments had to be chosen. The assignments would ask the students to write a set of programs that would implement the stages of compilation, starting from source language statements to machine executable instructions. It was preferable to choose a language the students were at least slightly familiar with and where the knowledge of the language would be useful in other courses and later in the workplace. This ruled out the "toy" languages that seemed to be favored by some due to the overhead of learning a new language. Learning the details of a "toy language" (although Stuntz[1] finds it useful in some contexts) would also be a wasted effort since it would be useless in other courses. The most reasonable choice therefore seemed to be a simple subset of C, denoted as C- as defined in the book "Engineering a Compiler."[7]

The choice was also motivated by:

> "C is as big a deal as you can get in computing. Created by Dennis Ritchie starting in the late 1960s at Bell Labs, it's the principal development language of the UNIX operating system. .... And everywhere that Unix went, C was sure to go."[9]
>
> Furthermore, "C is a simple language, simple like a shotgun that can blow off your foot. It allows you to manage every last part of a computer—the memory, files, a hard drive, which is great if you're meticulous and dangerous if you're sloppy. Software made in C is known for being Fast."[9]

C would therefore provide the tools used to write programs for the various stages of the compilation process. This meant that the process was started at a more advanced level than the original developers of compilers, where they only had the bare bones of machine instructions as they initially developed compilers.

I also decided to eschew use of tools like lex, yacc, and the like. While those tools are very useful when constructing a compiler as part of a practical compiler development, they hide some of the fundamental steps of the compilation process. Tool use also tends to bloat the final code for a compiler as well, an issue that is now coming to the foreground for all software development as noted by Hubert[11]: "Why Bloat is Still Software's Biggest Vulnerability. A 2024 plea for lean software."

The next question was how to specify C- precisely so that the compilation of C- programs could be implemented in C. A formal description was needed.

The tool for describing the most common computer languages is regular grammars. So, what are regular grammars? A precise definition was needed. First stop on the quest for a definition was a Wikipedia article.[12] This led, at once, to the most familiar problem with compiler teaching: *In theoretical computer science, and formal language theory, a regular grammar is a grammar that is right-regular or left-regular. While their exact definition varies from textbook to textbook, they all require that:*

> All production rules have at most one nonterminal symbol,
>
> That symbol is either always at the end, or always at the start, of the rule's right-hand side.
>
> Every regular grammar describes a regular language.

Note the innocent comment: "While their exact definition varies from textbook to textbook." This turns out to be a core problem with compiler studies. It is used with topics that are described differently depending on the textbook used, paper written, and lecture notes constructed. In addition to this notation, the terminology used can vary from source to source for the putative same topic.

This meant that the choice of descriptions and notations would, to some extent, restrict the material used for the teaching since for teaching purposes these items should be consistent to minimize confusion. For the regular grammars the quotation and description used for this version of teaching compilers was the one used in the book by Thain,[8] which then meant that other source material could only be used if it was reasonably consistent with the notation and description in Thain.[8]

Generally, the grammars are considered to have terminal and nonterminal symbols, operators, and expressions. The expressions consist of terminal and nonterminal symbols joined by operators. The operators are invoked by precedence rules, which can be overridden using parentheses.

The grammar for C- was adopted from Louden[13] Appendix A. The notation used was modified to the notation used in Thain's book.[8]

CONSTRUCTING THE ASSIGNMENTS

The planned assignments for the course were: scanner, parser, semantic analyzer, and code generator for the G language. The scanner assignment was:

> In this project, your task is to create a scanner for the C- programming language. You can refer to the C- language specification provided here for guidance. Your scanner output should consist of a list of tokens, each annotated by its token kind (such as identifier, keyword, number, and so on) and its corresponding source code location (line number). Print the corresponding lexeme if the token is an ID or a NUM. If invalid input is discovered, your scanner should produce a warning message and continue scanning the file.

This project involved taking a C- program text, scanning the characters one by one, and then recognizing the identifiers, keywords, and numbers occurring in the string and labeling them as tokens and lexemes.

An example input was provided:

```
void main (void) int x; int y; x: input 0; y: input 0; output (ged(x, y));
```

with the expected output so that the scanner could be tested uniformly for all the submissions. While I recommended that the students use C to implement the scanner, any other suitable computer language was also acceptable.

The assignment specified that the student should develop a consistent set of initializing tokens in the C definition. If they did not, one or more error messages should be output.

The second assignment was to create a parser for the C- programming language building on the scanner developed in assignment 1. The parser should read in a source file, determine whether the scanned items have valid sentences according to the C- grammar, and indicate success or failure. It should be evaluated for ambiguities given the C grammar and work to resolve problems such as the dangling-else. They were asked to create a set of complete tests to exercise all of the tricky corner cases.

A discussion of static-versus-dynamic semantics took place during the lectures at this point.

The third assignment was the semantic analyzer. Here, the task was to create a semantic analyzer for the C- programming language building on the scanner and the parser previously developed in assignments 1 and 2. The semantic analyzer should take the output from the parser and determine if the semantics are valid. When the input was semantically valid, the program should print the abstract semantic tree, with some additional information on the nodes, including types for expressions, and unique identifiers for IDs. Otherwise, the program should indicate failure and halt. A set of complete tests to exercise all of the tricky corner cases was to be created.

A fourth assignment was to take the output from the semantics analysis and generate low-level code close to the machine instructions. This was intended, but not implemented due to course time constraints.

Teaching a compiler course with minimal background in the subject is an engaging experience. It exposes the instructor to the various stages of compilation, provides experience in low-level programming (assuming no compiler-construction special tools are used), and provides a valuable learning experience for students and the instructor.

It should be noted that valuable support for the course in the form of interactive development of assignment and development of tutorial examples was provided by three highly qualified graduate teaching assistants.

REFERENCES

1. C. Stuntz, "On learning compilers and creating programming languages." Craig Stuntz. Accessed: Jun. 3, 2025. [Online]. Available: https://www.craigstuntz.com/posts/2023-10-13-learning-compilers-and-programming-languages.html
2. P. Chakraborty, "Is 'compiler construction' a dead subject?," *Current Sci.*, vol. 108, no. 5, pp. 777–778, Mar. 2015.
3. Aho, A. V. Alfred, V. Aho, M. S. Lam, R. Sethi, and D. Jeffrey, *Ullman-Compilers-Principles, Techniques, and Tools*, 2nd ed. Reading, MA, USA: Addison Wesley, 2006.
4. "When people recommend the Dragon Book, are they recommending the red dragon book, or the green dragon book?" Reddit. Accessed: Jun. 3, 2025. [Online]. Available: https://www.reddit.com/r/learnprogramming/comments/8p0tk5/when_people_recommend_the_dragon_book_are_they/
5. Barrkel, "What language-agnostic programming books should I read?" Hacker News. Accessed: Jun. 3, 2025. [Online]. Available: https://news.ycombinator.com/item?id=14487961
6. "Compiler books." Goodreads. Accessed: Jun. 3, 2025. [Online]. Available: https://www.goodreads.com/shelf/show/compiler
7. L. Torczon, and K. Cooper, *Engineering a Compiler*. San Mateo, CA, USA: Kaufmann Publishers Inc., 2007.
8. D. Thain, *Introduction to Compilers and Language Design*, 2nd ed. Morrisville, NC, USA: Lulu. com 2020.
9. P. Ford, "If you can't realize that, you'll better read this Coile: An Essay," Bloomberg Businessweek. 2015. Accessed: Jun. 3, 2025. [Online]. Available: https://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/#:~:text=This%20issue%20comprises%20a%20single,solid%20jokes%20and%20lasting%20insights
10. K. E. Iverson, "Computers Anil mathematical notation." jsoftware. Accessed: Jun. 3, 2025. [Online]. Available: https://www.jsoftrare.com/papers/camm.htm
11. B. Hubert, "Why bloat is still software's biggest vulnerability: A 2024 plea for lean software," *IEEE Spectrum*, vol. 61, no. 4, pp. 22–50, Apr. 2024, doi: 10.1109/MSPEC.2024.10491389
12. "Regular grammar," Wikipedia. Accessed: Jun. 3, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Regular_grammar0
13. K. C. Louden, *Compiler Construction: Principles and Practice*. Boston, MA, USA: PWS Publishing Company, Sep. 1997.

JON ROKNE is a professor in the Department of Computer Science at the University of Calgary, Calgary, AB T2N 1N4, Canada. Contact him at rokne@ucalgary.ca.

EDITORS
TIMOTHY JONES University of Cambridge, U.K., timothy.jones@cl.cam.ac.uk
ROBERT MULLINS University of Cambridge, U.K., robert.mullins@cl.cam.ac.uk

The Dark Side of Computing: Silent Data Corruptions

Dimitris Gizopoulos, University of Athens

Silent data corruptions due to defective silicon cause erroneous program results. Yet nobody knows how severe and frequent the problem is, how much we need to invest in solving it, and who should pay the bill.

Computing is unimaginably beautiful and powerful. In this essay, we touch on a problem that can undermine its beauty and might. We discuss what is probably the darkest side of computing: *silicon defects that corrupt our programs' data and we don't know it*.

Both the beauty and the might of computing result from two properties that all machine users demand.

Property #1: the computed result is *correct*. Program results are used to take important decisions and correctness is nonnegotiable. Depending on the domain, "correctness" may differ. The only correct answer may be "9,753," while in some cases any answer between "9,500" and "9,800" is equally correct.

Property #2: the computation is *fast*. Nobody needs a correct result if it comes late. The result must reach the user in milliseconds, seconds, an hour, or two weeks. In domains with strict timing requirements, speed means meeting deadlines or lives, property, money is lost. In other domains speed means "as fast as possible" or "faster than the previous product."

Humankind never stops pushing the limits of computing engines and always demands correct and fast computations. Often, computer designers must decide which of the two matters more—the choice depends on the criticality of the applications running on the system. Thus, as in all engineering disciplines, *cost matters*. Neither correctness nor speed comes for free. Both incur dramatic costs during design, manufacturing, and operation.

Correctness of operation is constantly in danger when we push the limits: from materials imperfections and flaws of the chip, board, system manufacturing, through the network errors, all of the way to programmer errors and arithmetic precision in parallel programs. Dealing with the aforementioned requires extra time and resources for correctness. For silicon chips, correctness is jeopardized at

Digital Object Identifier 10.1109/MC.2025.3554306
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

design time (hardware bugs), manufacturing time (silicon defects, variability of physical properties), and mission time (aging, radiation).

In an ideal world, the manufacturers of every part and the assembled whole of a computer would invest resources to deliver flawless products; this never happens. Verification of a new design, validation of few prototypes, and comprehensive testing of every chip, board, and system, have tight time and cost bounds. With products delivered in just six months or one year, some chips in the manufactured quantities will inevitably contain defects that are not screened at production. Even the most diligent manufacturing testing process can detect between 95% and 99% of the modeled defects. But even if we were able to screen every defective chip at production, there are new defects that can appear during chip lifetime due to aging or operating conditions.

SILICON DEFECTS AFFECTING PROGRAMS

What is a silicon defect and how it can affect our programs?

Let's take a simple example. Consider a two-input logic OR gate inside an integer multiplier of your CPU or GPU silicon chip. Due to imperfect manufacturing (a broken wire or a short circuit) this gate occasionally gives the wrong logic OR result when both inputs are at logic 0 (the gate output is 1 instead of the correct 0); the other three input combinations give the correct result 1. This defect, which may be either persistent or may depend on certain rare physical parameters, will affect our program execution when *all of the* following conditions are valid:

1. Our program uses multiplication instructions—the program used the defective arithmetic unit.
2. The operands of some of the multiplications feed the OR gate with both inputs at logic 0—the defective gate will produce the wrong result.
3. There is at least one path of logic gates from the defective gate to the multiplier outputs that propagates the wrong gate result—there is at least one wrong multiplication result.
4. The wrong multiplication result propagates through subsequent instructions in a way that harms execution: the program hangs, crashes, or raises an exception because the wrong multiplication result affects the control flow and changes the values of pointers to memory or ends with an incorrect result because the wrong multiplication affects the data flow.

When any of the conditions are not met, the defective OR gate does not matter for the program. Figure 1 visualizes the conditions, showing how a defective gate in a multiplier produces a wrong result which propagates through to erroneous program output, recognizing a dog instead of a cat.

Similarly to the multiplier, *every* single hardware unit inside a silicon chip of a CPU, GPU, or domain-specific accelerator (dynamic random-access memory chips too) can contain defects: registers, caches, buffers, control logic, and arithmetic units. Numbers manipulated by, stored in, and transferred through hardware units may be wrong due to silicon defects.

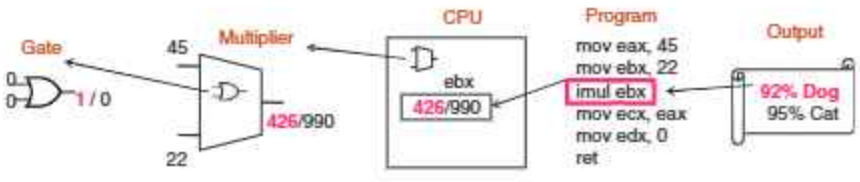


FIGURE 1. From a silicon gate to the program output. The defective gate must receive the inputs that give the wrong output *plus* the gate's output must propagate to the multiplier output *plus* the program must contain a multiplication instruction *plus* the wrong multiplication must affect the program output.

VISIBLE AND SILENT EFFECTS OF DEFECTIVE SILICON

When chips contain defects in their silicon that affect our programs, the first critical question is: "do we know the program ran incorrectly?" If defective silicon leads to a program generating an exception or a system crash (because hardware or software caught the erroneous operation), although we don't like it, at least we know it happened. We try again (hoping that erroneous execution was temporary), and if it happens again, we don't trust the computer for our subsequent work, and we either fix it or buy a new one. But such "visible" errors are not always the case.

Invisible errors [called *silent errors* or *silent data corruptions (SDCs)*] happen when a program runs on a defective chip, completes its execution, does not crash the process or the system, does not raise any exception, and does not produce an output that is obviously wrong. Still the program's result is incorrect and nobody knows! No hardware or software detection mechanisms caught it.

Many readers will, impulsively, think "then add hardware or software detection mechanisms!" Correct, but detection is very costly, and we need to think twice before implementing it. For example, hardware-based error detection and correction codes for memories can incur anything between 2% and 125% (!) extra silicon and power, while software-based redundant execution (double or triple redundancy) has a 100% or 200% extra performance and energy cost. Who is ready to take this?

SDC DISCLOSURES: PAST AND PRESENT

Systematic use of the term SDC likely started after a DSN 2008 panel where

panelists were wondering if it exists or if it is a myth.[1] For decades, we were convinced that SDCs were a myth, or, in the worst-case scenario, an SDC might affect the unlucky user of one in a million (or billion) computers. The reality is different according to the major users of computing chips of our days: data center operating companies.

Hyperscalers (Meta,[2] Google,[3] Alibaba[4]) have disclosed over the last few years an unexpectedly high number of CPUs (1 in 1000) that lead to SDCs—program executions that produce wrong results without any observable indication. Other types of processing chips [artificial intelligence (AI) accelerators from Google[5] and Meta,[6] and NVIDIA GPUs[7]] are also reported to generate SDCs. Reports agree that the root cause of such SDCs are chips which are born defective (escaped manufacturing testing), become defective (aging), or just differ from each other (variability).

A short note on terminology is important here. SDCs are neither detected nor corrected. Once the effect of a defect is detected, it is not silent anymore. An SDC is an eventual (terrible) result of a defect when a program runs. It is not detected; that's why it is called silent (unlike effects like crashes which can't be missed). Per the fault-tolerant computing terminology used for decades: We detect defects (or their models: the faults) and we correct their effects—the errors—at different abstraction layers (circuit, microarchitecture, software) so that the machine delivers the expected service. Public statements like "an SDC detection scheme" and "the SDC is corrected" (even if the speaker or author does know what they talk about) don't educate newcomers to the topic.

What the computing community needs to do for SDCs is to minimize or, ideally, zero their rate. In a perfect world, the result of a computation produced by chips should be either 1) correct (defect does not matter or correction worked), or 2) incorrect but known to be so (detection worked but no correction is possible). A perfect "zero-SDCs" world is impossible because it requires unaffordable protection costs.

> For decades, we were convinced that SDCs were a myth, or, in the worst-case scenario, an SDC might affect the unlucky user of one in a million (or billion) computers.

WHAT COMPANIES REALLY DISCLOSED?

Despite the hype of the incident's reports, it is still unclear how severe the problem is and what its specifics are. Table 1 summarizes what hyperscalers reported.

Clearly, there are still many open questions, such as follows:

- What are the real SDC rates, that is, corrupted program outputs over time when real programs run?
- Which chip microarchitectures are more likely to generate SDCs?
- Which hardware units (sizes, counts, designs) are more likely to generate SDCs when defective (Alibaba hinted floating-point and vector hardware units)?
- Which instructions are more likely to lead to SDCs?
- Which silicon technology nodes are more likely to generate SDCs?

Without knowing which hardware units are likely to generate SDCs, how can one protect them? Without knowing which instruction classes are mainly affected by defects, how can one harden them? Without knowing which CPU microarchitectures are more likely to generate SDCs, how can a hyperscaler decide which to purchase, and how can a chip vendor improve future designs?

Most importantly, how can anyone (chips vendor, system integrator, hyperscaler) take decisions for protection (and pay the costs) against SDCs if we don't know the *scale and severity of the problem*? Bottomline: How many SDCs are happening out there? How many programs per unit of time (hour, day) complete execution with the wrong output in a data center and nobody knows?

Thinking more carefully, the question does not make sense! "Measure the number of SDCs" is a contradiction in terms. We can't measure something that we can't observe because it is silent. So, essentially the mandate changes to: "Estimate the number of SDCs because you can never measure them."

The answer to the estimation request depends on many aspects of

TABLE 1. Hyperscalers disclosures.

	Meta	Google	Alibaba
Text about the number of CPUs that produce SDCs	"... hundreds of CPUs across hundreds of thousands of machines ..."	"... a few mercurial cores per several thousand machines ..."	"... 3.61% of the CPUs are identified to cause SDCs ..."
Interpretation of the text in defective parts (CPUs) per million (DPPM)	≈ 1000	< 1000	= 361

large-scale systems: their size, the defectivity of the chips, the workload types, the execution times, and so on. A first-order estimation (that employs microarchitecture-level fault injection in x86 CPU models) about the number of SDCs considering defects in the arithmetic units (scalar and vector, integer, and floating-point) of data center CPUs, underlines the urgency of the matter and calls for action.[8]

Figure 2 shows some examples of this first-order estimation for varying data center size (number of cores), workload execution time (minutes), and chips defectivity [defective parts per million (DPPM)]. It shows the estimated number of SDCs generated per day in the data center.

The number of incidents in Figure 2 do not refer to visible effects of CPUs defects but to silent ones. When a CPU in a data center causes crashes (or any visible effect), it is removed from the fleet and does not cause crashes anymore. The fleet of CPUs produces hundreds of erroneous program outputs delivered to the users that have absolutely no idea! Now, try to answer: How many SDCs can we afford? Zero! Let's see what we can do to reach this ideal.

REDUCING SDCs AND THE COST

There are multiple ways to reduce the rate of SDCs. Table 2 summarizes them and the corresponding costs.

Everything costs; who is going to pay?

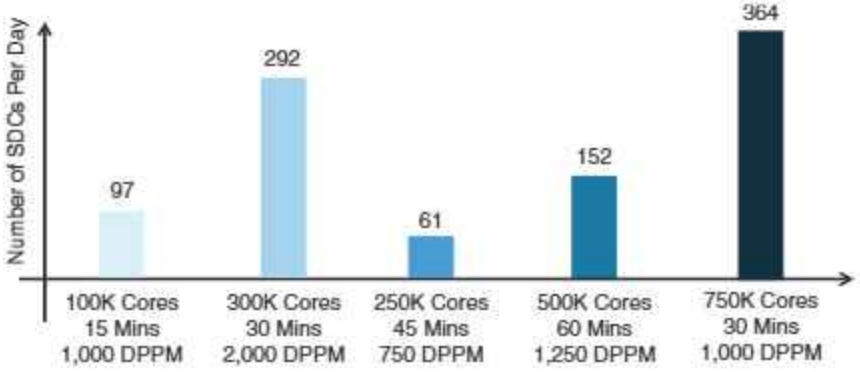


FIGURE 2. First-order estimation using microarchitectural simulation for the rate of SDCs due to defects in arithmetic units of data center x86 CPUs. Bars correspond to different data center size, workload time, and defectivity. These rates can't be measured in real production systems since the effect of the defects is silent.

MORE DISCLOSURES FROM INDUSTRY?

We must praise Meta for the first SDCs disclosure and subsequent action to engage with academia[9] and Google and Alibaba for confirming. We must also praise AMD, Intel, NVIDIA, and Arm for joining hyperscalers in the Open Compute Project (OCP)[10] and helping awareness about the problem and engagement with universities.[11] Initial disclosures were followed by professional organization initiatives (the IEEE Computer Society RAS in the Datacenter Summit 2024[12]) and panels,[13] keynotes,[14] special sessions,[15,16] publications,[8,17] and blog posts.[18,19]

However, beyond the current awareness state, it is unlikely that companies will provide more detailed, thus useful, information about defective silicon chips and the SDC rates they generate. Two reasons easily explain this pessimistic expectation.

First, companies make their profit out of their product features, the properties that the public perceive as good! Performance comes first; power and energy consumption are also important; reliability (reduced failure rates) is rarely (never!) a selling point. The fear that a machine may fail even once every many months, or, even worse, it may produce an unnoticed corrupted output, is terrifying. Nobody talks about the errors their system generates.

TABLE 2. Reducing the rate of SDCs and associated costs.

Approach	Why it reduces SDCs	Why it adds cost
Better manufacturing testing	Fewer manufacturing defects will escape into production. **Fewer SDCs!**	High volume manufacturing test time is expensive; chips **cost** more.
Reduced chip variability	Marginally nominal chips are not deployed in the systems. **Fewer SDCs!**	The production yield is reduced; chips **cost** more.
Enhanced hardware defect tolerance	Mission mode defects (due to radiation and aging for example) are detected or corrected at the hardware level. **Fewer SDCs!**	Every hardware redundancy scheme comes with design, verification, area, power, performance overheads; system **costs** more.
Enhanced software error tolerance	Erroneous calculations are detected and corrected. **Fewer SDCs!**	Every software redundancy scheme comes with performance and energy overheads; system **costs** more.

Second, failure rates (including SDC rates) are very difficult to define, measure, and reproduce. When a company makes claims about its product rating in speed or power, it is reasonably possible to reproduce the setup and verify that the claim is accurate. For failure rates (or SDC rates) there is no practical way to reproduce any claim: If a company states that its chip generates one silent data corruption per year in a population of 1,000,000 chips, how can one verify the statement?

SDC-AWARENESS IN COST MODELS

The recent rise of awareness about the existence of SDCs in cloud computing is very important. Hyperscalers now require better quality chips from silicon vendors while also focusing on detecting defective chips in fleets. The entire industry is looking for solutions.

It is probably time for hyperscalers to take the rates of SDCs into consideration in cost models. A data center user may occasionally receive erroneous execution outputs when their programs run in the cloud. For a potentially "high SDC rate" data center rental time, users should deal with the integrity of the program results themselves (and thus pay less). On the other hand, a guaranteed "low SDC rate" data center time means that the provider takes care of the correctness of the delivered results and, for this reason, can charge more.

It is again a matter of cost and who takes it. It can be the chip vendor, it can be the hyperscaler, it can be the user/customer. SDC rates should be added to the cost equation of data centers on both the operator side and the user side.

SDCS FROM DATA PARALLEL ARCHITECTURES

This short essay mainly focuses on SDCs from CPU chips because disclosures mainly refer to them. But who expects things to be better in massive data-parallel architectures like a GPU or a programmable AI accelerator?[20,21] The data-centric architecture of such chips and their massive deployment for machine learning/AI systems intuitively points to huge rates of SDCs. The discussion of the text so far and the research directions that conclude the essay apply to all different chip types. If a CPU contains a multiplier like the one discussed previously, how many multipliers does a GPU, or a custom AI accelerator, have?

> Tackling the challenge of SDCs requires contributions from researchers and practitioners across the entire computing systems stack.

PRACTICAL (COST-EFFECTIVE) RESEARCH DIRECTIONS

Dependable computing has always been an exciting field. In times that technology is pushed to its limits, statistics are not on our side. The scale of computing systems of our days and our greed for speed challenges our ability to guarantee correctness of computers operations. A (nonexhaustive) list of practical (cost-effective) research directions that can contribute to our quest for the holy grail of zero-SDC systems includes:

- *Estimations for the real SDC rates (not just the number of chips that occasionally produce SDCs)*: The single defective chip in a thousand may generate zero SDCs if the defective unit is not used; it can generate millions of SDCs per day if workloads use it massively. Ideally, the SDC rate estimations must correlate them to instruction or workload classes to assist software-based tolerance.
- *Effective periodic scanning of data centers fleets*: Manufacturing testing should constantly improve to screen defective parts, but continuous screening in the field can catch complex real-world SDC scenarios that the short, focused manufacturing testing process can't. After all, in data centers, what matters is the detection of defective chips and their replacement.
- *Tolerance at the hardware and the software layers*: Tolerance refers to approaches that aim at continuing the system operation despite the existence of defects. In the case of SDCs in large computing infrastructures where massive redundancy (the common denominator of all tolerance approaches) is rarely affordable, tolerance is not a very intuitive option. When a data center operator knows a chip is defective, the only responsible action is to replace it. The only meaningful alternative would be the degraded operation of a chip with a known and contained defect in a hardware unit which is not used by workloads.

Tackling the challenge of SDCs requires contributions from researchers and practitioners across the entire computing systems stack: physical design, logic and microarchitecture design, instruction set architectures, systems software, compilers, and application software. While we certainly live in the "Golden Age of Computer Architecture" per John Hennessy's and David Patterson's words,[22] we also live in the "Golden Age of Dependable Computing" and we all need to make sure computing remains as beautiful and as powerful as it should be.

ACKNOWLEDGMENTS

Thanks to the Computing Architectures column editors for the invitation to host the article in *Computer*, and all the team members of the Computer Architecture Lab of the University of Athens that work with me in the exciting research area. Research is supported by Meta, AMD, and the OCP.

REFERENCES

1. C. Constantinescu, I. Parulkar, R. Harper, and S. Michalak, "Silent data corruption — Myth or reality?" in *Proc. IEEE Int. Conf. Dependable Syst. Netw. (DSN)*, Anchorage, AK, USA, 2008, pp. 108–109, doi: 10.1109/DSN.2008.4630077.
2. H. D. Dixit et al., "Silent data corruptions at scale," 2021, *arXiv:2102.11245*.
3. P. H. Hochschild et al., "Cores that don't count," in *Proc. Workshop Hot Topics Operating Syst. (HotOS)*, New York, NY, USA: ACM, 2021, pp. 9–16, doi: 10.1145/3458336.3465297.
4. S. Wang, G. Zhang, J. Wei, Y. Wang, J. Wu, and Q. Luo, "Understanding silent data corruptions in a large production CPU population," in *Proc. 29th Symp. Operating Syst. Princ. (SOSP)*, 2023, pp. 216–230, doi: 10.1145/3600006.3613149.
5. Y. He, M. Hutton, S. Chan, R. De Gruijl, R. Govindaraju, and N. Patil, "Understanding and mitigating hardware failures in deep learning training systems," in *Proc. 50th Int. Symp. Comput. Archit. (ISCA)*, 2023, pp. 1–16, doi: 10.1145/3579371.3589105.
6. D. Ma et al., "Dr. DNA: Combating silent data corruptions in deep learning using distribution of neuron activations," in *Proc. 29th ACM Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, 2024, pp. 239–252, doi: 10.1145/3620666.3651349.
7. "Date center documentation – Version 535.129.03(Linux)/537.70(Windows)." NVIDIA Docs. Accessed: Apr. 4, 2024. [Online]. Available: https://docs.nvidia.com/datacenter/tesla/tesla-release-notes-535-129-03/index.html
8. O. Chatzopoulos, N. Karystinos, G. Papadimitriou, D. Gizopoulos, H. D. Dixit, and S. Sankar, "Veritas – Demystifying silent data corruptions: μArch-level modeling and fleet data of modern x86 CPUs," in *Proc. IEEE Int. Symp. High-Perf. Comput. Archit. (HPCA)*, 2025, pp. 1–14.
9. "Announcing the winners of the 2022 silent data corruptions at scale request for proposals," *Meta Research*, Apr. 4, 2025. [Online]. Available: https://research.facebook.com/blog/2022/6/announcing-the-winners-of-the-2022-silent-data-corruptions-at-scale-request-for-proposals/
10. "OCP's server resilience initiative: SDC academic research awards announced!" *Open Compute Project*, Jun. 4, 2024. [Online]. Available: https://www.opencompute.org/blog/ocps-server-resilience-initiative-sdc-academic-research-awards-announced
11. "Sudhanva Gurumurthi's post." LinkedIn. Accessed: Apr. 4, 2024. [Online]. Available: https://www.linkedin.com/posts/sudhanva-gurumurthi_emerging-fault-modes-challenges-and-research-activity-7086699729228607490-Yv1z/
12. "1st IEEE RAS in data centers summit." IEEE RAS 2024. Accessed: Apr. 4, 2024. [Online]. Available: https://ieee-ras.conferences.computer.org/2024/
13. ACM SIGARCH. *ISCA'22 - Lunch and Panel on Silent Data Corruptions*. (Jul. 13, 2022). Accessed: Apr. 4, 2024. [Online Video]. https://www.youtube.com/watch?v=j1IvGZZSm3k
14. S. Hesley. *ITC Wednesday Keynote*. (Dec. 15, 2024). Accessed: Apr. 4, 2024. [Online Video]. https://www.youtube.com/watch?v=aKnUAoCPkbE
15. G. Papadimitriou, D. Gizopoulos, H. D. Dixit, and S. Sankar, "Silent data corruptions: The stealthy saboteurs of digital integrity," in *Proc. IEEE 29th Int. Symp. On-Line Testing Robust Syst. Des. (IOLTS)*, Crete, Greece, 2023, pp. 1–7, doi: 10.1109/IOLTS59296.2023.10224870.
16. T. Macieira, S. Gurumurthy, S. Gurumurthi, A. Haggag, G. Papadimitriou, and D. Gizopoulos, "Silent data corruptions in computing: understand and quantify," in *Proc. IEEE 30th Int. Symp. On-Line Testing Robust Syst. Des. (IOLTS)*, Rennes, France, 2024, pp. 1–7, doi: 10.1109/IOLTS60994.2024.10616056.
17. N. Karystinos, O. Chatzopoulos, G.-M. Fragkoulis, G. Papadimitriou, D. Gizopoulos, and S. Gurumurthi, "Harpocrates: breaking the silence of CPU faults through hardware-in-the-loop program generation," in *Proc. ACM/IEEE 51st Annu. Int. Symp. Comput. Archit. (ISCA)*, Buenos Aires, Argentina, 2024, pp. 516–531, doi: 10.1109/ISCA59077.2024.00045.
18. S. Gurumurthi, V. Sridharan, and S. Gurumurthy, "Emerging fault modes: Challenges and research opportunities," *ACM SIGARCH*, Jul. 17, 2023. [Online]. Available: https://www.sigarch.org/emerging-fault-modes-challenges-and-research-opportunities/
19. D. Gizopoulos, "SDCs: A B C," *ACM SIGARCH*, Sep. 16, 2024. [Online]. Available: https://www.sigarch.org/sdcs-a-b-c/
20. A. Vahdat and M. Lohmeyer, "Enabling next-generation AI workloads: Announcing TPU v5p and AI hypercomputer," *Google Cloud*, Dec. 7, 2023. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-tpu-v5p-and-ai-hypercomputer
21. "MTIA v1: Meta's first-generation AI inference accelerator," *AI at Meta*, May 18, 2023. [Online]. Available: https://ai.meta.com/blog/meta-training-inference-accelerator-AI-MTIA
22. J. L. Hennessy and D. A. Patterson, "A new golden age for computer architecture," *Commun. ACM*, vol. 62, no. 2, pp. 48–60, doi: 10.1145/3282307.

DIMITRIS GIZOPOULOS is a professor at the Department of Informatics and Telecommunications, University of Athens, 15784 Athens, Greece. Contact him at dgizop@di.uoa.gr.

EDITOR **NIR KSHETRI**
The University of North Carolina at Greensboro;
nbkshetri@uncg.edu

Blockchain and Stablecoins: Driving the Future of Digital Finance

Nir Kshetri, The University of North Carolina at Greensboro

Stablecoins have emerged as a transformative digital asset in global finance. With regulatory standards evolving and adoption increasing, stablecoins are reshaping financial systems, especially in regions with currency instability, while addressing challenges related to scalability, transparency, and trust.

The first stablecoin was created in 2014 to enable crypto transactions when banks, wary of anonymity and illicit activity, avoided serving crypto firms. As the crypto market expanded, trust in stablecoins grew, prompting banks and regulators to engage and consider digital currencies.[1] Consequently, in the last decade, stablecoins have made a notable impact on global finance by offering a stable and efficient means of conducting transactions and facilitating cross-border payments.[2]

The stablecoin market was valued at US$230 billion in March 2024, with projections suggesting up to US$5 trillion in assets could move into stablecoins by 2030.[3]

Unsurprisingly, leading financial institutions are adopting stablecoins. PayPal, Stripe, and Bank of America are exploring stablecoin integration.[4] Likewise, Japan's major banks—MUFG, SMBC, and Mizuho—were reported to be developing Project Pax, a cross-border payment system that uses stablecoins instead of correspondent banks, integrating SWIFT messages so that corporate clients can trigger payments conventionally without directly handling stablecoins.[5] Likewise, companies such as Tesla, an electric vehicle and clean energy company, and Uniswap, a leading decentralized cryptocurrency exchange, have started holding stablecoins in their corporate treasuries to enhance liquidity and operational efficiency.[4]

Stablecoins are especially popular in regions like Latin America and Africa, where currency instability is a major concern.[6] By expanding financial services to underbanked populations, they enhance economic inclusion

Digital Object Identifier 10.1109/MC.2025.3554937
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

and strengthen their role in the global economy[7]. This trend is reshaping the financial system in the digital era. In countries like Kenya, which have strong payment networks, stablecoins are enhancing cross-border transactions. For instance, Sling Money, powered by the Pax Dollar (USDP) stablecoin, enables fast global transfers for Kenyans.[8] While M-Pesa excels domestically, it lacks efficient international remittance options. As a result, many Kenyans turn to Sling Money for cross-border transactions. Contrary to common belief, users seek not just North-South remittances but also seamless transfers between neighboring countries. Sling Money serves this purpose, bridging gaps in regional financial flows. Sending money from Uganda to Kenya traditionally incurs high fees (5–10% of the transfer amount), takes days to process, and requires navigating complex banking networks. Sling Money's stablecoin-powered solution reduces fees to near zero, completes transfers in seconds, and simplifies the process to resemble domestic peer-to-peer payments.[9]

Research shows that in emerging economies, accessing stablecoins incurs a premium—averaging 4.7% over the U.S. dollar (USD) price (up to 30% in countries like Argentina)—resulting in an estimated US$4.7 billion premium in 2024, projected to rise to US$25.4 billion by 2027.[10] This willingness to pay a premium highlights the strong demand for stablecoins as a reliable store of value and means of transaction, especially in economies facing currency instability and capital controls.

This article examines the role of stablecoins in the global financial system, focusing on their impact on cross-border transactions and financial stability. It explores the regulatory challenges, opportunities for adoption, and the implications of stablecoins in economies with volatile currencies.

CURRENT STATE OF THE STABLECOIN INDUSTRY AND MARKET

Stablecoins are cryptocurrencies designed to maintain a stable value by pegging to assets. There are various types, including fiat-backed stablecoins tied to government currencies, asset-backed stablecoins supported by tangible assets, crypto-collateralized stablecoins linked to crypto assets, and algorithmic stablecoins, which use algorithms to maintain stability.[11] For instance, stablecoins pegged to the USD maintain a fixed US$1 value, making their market capitalization closely align with their total supply. Stablecoins are private-sector alternatives to central bank digital currencies (CBDCs).[11]

Most stablecoins use smart contracts for issuance, collateralization, and price stability, ensuring decentralized and transparent functionality. Others, like those issued by centralized entities, primarily depend on off-chain reserves while using smart contracts for transactions and automation.[12]

Foresight Ventures suggests that stablecoins are still in a transitional phase, functioning more as a bridge between fiat currencies than a complete replacement. As digital payment infrastructures develop, non-crypto users are expected to gradually adopt stablecoins.[13] The market capitalization of stablecoins, expressed in billions of USDs, is shown in Figure 1, reflecting their growth trajectory.

Figure 1 shows the market capitalization of stablecoins. As of early 2025, there were more than 200 stablecoins in circulation. As of 22 March 2025, the five biggest stablecoins by market capitalization were Tether's USDT, Circle's USDC, Ethena's USDe (USDE), Dai (DAI), and First Digital USD (FDUSD) (https://coinmarketcap.com/view/stablecoin/).

As illustrated in Figure 2, the volume of transactions enabled by stablecoins is expanding rapidly. In 2023, stablecoins facilitated US$10.8 trillion in transactions,[16] which exceeded US$27.6 trillion in 2024, surpassing Visa and Mastercard's combined volumes.[6] However, this remains small compared to the US$1.8 quadrillion processed by the traditional payments industry, as estimated by McKinsey. Stablecoins are growing but still represent a fraction of global transaction volumes.[16]

REGULATORY LANDSCAPE

The fragmented regulatory landscape has led to inconsistencies in stablecoin oversight as governments worldwide struggle to classify and regulate them effectively.[18] For instance, stablecoin adoption is growing in Africa, but regulatory frameworks are still developing. In South Africa, the Financial Sector Conduct Authority (FSCA) classifies crypto assets as financial products, though there are no specific regulations for stablecoins yet.[19]

However, stablecoin regulations and standards are emerging to ensure their stability, transparency, and security. These standards focus on key areas

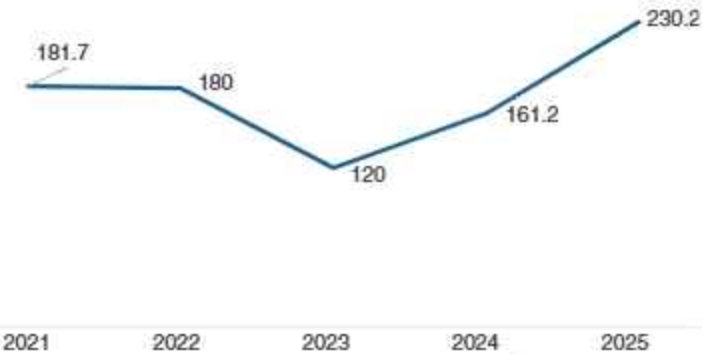


FIGURE 1. The market capitalization of stablecoins (billion US$). (Data sources: For 2021 and 2024, CoinGecko[14]; for 2022 and 2023, Jasperse and Hammer[11]; for 2025, Cooley[15].)

such as fully backed reserves, transparency through audits, clear governance and issuance guidelines, regulatory compliance [such as anti-money laundering (AML)/combating the financing of terrorism (CFT), and Know Your Customer], and consumer protection.[20] The goal is to establish a framework that supports stablecoin reliability, accountability, and user trust.[20] Major economies, such as the European Union (EU), the United States, the United Kingdom, Japan, Singapore, and the United Arab Emirates (UAE), are actively developing regulatory frameworks for stablecoins to ensure financial stability and market oversight.[3]

The EU

The EU's Markets in Crypto-Assets Regulation (MiCA), the first comprehensive regulatory framework for the crypto industry, took full effect on 30 December 2024.[20] MiCA requires stablecoins to be fully backed by liquid assets, held in separate reserve accounts to ensure user redemption at any time. Issuers must maintain transparency, regularly reporting their reserves and financial health while disclosing significant changes.[21] Under MiCA, stablecoin issuers are required to hold at least 60% of their reserve assets in European banks.[20]

European regulators mandate approval for stablecoin operations, enforcing strict financial and security compliance. Noncompliant stablecoins face delisting and restricted use, driving users toward MiCA-approved alternatives.[21] MiCA regulations ban algorithmic stablecoins, allowing only asset-backed stablecoins—e-money tokens (EMTs) and asset-referenced tokens (ARTs). EMTs must be 1:1 backed by fiat, while ARTs can be pegged to multiple assets but face stricter scrutiny. Issuers must be EU based, obtain authorization, and publish a whitepaper before offering. Major stablecoins, so-called "significant" stablecoins, are subject to enhanced oversight to ensure financial stability and protect monetary sovereignty.[22]

As of March 2025, 10 firms had received authorization to issue 15 stablecoins, officially recognized as EMTs under EU regulations.[21] As the EU's MiCA regulations took effect, Tether (USDT) and other noncompliant stablecoins were reported to be delisted from European exchanges.[21]

The United States

The lack of a thorough regulatory structure for stablecoins and digital assets in the United States has created an environment of uncertainty. Stablecoins are viewed as a key tool for expanding U.S. monetary influence, with 99% of stablecoin volume tied to the USD, promoting dollar use on global decentralized networks.[3] Treasury Secretary Scott Bessent stated that the Trump administration would prioritize using dollar-pegged stablecoins to maintain the USD's reserve currency status.[23]

It is being realized that without swift action, the United States risks falling behind in financial innovation. A regulated stablecoin market can bolster the USD's dominance in global finance. As digital economies grow, stablecoins ensure that the dollar remains the preferred currency.[3]

Two legislative proposals aim to regulate stablecoins with contrasting approaches. On 4 February 2025, Senators Bill Hagerty, Cynthia Lummis, and Kirsten Gillibrand and House Committee Chair Tim Scott introduced the Guiding and Establishing National Innovation for U.S. Stablecoins (GENIUS) Act,[24] which seeks a balanced regulatory framework, allowing smaller issuers to operate under state oversight while placing larger ones under federal supervision. It mandates 1:1 reserves, bans algorithmic stablecoins, and aims to strengthen the USD's dominance.[25]

In contrast, the Waters Bill, introduced by Rep. Maxine Waters, centralizes oversight, requiring all issuers to be federally regulated, prohibiting Big Tech from issuing stablecoins, tightening restrictions on offshore entities, and barring individuals with financial crime convictions from ownership. These bills highlight a shift from debating *whether* to regulate stablecoins to how to do so, with the outcome shaping the future of digital assets, financial inclusion, and U.S. leadership in crypto.[25]

Global stablecoin standards and the impact of U.S. and EU stablecoin regulations

As stablecoin regulations take shape in both the EU and the United States, the divergence between their regulatory approaches is becoming more apparent. Both the EU's MiCA and U.S. legislation seek to define stablecoins' role, highlighting their significance in digital assets.[26]

With U.S. and European regulation of digital assets, global crypto standards may emerge, though their origin—U.S. policy or MiCA—remains uncertain.[26] Similar to previous

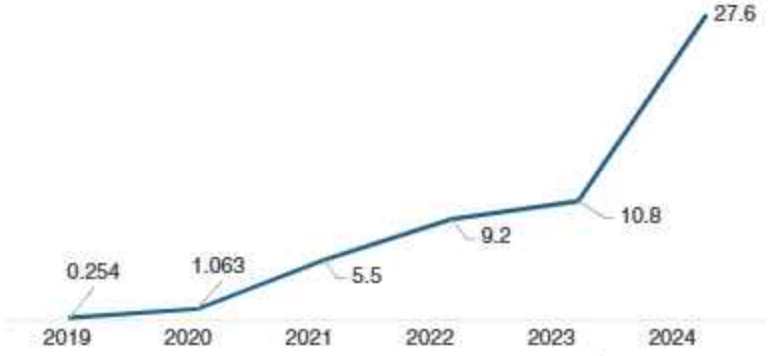


FIGURE 2. Transactions facilitated by stablecoins (trillion US$). (Data sources: For 2019, 2020, and 2021, Antonopoulos[17]; for 2022 and 2023, Cooley[16]; for 2024, Bradley[6].)

generations of innovations, stablecoin policies in the United States are more innovation friendly than in the EU. The head of the European Stability Mechanism noted that the U.S. administration has adopted a more favorable stance toward cryptocurrencies, particularly dollar-denominated stablecoins. He highlighted concerns in Europe that this shift could revive efforts by foreign and U.S. tech giants to develop large-scale payment solutions based on these stablecoins.[27]

The global dominance of USD-backed stablecoins can be attributed to the United States's market-friendly regulatory approach. As of 12 March 2025, the top 11 stablecoins by market capitalization were pegged to the USD, while the 12th-largest stablecoin, Stasis Euro, was pegged to the Euro. In terms of market capitalization, Stasis Euro ranked as the 258th largest cryptocurrency.[28]

As of March 2025, the stablecoin market exceeds US$234 billion, with Tether (USDT), pegged 1:1 to the USD, holding a dominant 61% market share (https://coinmarketcap.com/view/stablecoin/). Its success is largely due to a reserve strategy focused on U.S. Treasuries and other liquid dollar-denominated assets, which helps maintain its peg to the USD and builds market trust.[29] Tether holds around 85% of its reserves in cash and liquid assets, with U.S. Treasuries making up the majority (https://x.com/coinbureau/status/1850839929221243224). This strategy is driven by the stability and liquidity of Treasuries, which are backed by the U.S. Government and actively traded in a high-volume market. This deep liquidity ensures that the stablecoin's value is maintained without significant price fluctuations, fostering market confidence in its stability.[29]

Other stablecoins, such as USD Coin (USDC), similarly prioritize holding reserves in highly liquid low-risk assets, like U.S. Treasuries and cash in regulated financial institutions. For instance, the majority of USDC's reserve is held in the Circle Reserve Fund (USDXX), a U.S. Securities and Exchange Commission (SEC)-registered money market fund containing cash, short-dated U.S. Treasuries, and repurchase agreements. The reserve, backed by highly liquid stable assets, ensures liquidity, even under stress, with daily third-party reporting via BlackRock. The rest is held in cash at top global banks.[30]

Stablecoins like USDC and USDT thus maintain their peg by holding reserves in highly liquid assets, primarily U.S. Treasuries. This strategy ensures quick conversions to fiat during high redemption pressure, supporting their stability. The preference for U.S. Treasuries lies in their safety, liquidity, and market depth. In contrast, using European bonds or EU bank reserves could introduce higher transaction costs and systemic risks due to lower liquidity and potential localized financial instability. Diversified reserves, like those in USDC, help mitigate such risks.[29]

OPPORTUNITIES AND CHALLENGES IN STABLECOINS

Stablecoins offer significant opportunities, from enhancing financial inclusion to streamlining cross-border transactions. However, they also present challenges, including regulatory uncertainties and potential financial stability risks.

Opportunities

Stablecoins offer several advantages. They provide the ability to program money through smart contracts[7] and enable near-instant transactions, typically settling in less than a second,[31] at almost zero cost. This reduces friction for businesses and consumers compared to traditional payment systems.[32] Traditional cross-border payments, totaling US$150 trillion annually, incur US$2.5 trillion in intermediary fees, effectively taxing global commerce.[31] By reducing these costs, stablecoins enhance the efficiency of global transactions, making cross-border payments faster and more affordable.

As noted, stablecoins are transforming the global financial system by enhancing economic inclusion in developing economies, such as those in Latin America and Africa, where currency instability is a major concern. Sub-Saharan African currencies have depreciated significantly against global currencies like the USD and the pound, diminishing purchasing power. In countries like Nigeria and Zimbabwe, citizens favor saving in USDs due to the volatility of local currencies. While Bitcoin has gained popularity, stablecoins have become the preferred digital currency, particularly after the economic challenges brought on by the pandemic.[33] Stablecoins accounted for 43% of Sub-Saharan Africa's total crypto transaction volume in 2024, with Bitcoin at 18.1%.[19] In Brazil, stablecoins made up 26% of crypto purchases, driven by currency weakness and regulatory discussions. In Colombia, their use increased due to the declining peso and banking restrictions, rising by 17 percentage points in crypto portfolios. In Mexico, stablecoins gained traction, reaching 34% of crypto purchases, driven by remittances and avoiding high banking fees.[34]

Stablecoins act as a digital alternative to the USD, providing a crucial solution in economies facing foreign exchange (FX) shortages. About 70% of African countries are facing FX shortages, making it difficult for businesses to access the dollars they need. Stablecoins like USDT and USDC provide an alternative by offering easy conversion to hard dollars, bypassing traditional financial institutions. Platforms such as Yellow Card help African businesses access these stablecoins when banks can't meet their demand for USDs.[19]

Cryptocurrencies, in general, have been facilitating entrepreneurship by providing faster and more cost-effective payment solutions for businesses in developing countries. Small businesses in developing countries have found that using cryptocurrencies for payments, instead of major international

currencies like the USD or Euro, significantly improves speed and efficiency. A Nigerian vendor, sourcing products from China and the UAE, increased his profits by paying in cryptocurrency, avoiding the need to exchange naira for dollars or incur high money-transfer fees. This practical advantage has contributed to Bitcoin's growing use in these economies.[35] As noted previously, businesses in Africa are increasingly transitioning from established names like Bitcoin to stablecoins for making international payments. In February 2025, cryptocurrency platform GlobaChain announced plans to launch a stablecoin payment platform to improve cross-border transactions between Africa, Europe, and other regions. With support from Stellar, the platform aims to onboard more than 200 businesses and process US$50 million in monthly transactions.[36]

Finally, artificial intelligence (AI) is becoming a key driver in the rapid expansion of stablecoin adoption. AI enhances stablecoin stability and security by dynamically adjusting collateral reserves based on market conditions and using algorithmic auto-balancing to maintain the peg. Predictive risk management, such as sentiment analysis and transaction surveillance, allows for the early detection of potential de-pegging risks or destabilizing whale activity. Additionally, AI-driven fraud prevention measures identify wash trading and suspicious transactions, while AI-led smart contract audits detect and mitigate vulnerabilities, ensuring secure stablecoin operations and reducing risks.[4] GlobaChain leverages blockchain, AI optimization, and Circle products to eliminate delays and high fees in international payments.[36] For companies such as GlobaChain, AI-powered FX optimization helps remittance providers hedge against FX volatility and optimize stablecoin-to-fiat conversions, reducing costs, while automated compliance checks streamline AML processes and accelerate business-to-business settlements.[4] AI and blockchain converged to enhance financial management, with AI agents integrating stablecoins for optimized payments and savings. As of 11 February 2025, USDS was the third-largest stablecoin, maintaining a US$1 peg and offering an 8.75% savings rate, managed by AI-driven collateral decisions.[37] Likewise, AI trading agents like Buzz, which handled US$5.46 trillion in stablecoin transactions in 2024, are driving automated demand for programmable stablecoins.[4]

Challenges

The main challenges facing stablecoins can be understood through the Stablecoin Trilemma, which highlights the difficulty of achieving decentralization, scalability, and stability simultaneously. Achieving all three is difficult, often requiring tradeoffs. Decentralization involves distributing control among multiple parties, scalability ensures that the coin can handle high transaction volumes, and stability ensures that the coin's value remains pegged to an asset like the USD.[38]

Regarding scalability, during CBDC testing, the Federal Reserve Bank of Boston demonstrated that nonblockchain payment technology can process 10 times more transactions per second than a high-performance blockchain as the necessary ordering of transactions to prevent double-spending creates bottlenecks that limit scalability and hinder fast payments.[39] Ethereum hosts major stablecoins, like USDT, USDC, and DAI, but its scalability issues lead to high gas fees, especially during peak demand. In 2021, congestion caused transaction fees to exceed US$50, making Ethereum impractical for small payments. For example, South American freelancers receiving USDT faced US$40 fees, severely reducing their earnings and delaying access to essential funds.[40]

Looking at specific examples of stablecoins, the tradeoff between decentralization, scalability, and stability becomes clear. Tether (USDT), for example, prioritizes scalability and stability through centralization. However, this approach raises concerns about trust in the management of its reserves by the issuing entity.[38] The American watchdog Consumers' Research criticized Tether for lacking transparency on USDT reserves.[41] The watchdog expressed concerns over Tether's failure to conduct independent audits despite promises made since 2017, with the company repeatedly postponing them.[42]

Like USDT, other major stablecoins have struggled to gain customer trust because their centralization limits transparency. Many investors have reduced their stablecoin holdings due to concerns over transparency and rising interest rates offering higher returns from traditional assets. This shift has led to a decline in stablecoins' market capitalization from 2021 to 2023 (Figure 1).[11]

DAI, developed by MakerDAO, emphasizes decentralization while ensuring stability through its overcollateralized model. Unlike fiat-backed stablecoins, DAI is secured by a basket of cryptocurrencies, like Ether (ETH), locked in smart contracts. This approach offers a level of decentralization that centralized stablecoins cannot achieve, but it faces challenges. DAI's need for overcollateralization can result in inefficiencies and volatility, particularly during sharp market declines, impacting scalability.[38]

In addition to the Stablecoin Trilemma, there are other challenges that stablecoins face. As mentioned earlier, the regulatory landscape for stablecoins is evolving quickly and differs significantly around the world. Many countries lack regulatory clarity regarding digital money, and in some, such as the United Kingdom, regulated companies are prohibited from issuing stablecoins, leaving the market vulnerable to unregulated overseas operators.[43]

Regulators are also concerned that stablecoins could be exploited for money laundering or terrorist financing, necessitating stringent AML and CFT measures by issuers.[44] Illicit addresses received US$40.9 billion in cryptocurrency in 2024, with Chainalysis

projecting that the total could surpass US$51 billion as more criminal wallets are identified. Stablecoins accounted for 63% of all illicit transactions, a significant increase from previous years.[45]

Stablecoins combine the speed of tokenized transactions with price stability, making them a compelling choice for crypto payments. Stablecoins have thus rapidly evolved over the past decade, making a significant impact on global finance by enabling efficient cross-border payments and offering a stable digital alternative to traditional currencies. As adoption grows, particularly in emerging markets, stablecoins are reshaping financial systems, with applications in regions facing currency instability, such as Sub-Saharan Africa and Latin America. However, the regulatory landscape remains fragmented, with various global jurisdictions implementing different frameworks, such as MiCA in the EU and proposed legislation in the United States These regulations aim to ensure stability, transparency, and security in stablecoin operations, but challenges such as the Stablecoin Trilemma and concerns over transparency and illicit activity remain. Despite these obstacles, stablecoins continue to gain traction, and their future role in global finance will depend on how effectively these challenges are addressed and how the regulatory landscape evolves.

REFERENCES

1. J. Sahadi, "How stablecoin is different than other cryptocurrencies ... and how it's not," *CNN Bus.*, Oct. 27, 2021. Accessed: Mar. 21, 2025. [Online]. Available: https://www.cnn.com/2021/10/27/success/what-is-stablecoin-feseries/index.html
2. A. Jafri, "Emerging markets embrace stablecoins despite significant premiums," *CryptoSlate*, Aug. 21, 2024, Accessed: Mar. 21, 2025. [Online]. Available: https://cryptoslate.com/emerging-markets-embrace-stablecoins-despite-significant-premiums/
3. C. A. Makridis, "America must harness stablecoins to future-proof the dollar," *Fortune*, Mar. 21, 2025. Accessed: Mar. 22, 2025 [Online]. Available: https://fortune.com/2025/03/21/stablecoin-market-dollar-blockchain/
4. AELF, "Stablecoins are in vogue again. Here's how AI agents can help their development," *Blockster*, Mar. 7, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://blockster.com/stablecoins-are-in-vogue-again-heres-how-ai-agents-can-help-their-development
5. "Japan's big 3 banks to use stablecoins, Swift for cross border payments," *Ledger Insights*, Sep. 5, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.ledgerinsights.com/japans-big-3-banks-to-use-stablecoins-swift-for-cross-border-payments/
6. T. Bradley, "Stablecoins seek to rewire the financial system," *Forbes*, Mar. 19, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.forbes.com/sites/tonybradley/2025/03/19/stablecoins-seek-to-rewire-the-financial-system/
7. A. Aziz, "Ethereum dominates stablecoin market with $850 billion in monthly volume, led by USDC and USDT," *CoinMarketCap Acad.*, Mar. 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://coinmarketcap.com/academy/article/ethereum-dominates-stablecoin-market-with-dollar850-billion-in-monthly-volume-led-by-usdc-and-usdt
8. R. T. Watson, "Sling Money, a stablecoin-based global payment app, raises $15 million in Series A," *The Block*, Aug. 14, 2024. Accessed: Mar. 22, 2025. [Online]. Available: https://www.theblock.co/post/311223/sling-money-a-stablecoin-based-global-payment-app-raises-15-million-in-series-a
9. B. Sobrado, "Beyond crypto trading: The real-world impact of stablecoins," *Forbes*, Mar. 21, 2025. Accessed: Mar. 22, 2025. [Online]. Available: https://www.forbes.com/sites/boazsobrado/2025/03/21/beyond-crypto-trading-the-real-world-impact-of-stablecoins/
10. B. Reynolds, "The decade of digital dollars: Unlocking economic efficiency with stablecoins," BVNK, Accessed: Mar. 21, 2025. [Online]. Available: https://www.bvnk.com/report/decade-of-digital-dollars
11. Moody's Investors Service, "Stablecoins have been unstable. Why?" Moody's, Oct. 18, 2023. Accessed: Mar. 21, 2025. [Online]. Available: https://www.moodys.com/web/en/us/about/insights/data-stories/stablecoins-instability.html
12. J. Jasperse and S. Hammer, "State stablecoin regulation and the emergence of global principles," *Bus. Law Today*, Sep. 20, 2024. Accessed: Mar. 22, 2025. [Online]. Available: https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-september/state-stablecoin-regulation-emergence-global-principles/
13. E. Mitchell, "Stablecoins move twice as much as visa—And they're just warming up," *CCN*, Mar. 20, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.ccn.com/news/crypto/stablecoins-twice-visa-warming-up/#:~:text=Stablecoins%20and%20Payments-,Stablecoins%20Outpace%20Visa,%244.1%20trillion%20during%20this%20period
14. CoinGecko, "State of stablecoins: 2024," *CoinGecko*, Sep. 10, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.coingecko.com/research/publications/state-of-stablecoins-2024
15. D. Llama, "Total stablecoins market cap," *DeFi Llama*, Accessed: Mar. 21, 2025. [Online]. Available: https://defillama.com/stablecoins
16. P. Cooley, "What role do stablecoins play in the payments industry?" *Payments Dive*, Dec. 17, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.paymentsdive.com/news/

what-role-do-stablecoins-play-in-the-payments-industry/735718/
17. S. Antonopoulos, "Stabilizing stablecoins—Can regulation stimulate growth?" Guidehouse, McLean, VA, USA, 2021. Accessed: Mar. 21, 2025. [Online]. Available: https://guidehouse.com/-/media/www/site/insights/financial-services/2021/fs_stabilizingstablecoins.pdf
18. W. Park, "Stablecoins: The next big thing in crypto!" *AInvest*, Mar. 20, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.ainvest.com/news/stablecoins-big-crypto-2503/
19. Chainalysis Team, "Sub-Saharan Africa: Nigeria takes #2 spot in global adoption, South Africa grows crypto-TradFi nexus," Chainalysis, New York City, NY, USA, Oct. 2, 2024. Accessed: Mar. 17, 2025. [Online]. Available: https://www.chainalysis.com/blog/subsaharan-africa-crypto-adoption-2024/
20. Z. Vardai, "MiCA rules pose 'systemic' banking risks for stablecoins — Tether CEO," *Cointelegraph*, Oct. 28, 2024. Accessed: Mar. 17, 2025. [Online]. Available: https://cointelegraph.com/news/mica-systemic-banking-risks-stablecoin-tether-ceo
21. L. Nessi, "USDT being delisted in Europe? MiCA-compliant stablecoin alternatives to consider," *CCN*, Mar. 17, 2025. Accessed: Mar. 17, 2025. [Online]. Available: https://www.ccn.com/education/crypto/usdt-delisting-in-europe-mica-stablecoin-alternatives-for-european-users/
22. Legal Nodes Team, "The EU markets in crypto-assets (MiCA) regulation explained," Legal Nodes, Harpenden, U.K., 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://legalnodes.com/article/mica-regulation-explained
23. M. Keiser, "Gold-backed stablecoins will outcompete USD stablecoins," *Cointelegraph*, Mar. 22, 2025. Accessed: Mar. 22, 2025. [Online]. Available: https://cointelegraph.com/news/gold-backed-stablecoins-outcompete-dollar-stables-max-keiser
24. T. Orme-Claye, "Introduction of GENIUS act stablecoin bill," *Reuters*, Mar. 1, 2025, Accessed: Mar. 15, 2025. [Online]. Available: https://www.reuters.com/practical-law-the-journal/government/introduction-genius-act-stablecoin-bill-2025-03-01/
25. T. M. Evans, "What the newest crypto bills in Congress mean for stablecoins," *Forbes*, Feb. 17, 2025. Accessed: Mar. 22, 2025. [Online]. Available: https://www.forbes.com/sites/tonyaevans/2025/02/17/what-the-newest-crypto-bills-in-congress-mean-for-stablecoins/
26. R. Wolfson, "US crypto regulations vs. MiCA rules: Are global standards underway?" *Cryptonews*, Feb. 6, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://cryptonews.com/news/us-crypto-regulations-vs-mica-rules-are-global-standards-underway/
27. Reuters, "EU worries US embrace of crypto assets could impact Europe financial stability," *Reuters*, Mar. 11, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.reuters.com/markets/europe/eu-worries-us-embrace-crypto-assets-could-impact-europe-financial-stability-2025-03-10/
28. A. Borovets, "EU top-level official says the US stablecoins challenge the EU financial stability, names the remedy," *Crypto News*, Mar. 12, 2025. Accessed: Mar. 12, 2025. [Online]. Available: https://crypto.news/eu-top-level-official-says-the-us-stablecoins-challenge-the-eu-financial-stability-names-the-remedy/
29. A. Lian, "MiCA's stablecoin gamble: How Europe's bank mandate could backfire," *Int. Policy Digest*, Jan. 28, 2025. Accessed: Mar. 17, 2025. [Online]. Available: https://intpolicydigest.org/mica-s-stablecoin-gamble-how-europe-s-bank-mandate-could-backfire/
30. "Transparency & stability," Circle, Boston, MA, USA, Mar. 20, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.circle.com/transparency
31. J. Allaire, "AI and stablecoins: A pairing for a more intelligent era of online business," *World Economic Forum*, Jan. 16, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.weforum.org/stories/2025/01/stablecoin-ai-business/
32. S. Broner, "How stablecoins will eat payments, and what happens next," *a16z Crypto*, Dec. 12, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://a16zcrypto.com/posts/article/how-stablecoins-will-eat-payments/
33. Adaverse Accelerator, "Stablecoins: Africa's gateway to economic growth and global integration," *Medium*, Jul. 1, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://adaverseaccelerator.medium.com/stablecoins-africas-gateway-to-economic-growth-and-global-integration-4301ee58e6d6
34. "Stablecoins used in Latin America as store of value," *CoinsPaid Media*, Mar. 14, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://coinspaidmedia.com/news/stablecoins-used-latin-america-savings
35. N. Kshetri, *Blockchain in the Global South: Opportunities and Challenges for Businesses and Societies*, Cham, Switzerland: Springer-Verlag, 2023.
36. A. Udugba, "Globachain launches stablecoin platform for Africa-Europe cross-border payments," *BusinessDay*, Feb. 24, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://businessday.ng/bd-weekender/tech/article/globachain-launches-stablecoin-platform-for-africa-europe-cross-border-payments/
37. B. McGleenon, "Stablecoins key to new era of AI-driven financial automation, says Sky Protocol's Rune Christensen," *The Block*, Feb. 11, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://www.theblock.co/post/339948/stablecoins-key-to-new-era-of-ai-driven-financial-automation-says-sky-protocols-rune-christensen
38. Komodo Team, "What is the stablecoin trilemma and how does it work?" *Komodo Platform*, Oct. 29, 2024. Accessed: Mar.

21, 2025. [Online]. Available: https://komodoplatform.com/en/academy/what-is-the-stableco in-trilemma-and-how-does-it-work/
39. M. Adachi et al., "Stablecoins' role in crypto and beyond: Functions, risks and policy," European Central Bank, Frankfurt am Main, Germany, 2022. Accessed: Mar. 21, 2025. [Online]. Available: https://www.ecb.europa.eu/press/financial-stability-publications/macroprudential-bulletin/html/ecb.mpbu202207_2~836f682ed7.en.html
40. B. Shell, "The risks of current stablecoin networks: Why bitcoin and lightning offer a safer alternative," *Voltage Cloud*, Sep. 9, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.voltage.cloud/blog/the-risks-of-current-stablecoin-networks-why-bitcoin-and-lightning-offer-a-safer-alternative
41. "Tether: Refusal to open its books puts consumers at risk – Could tether be the next FTX?" Sep. 2024. Accessed: Mar. 21, 2025. [Online]. Available: http://tetherwarning.com/wp-content/uploads/2024/09/CW_Tether-Report.pdf
42. P. Jha, "Tether hit with scathing report over USDT reserve transparency," *CCN*, Sep. 13, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.ccn.com/news/crypto/tether-usdt-reserve-transparency-concerns/
43. J. Fry, "The evolution of stablecoins," *Medium*, Feb. 13, 2025. Accessed: Mar. 21, 2025. [Online]. Available: https://jonnyfry175.medium.com/the-evolution-of-stablecoins-02f166ba522c
44. Merkle Science, "Stablecoin regulation: Addressing risks and compliance challenges," *Merkle Sci.*, 2024. Accessed: Mar. 21, 2025. [Online]. Available: https://www.merklescience.com/blog/stablecoin-regulation-addressing-risks-and-compliance-challenges
45. R. Karim, "Stablecoins: The new epicentre of crypto fraud," International Compliance Association, Birmingham, U.K., Mar. 3, 2025, Accessed: Mar. 21, 2025. [Online]. Available: https://www.int-comp.org/insight/stablecoins-the-new-epicentre-of-crypto-fraud/

NIR KSHETRI is a professor of management at the Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC 27412 USA. Contact him at nbkshetr@uncg.edu.

IEEE Computer Graphics and Applications bridges the theory and practice of computer graphics. Subscribe to *CG&A* and

- stay current on the latest tools and applications and gain invaluable practical and research knowledge,
- discover cutting-edge applications and learn more about the latest techniques, and
- benefit from *CG&A*'s active and connected editorial board.

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3562023

Get Published in the *IEEE Open Journal of the Computer Society*

Get more citations by publishing with the *IEEE Open Journal of the Computer Society*

Your research on computing and informational technology will benefit from 5 million unique monthly users of the *IEEE Xplore*® Digital Library. Plus, this journal is fully open and compliant with funder mandates, including Plan S.

Submit your paper today!
Visit www.computer.org/oj to learn more.

IEEE COMPUTER SOCIETY | IEEE

Digital Object Identifier 10.1109/MC.2025.3562104

Shaping the Future Through Artificial Intelligence Standardization Efforts

Jyotika Athavale, Synopsys
Richard Tong, NEOLAF Inc.

At the IEEE Computer Society, we see firsthand the challenges and opportunities that artificial intelligence presents, and are partnering with industry stakeholders, academia, and government bodies to create consensus-based standards.

Our article in the April 2025 issue[1] focused on key standardization initiatives related to functional safety across application domains. Today, artificial intelligence (AI) technologies are making significant strides in tackling key issues in automated vehicle operations. The early obstacles encountered in rolling out automated driving technology highlight the vital need to emphasize safety at every stage of progress. Addressing risks in the complex realm of diverse functionalities in real-world scenarios requires harnessing cutting-edge solutions and reaching consensus across the industry on the best safety practices. As a result, considerable effort has been devoted to defining safety standards for AI-powered vehicle functions and automated driving systems as a whole.

As we know, the IEEE Computer Society, in collaboration with the IEEE Standards Association, is a global leader in setting standards for engineering and technology. It partners with industry stakeholders, academia, and government bodies to create consensus-based standards. Within the context of standards development related to AI technology topics, two key committees are involved:

- *Functional Safety Standards Committee (FSSC):* This committee focuses on functional safety in areas

Digital Object Identifier 10.1109/MC.2025.3557485
Date of current version: 29 May 2025

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

EDITOR **JYOTIKA ATHAVALE**
SYNOPSYS; jyotika@synopsys.com

like automotive, industrial, and avionics. It develops standards related to AI safety architectures, methodologies, and tools for automated driving functions, such as IEEE P2851 and others.

- *Artificial Intelligence Standards Committee (C/AISC):* This committee develops standards for AI governance, machine learning, algorithms, and data usage. Its work includes creating guidelines for evaluating the explainability and robustness of AI models, particularly in areas like image recognition.

The pace of innovation in AI is staggering, and we want to share how AISC is helping shape this future. We've seen firsthand the challenges and opportunities that AI presents, and we are convinced that a balanced approach to standardization is key to ensuring that these technologies are both groundbreaking and responsibly managed.

INTRODUCING AISC AND ITS PORTFOLIO

Established in 2020, AISC has rapidly become a pivotal force in creating robust standards that guide every aspect of AI, from foundational models and agents to domain-specific applications. AISC has assembled a global network of experts from academia, industry, and government. Our leadership team is committed to transforming how AI is standardized and deployed. Our officers include: Chair Richard Tong, Vice Chair Jeanine DeFalco, Secretary Randy Soper, and Program Manager Christy Bahn. Our portfolio covers more than 60 standards, guidelines, and recommended practices and spans several crucial areas:

- *Foundation Standards:* These establish the fundamental architectures and terminologies essential for building AI systems.
- *Governance Standards:* These are focused on risk, safety, trustworthiness, and ethical considerations.
- *Technology Stack Standards:* These target the technical underpinnings that enable seamless implementation and integration.
- *Agent and Domain Application Standards:* These address the rapidly evolving field of AI agents and sector-specific use cases.

Our vision is not only to facilitate wide deployment, adoption, and application of AI, but also to ensure that AI innovation happens in an equitable, responsible, and trustworthy manner.

INTEROPERABILITY: THE ENGINE OF INNOVATION

One of the cornerstones of our work at AISC is enhancing the interoperability, efficiency, and interconnectedness of AI systems. Today, AI isn't a monolithic entity: It's a dynamic ecosystem where foundational models, AI agents, and specialized applications interact continuously. Without a unified approach to implementation, these systems risk becoming isolated silos that inhibit progress.

Take, for example, our efforts around IEEE 3652.1 Federated Machine Learning (published) and the P3394 LLM Agent Interface Standard (being developed). IEEE 3652.1 defines a framework for federated machine learning that enables collaborative model development on distributed data, while safeguarding privacy, security, and compliance. P3394 is designed to define the protocol for AI agents to communicate seamlessly with one another and with external systems. Standards like these are crucial because they allow diverse AI components—from health-care applications to educational platforms—to work together, fostering an environment where innovation isn't hindered by technical fragmentation.

We believe that a cohesive, interoperable AI ecosystem is the bedrock upon which future advancements will be built. When we standardize the ways in which AI systems interact, we lay the groundwork for a collaborative technological landscape. This means that when a new AI innovation emerges, it doesn't have to reinvent the wheel in terms of integration; it can plug into a preexisting, reliable framework. The result is a more efficient development process and a faster path to real-world applications that benefit society as a whole.

GOVERNANCE: BALANCING RISK AND REWARD

While technological interoperability drives innovation, robust governance ensures that such progress is sustainable and responsible. The conversation around AI is often polarized, caught between the excitement of innovation and the fear of unbridled technological risk. This dichotomy is particularly evident in regulatory environments. The European Union's AI Act, much like the General Data Protection Legislation (GDPR) in its early days, serves as a stark reminder of how overregulation can inadvertently stifle the very progress it seeks to protect.

At AISC, our governance initiatives focus on establishing standards that help define and mitigate risks without imposing draconian restrictions. For instance, our P3396 standard on AI risk, trustworthiness, safety, and responsibility provides a framework that not only evaluates potential risks but also offers clear guidelines for addressing them and helps AI developers and operators to be protected from unnecessary burden. Alongside this, standards like P3417 for differential privacy and P3445 for privacy engineering empower organizations to integrate robust security and ethical practices right from the design stage.

We advocate "first-principles thinking" in AI governance. Instead of

letting fear drive regulatory overreach, we should focus on distilling the essence of risk: identifying its core elements and addressing them directly. This approach not only encourages innovation but also builds public trust. It's about ensuring that AI systems are developed with a deep understanding of both their potential and their limitations, so that they can be safely integrated into our daily lives without sacrificing ethical or societal values.

BRIDGING THE TWO PILLARS: A UNIFIED STRATEGY FOR AI

The real challenge—and opportunity—lies in harmonizing these two pillars of interoperability and governance. It's not enough to develop standards that allow AI systems to interact seamlessly; we must also ensure that this interaction happens within a framework of accountability and ethical responsibility. The ultimate goal of AI standardization should be to create a landscape where innovation is encouraged, risks are managed at their core, and regulatory measures are proportionate and thoughtful.

This balanced approach is what sets AISC apart: It's about creating a synergy between the technical and the regulatory, between innovation and responsibility.

OUR CALL TO ACTION

The work of AISC is not just about designing, writing, and promoting standards; it's about fostering a global dialogue on how we want our future with AI to look. We are at a pivotal moment where our decisions can shape the trajectory of AI development for generations. We urge industry leaders, policymakers, researchers, and the broader public to engage with our efforts. Whether you're providing feedback on draft standards, collaborating on research, or simply staying informed about the latest developments, your participation is crucial.

In the face of rapid technological change, our commitment at AISC is to ensure that AI is both a catalyst for innovation and a tool for societal good. By embracing interoperability, we unlock the full potential of AI. By enabling informed and balanced governance, we build a foundation of trust and accountability. Together, these efforts will pave the way for a future where AI is seamlessly integrated into every aspect of our lives: efficient, responsible, and truly transformative for the benefit of humanity.

REFERENCE

1. J. Athavale and D. Galpin, "Functional safety standards: IEEE P2851 road map," *Computer*, vol. 58, no. 4, pp. 158–160, Apr. 2025, doi: 10.1109/MC.2025.3536796.

JYOTIKA ATHAVALE is the director of RAS architecture at Synopsys, Sunnyvale, CA 94085 USA. Contact her at jyotika@synopsys.com.

RICHARD TONG is with NEOLAF, Inc., La Puente, CA 91746 USA. Contact him at richard.tong@ieee.org.

IT Professional
TECHNOLOGY SOLUTIONS FOR THE ENTERPRISE

CALL FOR ARTICLES

IT Professional seeks original submissions on technology solutions for the enterprise. Topics include

- emerging technologies,
- cloud computing,
- Web 2.0 and services,
- cybersecurity,
- mobile computing,
- green IT,
- RFID,
- social software,
- data management and mining,
- systems integration,
- communication networks,
- datacenter operations,
- IT asset management, and
- health information technology.

We welcome articles accompanied by web-based demos. For more information, see our author guidelines at www.computer.org/itpro/author.htm.

WWW.COMPUTER.ORG/ITPRO

IEEE COMPUTER SOCIETY IEEE

Digital Object Identifier 10.1109/MC.2025.3562105

EDITORS
NORITA AHMAD American University of Sharjah; nahmad@aus.edu
PREETI CHAUHAN IEEE Reliability Society; preeti.chauhan@ieee.org

Redefining Human Resource Practices With AI Agents and Agentic AI: Automated Compliance and Enhanced Productivity

Nir Kshetri, University of North Carolina at Greensboro

This article analyzes how agentic artificial intelligence is revolutionizing human resource management through automated workflows, enhanced decision making, and improved employee experiences while addressing implementation challenges like security risks, regulatory compliance, and workforce adoption.

Digital Object Identifier 10.1109/MC.2025.3559745
Date of current version: 29 May 2025

Agentic artificial intelligence (AI) is revolutionizing human resource management (HRM) by automating routine tasks such as recruitment, thereby streamlining workflows, minimizing bias, and enhancing overall efficiency.[1] Additionally, it uses predictive analytics to assess turnover risks, enabling proactive intervention and cost savings. This technology is also likely to boost employee satisfaction and improves retention strategies.[1] Agentic AI thus enables strategic HRM with tailored, goal-oriented services.[2]

Agentic AI allows users to define a task, which the AI autonomously breaks into steps, designs workflows, and selects appropriate AI or IT services for execution.[3] As an illustrative case, we can consider a task to improve employee retention. An AI HR agent autonomously streamlines HR processes by identifying improvement strategies, generating task lists, and collecting relevant data, from analyzing exit interviews

0018-9162 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

to benchmarking industry best practices. It systematically stores insights in a centralized knowledge base, refining its approach through iterative feedback loops. For instance, after conducting employee satisfaction surveys, the agent cross-references findings with turnover metrics, updates retention strategies, and adapts future actions based on outcomes. This closed-loop automation enhances both operational efficiency and evidence-based decision-making in HR.[4]

Agentic AI solutions benefit not only large enterprises but also small and medium-sized enterprises (SMEs), enhancing operational efficiency and competitiveness across the board. SMEs can enhance their HR operations by implementing agentic AI, which automates and refines various HR functions. In talent acquisition, AI agents can autonomously screen resumes, evaluate candidate qualifications, and generate shortlists of top applicants, thereby expediting the recruitment process and minimizing manual effort. Additionally, agentic AI can streamline onboarding by guiding new hires through necessary documentation and training modules, ensuring a smooth integration into the company.[5]

Agentic AI solutions benefit not only large enterprises but also small and medium-sized enterprises, enhancing operational efficiency and competitiveness across the board.

Agentic AI not only automates tasks to boost employee productivity but also offers job seekers personalized career guidance, enhancing their empowerment beyond what generative AI provides. Generative AI assists by generating a cover letter based on a provided résumé and job description. In contrast, agentic AI autonomously creates a LinkedIn account, drafts applications, applies to all suitable jobs, and provides a secured record of its actions.[6]

This article explores how agentic AI is revolutionizing HR, automating workflows, enhancing decision making, and improving employee experiences, while addressing key challenges like security, bias, and adoption resistance. It highlights real-world implementations, regulatory impacts, and future-ready strategies for ethical AI integration in HRM.

AGENTIC AI SOLUTIONS FROM LEADING HR TECHNOLOGY COMPANIES

HR technology vendors are increasingly integrating agentic AI into their solutions to enhance various human resource functions and streamline operations (Table 1). This trend signifies a major technological shift in HR practices. These innovations include AI agents that coordinate tasks across various systems and departments, manage AI agents throughout the enterprise, and automate routine HR tasks, allowing professionals to focus on more strategic work.

TABLE 1. Agentic AI and AI agents developed by leading HR technology companies and their applications.

Company	Agentic AI/AI agents	HRM function	Functionality
ServiceNow	In AI Agent Studio, the AI Agent Orchestrator serves as the central coordinator, managing workflows across multiple AI agents and enterprise systems.[27]	Onboarding	Automatically delivers pertinent information to new employees and addresses their inquiries, streamlining the onboarding process.[28]
Workday	Recruiting Agent	Recruitment	Enhances talent acquisition by automating tasks such as job description creation, candidate sourcing, and interview scheduling; integrates with platforms like Microsoft Teams for streamlined communication.[29]
UKG	Bryte AI Agents	Employee training and development	Proactively solve problems and make recommendations through multistep processes that learn and improve over time, assisting in managing promotions and identifying impacts of new tax regulations.[7]
Oracle	Career Planning Guide, Performance and Goals Assistant	Career development, performance management, learning and development	Recommends career paths tailored to employees' skills and experience; monitors goal progression, offering timely suggestions to ensure employees remain aligned with their performance objectives ahead of evaluations.[30]

ServiceNow has launched the AI Agent Orchestrator to coordinate specialized AI agents across tasks, systems, and departments, along with thousands of prebuilt agents for HR, plus the AI Agent Studio for creating custom agents.[7] The recruitment sector is at the forefront of leveraging advanced AI agents.[8] Workday offers AI-powered agents to improve HR processes, including recruiting and succession planning, and has introduced the Agent System of Record to manage these agents effectively.[7] Oracle has developed AI agents within its Cloud HCM to streamline HR processes, assisting employees with career development, time-off requests, and workforce analytics, while providing HR teams with centralized data insights.[9]

Traditional HR systems, primarily databases for payroll and compliance, are thus evolving into integrated talent intelligence systems through the incorporation of intelligent agents into core human capital management platforms.[10] Beyond the HR technology providers highlighted in Table 1, numerous other enterprises have deployed agentic AI solutions to automate and optimize HRM functions. NEC launched agentic AI to automate talent management decisions. The system conducts comprehensive internal and external data searches to optimize these processes.[3] LinkedIn's 2024 release of the Hiring Assistant marked a shift toward AI-driven automation in talent acquisition.[8]

ENHANCING HR SELF-SERVICE AND EMPLOYEE EXPERIENCES

Employee experience covers all workplace interactions, from onboarding to daily work and advancement, impacting engagement and retention.[11] Fragmented HR systems often lead to employee frustration, but integrating HR self-service can enhance the employee experience by providing accessible information and empowering employees.[12] Modern employees expect user-friendly technologies and tend to reject those they dislike. Demonstrating a dedication to enhancing employee experience, 87% of HR managers aim to raise their HR tech budgets in 2024.[16]

In general, AI-driven systems enable employees to swiftly access information via simple queries, eliminating the need to navigate multiple platforms or submit service requests.

In general, AI-driven systems enable employees to swiftly access information via simple queries, eliminating the need to navigate multiple platforms or submit service requests. This enhances satisfaction and reduces operational costs, streamlining HR processes while improving user experience.[16] Building upon earlier AI, agentic AI automates tasks and streamlines workflows, providing a smoother, more engaging technological experience.[16] AI agents such as Galileo, Microsoft Copilot, Workday Assistant, and Eightfold's AI agent streamline HR operations by automating tasks, improving employee experiences, and reducing integration efforts for HR teams.[13] Integrating AI agents into HR systems will eliminate the need for employees to navigate multiple platforms, significantly enhancing the HR tech experience.[10]

Additionally, AI agents enhance employee experiences by personalizing training and career development recommendations, assigning role-specific modules to new hires to equip them with necessary skills.[1] For instance, ZBrain's Training Module Assignment Agent analyzes job roles and employee data to deliver customized and efficient learning experiences.[4]

Agentic AI also enhances employee experiences by ensuring employees maximize their benefits, such as paid vacation. This personalization fosters a positive employee experience. Traditional HR chatbots answer basic questions like available vacation days. In contrast, AI agents not only provide this information but also guide employees through time-off requests and assist in booking travel arrangements. Seamlessly integrating into HR systems, these agents offer a dynamic and proactive approach to workforce management.[17]

BOOSTING OPERATIONAL EFFICIENCY AND PRODUCTIVITY

Agentic AI is transforming business operations by automating complex tasks and integrating AI-driven processes across multiple systems, leading to significant efficiency and productivity gains across various industries.[10] Accenture deploys AI-powered scheduling assistants (Agentforce) to optimize employee productivity through automated work planning, priority identification, and office time utilization.[14] Agentic AI tools also automate compliance processes by monitoring policy adherence, flagging violations, and providing real-time guidance, dramatically reducing manual oversight while improving accuracy. For instance, ZBrain's compliance agent autonomously audits financial transactions against corporate policies, detecting noncompliant activity with precision to mitigate risks and operational inefficiencies.[4]

AI-powered chatbots have transformed HR by streamlining employee queries and reducing routine tasks. Now, AI agents promise even greater autonomy and operational efficiency.[17] For instance, Chipotle cut hiring time by 75% using Paradox's AI agent, enabling fully automated

hiring; some new hires even mistake the bot ("Amelia") for a human.[8]

AI agents with generative capabilities enhance HR modules, improving decision making and operational efficiency. This transformation enables HR systems to proactively offer deeper insights and automate functions like talent management and recruitment.[10] AI-powered scheduling tools can boost efficiency, with some leaders predicting they will enable widespread four-day workweeks.[18]

WORKPLACE SAFETY AND COMPLIANCE

HRM researchers have found that employee outcomes in high-risk workplaces lacking safety measures and compliance are adversely affected, leading to increased injuries, decreased job satisfaction, and higher turnover rates.[15] HRM's integration of safety culture into organizational policies shapes employee values and behaviors toward proactive safety practices.[16]

Despite engineering advancements that have reduced workplace hazards, employee safety remains a significant concern. While the physical environment plays a role, human behavior is a critical factor in most workplace accidents and injuries. Risky actions, failures in detecting hazards, and a lack of proactive safety measures contribute to these incidents. Implementing behavior-based safety programs can address these issues by systematically observing and modifying unsafe behaviors, thereby enhancing overall workplace safety.[17] In 2023, private industry employers reported approximately 2.6 million nonfatal workplace injuries and illnesses. Additionally, there were 5,283 fatal work-related injuries across all sectors.[18]

> Agentic AI also enhances employee experiences by ensuring employees maximize their benefits, such as paid vacation.

AI boosts workplace safety by predicting hazards via real-time data analysis, monitoring compliance, and preemptively mitigating risks ("Exhibit 1: From Detection to Action—viAct's Integrated Artificial Intelligence System for Real-Time Construction Risk Management"). In construction, AI agents analyze video feeds to detect unsafe behaviors (for example, missing personal protective equipment) and historical data to forecast dangers, while adaptive learning improves responses over time. For example, safety AI agents use the Internet of Things (IoT) and cameras to automate incident reporting and enforce protocols, ensuring a safer, compliant worksite.[19]

Occupational safety AI agents surpass basic chatbots by operating as adaptive, intelligent systems. Leveraging large language models (LLMs)

EXHIBIT 1: FROM DETECTION TO ACTION—viAct'S INTEGRATED ARTIFICIAL INTELLIGENCE SYSTEM FOR REAL-TIME CONSTRUCTION RISK MANAGEMENT

viAct's artificial intelligence (AI)-powered safety system for construction sites implements a three-phase evolution: predictive AI detects hazards through data analysis, generative AI formulates customized risk solutions, and agentic AI (viGent) executes autonomous safety interventions. This integrated approach progressively transforms passive monitoring into proactive protection, enhancing both worker safety and worksite efficiency through intelligent automation. The system's adaptive capabilities enable continuous improvement in hazard prevention and compliance enforcement.[25]

viAct's integrated agentic AI system (viGent) transforms construction safety through multilayered protection: computer vision detects accidents and unsafe behaviors (slips, falls, and personal protective equipment violations), while AI video analytics enable real-time monitoring via existing CCTV. The system's predictive capabilities analyze historical/real-time data to preempt risks: during emergencies, it autonomously maps evacuation routes and alerts personnel. For vehicle operations, viGent prevents collisions by dynamically adjusting forklift paths and logging near-miss data to optimize traffic flow. In confined spaces, it monitors environmental thresholds and worker duration, triggering automatic evacuations when hazards exceed limits. This centralized platform provides safety officers with comprehensive oversight while enabling immediate AI-driven interventions that surpass human response capabilities.[25]

and IoT integration, they dynamically analyze real-time data to predict hazards and enact proactive responses, like instantly alerting workers of slippery floors near machinery while rerouting foot traffic and notifying supervisors.[20] In the construction industry, for instance, agentic AI serves as a valuable tool for risk management by autonomously analyzing site hazards, optimizing workflows, and facilitating human-like decision making. This technology enhances safety protocols and operational efficiency, leading to more secure and productive construction environments.[21]

CHALLENGES AND WAYS FORWARD

The rise of AI agents offers significant automation benefits in HR, but it also introduces several challenges. The rapid proliferation of agentic AI without adequate governance can lead to inconsistent behaviors and decision making, complicating regulatory compliance and alignment with business objectives. For instance, the EU AI Act regulates all AI applications, including agentic AI, by risk level to ensure ethical, transparent, and accountable deployment. The EU AI Act mandates strict data protection for agentic AI, requiring data minimization and user control over personal data. High-risk applications must undergo bias assessments, rigorous testing, and external review before deployment.[22]

Agentic AI risks enabling adaptive "smart malware" that autonomously evades security defenses.[23] In HRM, agentic LLMs handling sensitive employee and candidate data present major security risks. Their extensive access to personal information and system controls makes them attractive targets for cyberattacks.[24]

Employee resistance to AI adoption persists, particularly in HR and administrative roles, where fears of job displacement remain prevalent.[27] Some employees remain hesitant to adopt AI agents, concerned about potential role displacement.[18] Despite proposals to integrate AI agents as "colleagues," HR professionals have largely rejected this approach. For example, Lattice's 2024 attempt to incorporate virtual workers into org charts sparked immediate backlash, forcing the company to abandon the initiative within days.[25]

Agentic AI's success hinges on high-quality, unbiased data; inaccuracies can lead to biased outcomes and erode trust. Ensuring data integrity is crucial to prevent reinforcing existing biases and to support effective AI adoption in HRM.[26]

To address these challenges, organizations must implement strong safeguards, such as clear governance frameworks, rigorous data controls, and enhanced cybersecurity measures. Human oversight remains essential to ensure AI aligns with ethical standards and business objectives.

Agentic AI is transforming HR by automating workflows, enhancing decision making, and improving employee experiences, from recruitment to safety compliance. While it boosts efficiency and productivity, challenges like data security, bias, and employee resistance require robust governance and human oversight. By balancing innovation with ethical safeguards, organizations can harness agentic AI's full potential to create strategic, adaptive, and human-centric HR systems for the future.

REFERENCES

1. "Agentic AI and its impact on HR: Shift in the future of work," *Fresh Thinking*, Accessed: Jan. 15, 2025. [Online]. Available: https://www.freshthinking.ie/blog/freshthinking-suite/agentic-ai-and-its-impact-on-hr-shift-in-the-future-of-work/
2. Oracle Corporation, "Oracle AI agents for oracle cloud HCM," Oracle, Austin, TX, USA, 2024. Accessed: Jan. 17, 2025. [Online]. Available: https://www.oracle.com/a/ocom/docs/applications/human-capital-management/ai-agents-for-oracle-cloud-hcm.pdf
3. "NEC develops Agentic AI to boost productivity through automation of advanced specialized tasks," *NEC Newsroom*, Nov. 27, 2024. [Online]. Available: https://www.nec.com/en/press/202411/global_20241127_01.html
4. A. Takyar. "AI agents for HR: Use cases, components, benefits, capabilities and implementation." LeewayHertz. Accessed: Jan. 16, 2025. [Online]. Available: https://www.leewayhertz.com/ai-agents-for-hr/#key-components-hr
5. A. Sankaran, "How small and medium businesses can take advantage of the emerging agentic AI era," *Forbes*, Apr. 1, 2025. Accessed: Apr. 1, 2025. [Online]. Available: https://www.forbes.com/councils/forbesbusinesscouncil/2025/04/01/how-small-and-medium-businesses-can-take-advantage-of-the-emerging-agentic-ai-era/
6. K. Köse, "Yes, you will be replaced: An agentic AI manifesto," *Forbes Technol. Council*, Jan. 28, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2025/01/28/yes-you-will-be-replaced-an-agentic-ai-manifesto/
7. S. Boese, "Agentic AI: What HR must know about the next evolution of HR tech," *Human Resour. Executive*, Mar. 12, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://hrexecutive.com/agentic-ai-what-hr-must-know-about-the-next-evolution-of-hr-tech/
8. S. Forsdick, "What does the advent of AI agents mean for HR?" *Raconteur*, Nov. 12, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.raconteur.net/future-of-work/hr-ai-agent-impact
9. A. DeRose, "Oracle releases new agentic AI features for HR," *HR Brew*, Feb. 7, 2025, Accessed: Apr. 2, 2025. [Online]. Available: https://www.hr-brew.com/stories/2025/02/07/oracle-releases-new-agentic-ai-features-for-hr

10. R. Maurer, "The AI agents are coming," *SHRM*, Oct. 21, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.shrm.org/topics-tools/news/technology/the-ai-agents-are-coming
11. L. Soon, "The rise and impact of agentic AI: Transforming the human experience," *CMSWire*, Nov. 26, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.cmswire.com/customer-experience/the-rise-and-impact-of-agentic-ai-transforming-the-human-experience/
12. R. Maurer, "AI-powered transformation will lead the shift to a more productive future," *SHRM*, Nov. 13, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.shrm.org/topics-tools/news/technology/ai-powered-transformation-will-lead-the-shift-to-a-more-productive-future
13. HRM Asia Newsroom, "Josh Bersin's top AI headline for 2025? It's all about the agents," *HRM Asia*, Sep. 27, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://hrmasia.com/josh-bersins-top-ai-headline-for-2025-its-all-about-the-agent
14. "AI agents and productivity: Revolutionizing workflow efficiency," *AI Today*, Feb. 7, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://aitoday.com/artificial-intelligence/ai-agents-and-productivity-revolutionizing-workflow-efficiency/
15. T.-V. Vu, T. Vo-Thanh, H. Chi, N. P. Nguyen, D. V. Nguyen, and M. Zaman, "The role of perceived workplace safety practices and mindfulness in maintaining calm in employees during times of crisis," *Human Resour. Manage.*, vol. 61, no. 3, pp. 315–333, 2022, doi: 10.1002/hrm.22101.
16. A. Kellner, K. Townsend, R. Loudoun, and A. Wilkinson, "High reliability human resource management (HRM): A system for high risk workplaces," *Human Resour. Manage. J.*, vol. 33, no. 1, pp. 170–186, 2023, doi: 10.1111/1748-8583.12424.
17. M. T. Ford, and L. E. Tetrick, "Safety motivation and human resource management in North America," *Int. J. Human Resour. Manage.*, vol. 19, no. 8, pp. 1472–1485, 2008, doi: 10.1080/09585190802200231.
18. "Injuries, illnesses, and fatalities," U.S. Dept. of Labor, Bureau of Labor Statist., Washington, DC, USA. Accessed: Apr. 1, 2025. [Online]. Available: https://www.bls.gov/iif/
19. D. Vincent. "AI agents in construction: Revolutionizing the industry," Mastt. Accessed: Apr. 2, 2025. [Online]. Available: https://www.mastt.com/guide/ai-agents
20. S. Ali, "Why AI agents outperform chatbots as AI safety compliance tools," *viAct*, Jan. 11, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.viact.net/post/why-ai-agents-outperform-chatbots-as-ai-safety-compliance-tools
21. "Agentic AI in construction: The AI-powered ally in risk management," *viAct*, Dec. 23, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.viact.ai/post/agentic-ai-in-construction-the-ai-powered-ally-in-risk-management
22. M. C. Borrelli, and S. Musch. "How to use agentic AI in line with the EU AI act." CX Network. Accessed: Apr. 2, 2025. [Online]. Available: https://www.cxnetwork.com/artificial-intelligence/articles/how-to-use-agentic-ai-in-line-with-the-eu-ai-act
23. T. Coshow. "Intelligent agents in AI really can work alone. Here's how." Gartner. Accessed: Apr. 2, 2025. [Online]. Available: https://www.gartner.com/en/articles/intelligent-agent-in-ai
24. E. Kron· "Five privacy concerns around agentic AI," *SC World*, Feb. 19, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://www.scworld.com/perspective/five-privacy-concerns-around-agentic-ai
25. J. Peters, "This HR company tried to treat AI bots like people — It didn't go over well," *The Verge*, July 15, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://www.theverge.com/2024/7/15/24199054/lattice-digital-workers-ai
26. HRM Asia Newsroom, "The great AI-in-HR balancing act: Finding your organisation's way," *HRM Asia*, Feb. 26, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://hrmasia.com/the-great-ai-in-hr-balancing-act-finding-your-organisations-way/
27. J. Zhang and T. Sage, "Agentic AI (AI Agent) development guidelines and use cases (hands-on experience)," *ServiceNow Community*, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://www.servicenow.com/community/in-other-news/agentic-ai-ai-agent-development-guidelines-and-use-cases-hands/ba-p/3206822
28. FROX AG, "ServiceNow agentic AI: Use cases and areas of application," *frox.ch*, Feb. 12, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://www.frox.ch/en/newsroom/blog-articles/servicenow-agentic-ai-use-cases/
29. "Workday announces new AI agents to transform HR and finance processes," *Workday*, Sep. 17, 2024. Accessed: Apr. 2, 2025. [Online]. Available: https://newsroom.workday.com/2024-09-17-Workday-Announces-New-AI-Agents-to-Transform-HR-and-Finance-Processes
30. M. Brue, "Inside Oracle's new AI agents for human capital management," *Forbes*, Feb. 5, 2025. Accessed: Apr. 2, 2025. [Online]. Available: https://www.forbes.com/sites/moorinsights/2025/02/05/inside-oracles-new-ai-agents-for-human-capital-management/

NIR KSHETRI is a professor of management in the Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC 27412 USA. Contact him at nbkshetr@uncg.edu.

IEEE COMPUTER SOCIETY CAREER CENTER

Career Accelerating Opportunities

Explore new options—upload your resume today

careers.computer.org

Changes in the marketplace shift demands for vital skills and talent. The IEEE Computer Society Career Center is a valuable resource tool to keep job seekers up to date on the dynamic career opportunities offered by employers.

Take advantage of these special resources for job seekers:

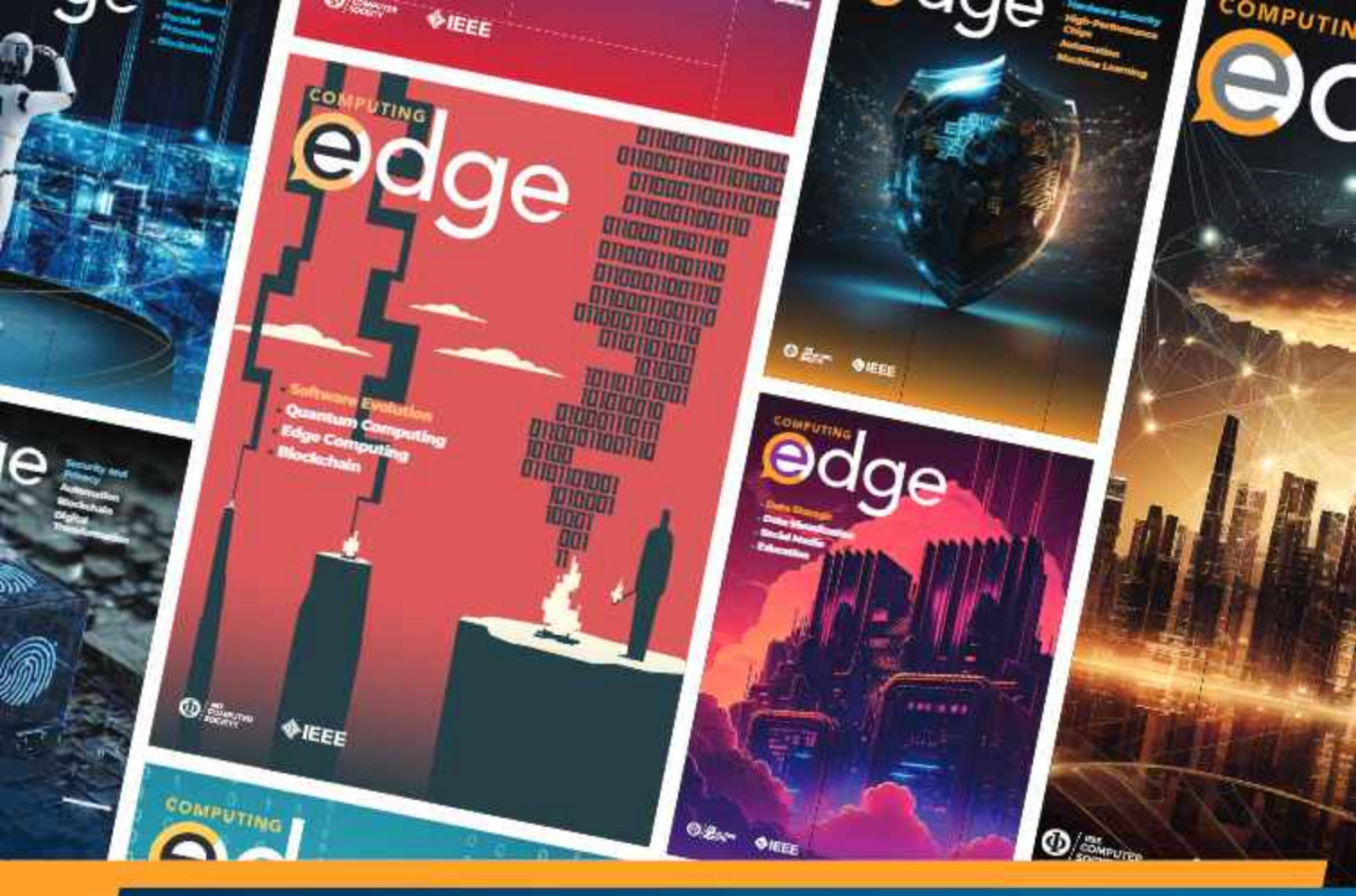
- JOB ALERTS
- TEMPLATES
- WEBINARS
- CAREER ADVICE
- RESUMES VIEWED BY TOP EMPLOYERS

No matter what your career level, the IEEE Computer Society Career Center keeps you connected to workplace trends and exciting career prospects.

IEEE COMPUTER SOCIETY

IEEE

Digital Object Identifier 10.1109/MC.2025.3562106



ComputingEdge

Your one-stop resource for industry hot topics, technical overviews, and in-depth articles.

- Cutting-edge articles from the IEEE Computer Society's portfolio of 12 magazines.
- Unique original content by computing thought leaders, innovators, and experts.
- Keeps you up to date on what you need to know across the technology spectrum.

Subscribe for free

www.computer.org/computingedge

IEEE COMPUTER SOCIETY

IEEE